On Adaptive HMM State Estimation

Jason J. Ford and John B. Moore, *Fellow, IEEE*

***Abstract*— In this paper new online adaptive hidden Markov model (HMM) state estimation schemes are developed, based on extended least squares (ELS) concepts and recursive prediction error (RPE) methods. The best of the new schemes exploit the idempotent nature of Markov chains and work with a least squares prediction error index, using *a posteriori* estimates, more suited to Markov models then traditionally used in identification of linear systems.**

These new schemes learn the set of N Markov chain states, and the *a posteriori* probability of being in each of the states at each time instant. They are designed to achieve the strengths, in terms of computational effort and convergence rates, of each of the two classes of earlier proposed adaptive HMM schemes without the weaknesses of each in these areas. The computational effort is of order N.

Implementation aspects of the proposed algorithms are discussed, and simulation studies are presented to illustrate convergence rates in comparison to earlier proposed online schemes.

***Index Terms*— Hidden Markov model, parameter estimation, recursive estimation.**

NOMENCLATURE

$\alpha_{k\|k,\theta}$, $\alpha_{k\|k-1,\theta}$	Unnormalized conditional estimates.
θ	Unknown parameters, i.e., state levels.
$\hat{\theta}_k$, $\hat{\theta}_k^*$	Estimates of state levels.
$\hat{\Theta}_k$	Collection of estimates.
$\tilde{\theta}_k$	Error in estimate.
κ_k, ϕ_k	Gradients.
$\lambda(\theta)$	Parameterized HMM model.
$\underline{1}$	Column vector of all ones.
A, a_{ij}	Stochastic transition matrix, elements.
$B(y_k, \theta)$, $b_k(i)$	Matrix of observation probabilities, elements.
e_i	Unit vector.
$\mathcal{F}_k$	Filtration of X_k.
M_k	Martingale increment.
N	Number of states.
N_θ	Number of parameters.
N_k	Normalization factor.
P_k	Approximation of the Hessian.
$\hat{P}_k$, $\hat{P}_k^*$	Estimates of P_k.
X_k	Markov state at time k.
$X_k^{(i)}$, θ^i	ith element.
$\mathcal{X}$	Diagonal matrix with X on the diagonal.
$\hat{X}_{k\|k,\theta}$, $\hat{X}_{k\|k,\hat{\Theta}_k}$, $\hat{X}_{k\|k-1,\theta}$, $\hat{X}_{k\|k-1,\hat{\Theta}_{k-1}}$	Conditional expectations of X_k.
$V_k(\theta)$, $\overline{V}_k(\theta)$	Cost functions.
w_k, n_k	Noise terms.
y_k	Observations.
Y_k	Collection of observations up until time k.
$\mathcal{Y}_k$	Filtration of Y_k.
ELS	Extended least squares.
EM	Expectation-maximization.
HMM	Hidden Markov model.
ODE	Ordinary differential equation.
RLS	Recursive least squares.
RPE	Recursive prediction error.
QAM	Quadrature amplitude modulation.
WGN	White Gaussian noise.
$\langle ., . \rangle$	Inner product.
diag(.)	Diagonal matrix from a vector.
$E[.]$, $E[.\|.]$	Expectation operation.

Manuscript received November 7, 1995; revised August 12, 1997. This work supported by the activities of the Cooperative Research Centre for Robust and Adaptive Systems by the Australian Government under the Cooperative Research Centres Program and by the EPSRC (U.K.) for the Senior Visiting Fellowship held at the Centre for Process Systems Engineering at Imperial College, London, U.K. The associate editor coordinating the review of this paper and approving it for publication was Dr. Petar M. Djuric.

The authors are with the Department of Systems Engineering and Cooperative Research Centre for Robust and Adaptive Systems, Research School of Information Sciences and Engineering, Australian National University, Canberra ACT 0200 Australia (e-mail: Jason.Ford@anu.edu.au).

Publisher Item Identifier S 1053-587X(98)01348-8.

I. INTRODUCTION

HIDDEN Markov models (HMM's) are a powerful tool in the field of signal processing [1], [2] with application to speech processing [4], digital communication systems [3], and biological signal processing [6]. The major limitation of schemes for the estimation of HMM parameters revolve around computation and memory requirements.

HMM's in discrete time can be viewed as having a state X_k at time k belonging to a discrete set, without loss of generality, denoted as $S = \{e_1, e_2, \cdots, e_N\}$, where e_i is a vector that is zero everywhere excepting the ith element, which is one. There are transitions between states described by fixed probabilities that form a matrix $A = (a_{ij})$, where a_{ij} is the probability of transferring from state e_j to state e_i. The model measurements are an output mapping from the Markov states $\theta' X_k$ contaminated by additive noise. The corruption of the output mapping is the reason the model is termed hidden.

The expectation-maximization (EM) algorithm [4] is a popular off-line locally convergent scheme for obtaining maximum likelihood estimates of the HMM parameters. A major limitation of off-line multipass estimation schemes is the "curse of dimensionality," where the computational effort

and memory requirements are in proportion to the square of the number of Markov states and proportional to the length of the signal to be processed. One avenue to improve the computational and memory requirements would appear to be through the investigation of on-line schemes. It should also be said that in learning the model parameters in a multipass arrangement, convergence rates are linear, meaning of order $1/N$ with respect to the number of passes N through the data.

The two notable examples of on-line adaptive schemes for HMM parameters estimation are the recursive Kullback–Leibler (RKL) scheme [5] and the recursive prediction error (RPE) scheme [8]. The RKL scheme converges linearly, and each iteration of the parameter update equation has computational complexity of $O(N_\theta)$, where N_θ is the number of parameters to be estimated. The RPE scheme [8] was developed with the aim to provide improved convergence rates. This scheme is known to be asymptotically efficient and provide quadratic convergence (of $O(1/N^2)$). However, each iteration of the parameter update equation has computational complexity of $O(N_\theta^2)$, which can be prohibitive for large N_θ.

The key contributions of this paper are the proposal of several new on-line schemes for HMM parameter estimation, based on extended least squares (ELS) and recursive prediction error (RPE) concepts with the ELS approach rationalized via martingale convergence results, and convergence results shown for the RPE schemes via an ordinary differential equation (ODE) approach. The best of these new schemes are based on a least squares prediction error index that uses *a posteriori* estimates rather than prediction estimates.

A typical application in the simplest of contexts, under study in a companion paper, is the demodulation of coded QAM signals with known transition probabilities in a noisy fading channel. The state transition probabilities and channel noise statistics would be assumed known, but the channel gain and phase changes are unknown and possibly time varying. The problem of estimating the transition probabilities is considered in a companion paper [7].

The paper is organized as follows. In Section III, we formulate the signal model and introduce an information state model. In Section IV, we introduce first the simplest case of the adaptive estimation task, namely, when the state sequence is measured directly, and then apply the least squares approach familiar in linear system identification. When the state sequence is not measured directly, the least squares approach leads to the ELS algorithms. Some convergence results are presented. In Section V, we generalize the ELS algorithms by introducing RPE recursion schemes, with new search directions, and ODE convergence results are presented. A new cost function, which is suggested by the least squares approach, and the *a posteriori* weighted RPE scheme is also presented. In Section VI, implementation considerations are discussed, and simulation examples are given. Finally, some conclusions are presented in Section VII.

II. Problem Formulation

In this section, we describe the HMM signal model in state space form, discuss its parameterization, and reformulate it as an information state model.

A. HMM State Space Model

Let X_k be a discrete-time homogeneous, first-order Markov process belonging to a finite set. The state space X, *without loss of generality*, can be identified with a set of unit vectors $S = \{e_1, e_2, \cdots, e_N\}$, $e_i = (0, \cdots, 0, 1, 0, \cdots, 0)' \in \mathrm{R}^N$ with 1 in the ith position. The transition probability matrix is $\mathbf{A} = (a_{ij})$ for $1 \le i, j \le N$, where

$$a_{ij} = P(X_{k+1} = e_i | X_k = e_j) \tag{2.1}$$

so that

$$E[X_{k+1}|X_k] = AX_k \tag{2.2}$$

where $E[.]$ denotes the expectation operator. We also denote $\{\mathcal{F}_l, l \in \mathcal{Z}^+\}$ to be the complete filtration generated by X, that is, for any $k \in \mathcal{Z}^+$, $\mathcal{F}_k$ is the complete filtration generated by X_ℓ, $\ell \le k$. For a brief introduction of the concept of filtration see [2, Appendix A].

Lemma 1: The dynamics of X_k are given by the state equation

$$X_{k+1} = AX_k + M_{k+1} \tag{2.3}$$

where M_{k+1} is a $(A, \mathcal{F}_k)$ martingale increment in that $E[M_{k+1}|\mathcal{F}_k] = 0$.

Proof [2], [11]:

$$\begin{aligned} E[M_{k+1}|\mathcal{F}_k] &= E[X_{k+1} - AX_k | X_k, A] \\ &= E[X_{k+1}|X_k, A] - AX_k = 0. \end{aligned}$$

□

We assume X_k is hidden, that is, indirectly observed by measurements y_k in a continuous range $\mathbf{R}$. The *observation process* y_k is assumed to be scalar (for simplicity of presentation only) and to have the form

$$y_k = \theta' X_k + w_k = X_k'\theta + w_k \in \mathrm{R} \tag{2.4}$$

where $\theta \in \mathrm{R}^N$ is the vector of state values of the Markov chain. We also define $Y_k \triangleq (y_0, \cdots, y_k)$. We assume w_k is i.i.d., with zero mean and Gaussian density, i.e., $w_k \sim N[0, \sigma_w^2]$ and $E[w_{k+1}|\mathcal{F}_k \vee \mathcal{Y}_k] = 0$, where $\mathcal{Y}_l$ is the complete filtration generated by y_k, $k \le l$.

We shall define the vector of parameterized probability densities (or symbol probabilities) as $\mathbf{b}_k = (b_k(i))$, for $b_k(i) \triangleq P(y_k|X_k = e_i)$. In the special case as here when w_k is i.i.d. and $N[0, \sigma_w^2]$, we can write

$$b_k(i) = \frac{1}{\sqrt{2\pi\sigma_w^2}} \exp\left[\frac{-(y_k - \theta' e_i)^2}{2\sigma_w^2}\right]. \tag{2.5}$$

We can also write that the initial state probability vector for the Markov chain $\pi = (\pi_i)$ is denoted by $\pi_i = P(X_1 = e_i)$. The HMM is denoted $\lambda = (A, \theta, \pi, \sigma_w^2)$.

1) Model Parameterization: For simplicity, in this paper, we shall be considering the problem of estimating unknown state values, assuming knowledge of A, π, and σ_w^2, as in communication channels with known coding. Let λ be parameterized by an unknown vector θ so that the parameterized HMM is denoted by

$$\lambda(\theta) = (A, \theta, \pi, \sigma_w^2).$$

B. Information State Model

Let $\hat{X}_{k|k,\theta}$ and $\hat{X}_{k|k-1,\theta}$ denote the conditional filtered state estimate and one-step-ahead state prediction of X_k at time k, given measurements $\mathbf{Y}_k$, up until time k and the parameter θ i.e.,

$$\begin{aligned}\hat{X}_{k|k,\theta} &\triangleq E[X_k|\mathcal{Y}_k, \theta] \\ \hat{X}_{k|k-1,\theta} &\triangleq E[X_k|\mathcal{Y}_{k-1}, \theta].\end{aligned} \tag{2.6}$$

Note that $\hat{X}_{k|k-1,\theta} = \mathbf{A}\hat{X}_{k-1|k-1,\theta}$.

It is often convenient to work with an unnormalized conditional estimates (or the so-called "forward" variables) $\alpha_{k|k,\theta}$ and $\alpha_{k|k-1,\theta}$, which are defined as $\alpha_{k|k,\theta} = (\alpha_{k|k,\theta}(i))$ for $\alpha_{k|k,\theta}(i) \triangleq P(Y_k, X_k = e_i|\theta)$ and $\alpha_{k|k-1,\theta} = A\alpha_{k-1|k-1,\theta}$. These unnormalized conditional estimates are computed using the "forward" recursion

$$\begin{aligned}\alpha_{k+1|k+1,\theta} &= \mathbf{B}(y_{k+1}, \theta)\mathbf{A}\alpha_{k|k,\theta} \\ \alpha_{k+1|k,\theta} &= \mathbf{AB}(y_k, \theta)\alpha_{k|k-1,\theta}\end{aligned} \tag{2.7}$$

where $\mathbf{B}(y_k, \theta) = \text{diag}([b_k(1), \cdots, b_k(N)])$.

We can now write the conditional filter estimate and one-step-ahead prediction from the unnormalized conditional estimates as

$$\begin{aligned}\hat{X}_{k|k,\theta} &= \langle \alpha_{k|k,\theta}, \underline{1}\rangle^{-1}\alpha_{k|k,\theta} \\ \hat{X}_{k|k-1,\theta} &= \langle \alpha_{k|k-1,\theta}, \underline{1}\rangle^{-1}\alpha_{k|k-1\theta}\end{aligned} \tag{2.8}$$

where $\langle ., . \rangle$ is an inner product, and $\underline{1}$ is the column vector containing all ones.

1) Parameterized Filtered Estimate: We now seek to express the observations y_k in terms of the conditional filter estimate at time k.

Lemma 2 [8]: The conditional measurements y_k are given by

$$y_k = \theta'\hat{X}_{k|k,\theta} + n_k$$

where

$$n_k = \theta'[X_k - \hat{X}_{k|k,\theta}] + w_k.$$

The parameterized filtered estimates $\hat{X}_{k|k,\theta}$ are given by

$$\begin{aligned}\hat{X}_{k|k,\theta} &= N_k\mathbf{B}(y_k, \theta)\mathbf{A}\hat{X}_{k-1|k-1,\theta} \\ y_k &= \theta'\hat{X}_{k|k,\theta} + n_k\end{aligned} \tag{2.9}$$

where $N_k = \langle \mathbf{B}(y_k, \theta)\mathbf{A}\hat{X}_{k-1|k-1,\theta}, \underline{1}\rangle^{-1}$ is a normalization factor.

2) Parameterized One-Step-Ahead Prediction: We now seek to express the observations y_k in terms of a prediction based on the conditional filter estimates at time $k-1$.

Lemma 3 [8]: In the above notation, the measurements y_k are given by

$$y_k = \theta'\hat{X}_{k|k-1,\theta} + n_k$$

where

$$n_k = \theta'[X_k - \hat{X}_{k|k-1,\theta}] + w_k \tag{2.10}$$

and n_k is a white $(\theta, \mathcal{Y}_{k-1})$ martingale increment.

Proof [8]: Following the standard arguments, since $\hat{X}_{k|k-1\theta}$ is measurable with respect to θ, $\mathcal{Y}_{k-1}$, $E[w_{k+1}|\mathcal{Y}_k] = 0$ and $E[M_{k+1}|\mathcal{Y}_k] = 0$, then

$$E[n_k|\theta, \mathcal{Y}_{k-1}] = \theta'\hat{X}_{k|k-1,\theta} - \theta'\hat{X}_{k|k-1,\theta} = 0.$$

□

In summary, the parameterized predictor-based signal model for an HMM parameter θ and with states $\hat{X}_{k|k-1,\theta}$ is given by

$$\begin{aligned}\hat{X}_{k+1|k,\theta} &= N_k\mathbf{AB}(y_k, \theta)\hat{X}_{k|k-1,\theta} \\ y_k &= \theta'\hat{X}_{k|k-1,\theta} + n_k\end{aligned} \tag{2.11}$$

where n_k is a $(\theta, \mathcal{Y}_{k-1})$ martingale increment, and $N_k = \langle \mathbf{AB}(y_k, \theta)\hat{X}_{k|k-1,\theta}, \underline{1}\rangle^{-1}$ is a normalization factor.

We now proceed to consider the problem of estimating θ given a sequence of observations Y_k.

III. Least Squares and Extended Least Squares

This section has two parts. In the first part, to introduce the problem, we consider the simplified adaptive estimation task for the case when X_k is measured. The familiar least squares algorithm from linear system identification theory is presented, and the idempotent nature of indicator vectors is exploited. The least squares cost is introduced, both in its original form and in an alternative form appropriate to the new least squares recursion. Convergence results are presented.

In the second part of this section, the assumption that X_k is measured is relaxed, and the *ad hoc* idea of ELS is introduced. The two least squares algorithms are converted, via various assumptions, to a collection of ELS algorithms. Computational and convergence aspects are discussed. The importance of studying these ELS algorithms is both as a motivation for locally convergence RPE algorithms presented in Section IV and as suboptimal computationally efficient approximations to these same RPE algorithms.

A. Least Squares

In this subsection, we consider signal model (2.4) given in the previous section and the idealized estimation task to estimate θ given a sequence of observations Y_k and the state sequence $X_1, X_2, \cdots, X_k$. Subsequently, we will consider the case when X_k must be estimated from $\mathcal{Y}_k$.

1) Off Line: In a familiar approach, premultiplication of (2.4) by X_k and some algebraic manipulation leads to the off-line estimate for θ ($\hat{\theta}_k$) based on k data points

$$\hat{\theta}_k = \left(\sum_{i=1}^{k} X_iX_i'\right)^{-1}\sum_{i=1}^{k} X_iy_i \tag{3.1}$$

where $\hat{\theta}_k$ is an estimate of θ given k data points. The estimation error is

$$\tilde{\theta}_k = \theta - \hat{\theta}_k = \left(\sum_{i=1}^{k} X_iX_i'\right)^{-1}\sum_{i=1}^{k} X_iw_i. \tag{3.2}$$

By exploiting the idempotent nature of indicator vectors, the above estimates (3.1) and (3.2) can be written as scalar

equations

$$\hat{\theta}_k^{(j)} = \left[\sum_{i=1}^{k} X_i^{(j)}\right]^{-1} \sum_{i=1}^{k} X_i^{(j)} y_i$$
$$\tilde{\theta}_k^{(j)} = \left[\sum_{i=1}^{k} X_i^{(j)}\right]^{-1} \sum_{i=1}^{k} X_i^{(j)} w_i, \qquad \text{for } j = 1, 2, \cdots, N \tag{3.3}$$

where $\hat{\theta}_k^{(j)}$, $\tilde{\theta}_k^{(j)}$, and $X_i^{(j)}$ denote the jth element of $\hat{\theta}_k$, $\tilde{\theta}_k$, and X_i, respectively.

We are led to the following lemma.

Lemma 4: In the above notation, and with w_k a martingale increment with respect to the σ-algebra $\mathcal{F}_{k-1}^0 = \{X_1, \cdots, X_k\}$, with $\mathcal{F}_{k-1}$ as the complete filtration, in that $E[w_k|\mathcal{F}_{k-1}] = 0$, then

$$\lim_{k\to\infty} \hat{\theta}_k \text{ exist a.s.} \tag{3.4}$$

Moreover, for each j in (3.3)

$$\lim_{k\to\infty} \hat{\theta}_k^{(j)} = \theta^{(j)} \text{a.s.} \iff \lim_{k\to\infty} \left[\sum_{i=1}^{k} X_i^{(j)}\right]^{-1} = 0. \tag{3.5}$$

Proof: We proceed first to prove the second result by examining subsequences of the chain on which each state is active. Let $n(k)^j$ denote the number of times state j is active up until time k, i.e., $n(k)^j = \sum_{i=1}^{k} X_i^{(j)}$, and let $a(i)^j$ denote the time k at which the state j is active for the ith time.

From (3.3), we define $W_k^j \triangleq \sum_{i=1}^{k} [\sum_{i=1}^{k} X_i^{(j)}]^{-1} X_i^{(j)} w_i$, which is a martingale adapted to $\mathcal{F}_k$ since $E[W_{k+1}|\mathcal{F}_k] = W_k$. Note that $W_k^j = \sum_{i=1}^{n(k)^j} (1/i) w_{a(i)^j}$ by summing only over the subsequence with active j; then, it follows that W_k^j is bounded in L_2 for each j by

$$E[(W_k^j)^2] = E\left[\left(\sum_{i=1}^{n(k)^j} \frac{1}{i} w_a(i)^j\right)\left(\sum_{\ell=1}^{n(k)^j} \frac{1}{\ell} w_a(\ell)^j\right)\right]$$
$$= E\left[\sum_{i=1}^{n(k)^j} \frac{1}{i^2} w_a(i)^j\right] \le B \sum_{i=1}^{n(k)^j} \frac{1}{i^2} < \infty$$

where we have used that $E[w_i w_\ell] = 0 \ \forall i \neq \ell$ and that $E[w_i w_i] \le B$ for all i.

For the only if direction of the second lemma result, note that under the lemma condition $\lim_{k\to\infty} [\sum_{i=1}^{k} X_i^{(j)}]^{-1} = 0$, we have $n(k)^j \to \infty$ as $k \to \infty$. Hence, by the martingale convergence [11], [12]

$$\lim_{k\to\infty} W_k^j \text{ exist a.s. for each } j.$$

Using the Kronecker lemma [11], [13], we have

$$\lim_{k\to\infty} \frac{1}{n(k)^j} \sum_{i=1}^{n(k)^j} w_i = 0, \qquad \text{a.s. for each } j.$$

Hence, by rewriting as a summation over k, we have

$$\lim_{k\to\infty} \frac{1}{n(k)^j} \sum_{i=1}^{k} X_i w_i = \lim_{k\to\infty} \tilde{\theta}_k^{(j)} = 0 \qquad \text{a.s. for each } j.$$

The result follows. The if direction of the second lemma result follows from noting that $\lim_{k\to\infty} [\sum_{i=1}^{k} X_i^{(j)}]^{-1} \neq 0$ implies that $n(k)^j$ is finite. Hence

$$0 < \frac{1}{n(k)^j} \sum_{i=1}^{n(k)^j} w_i < \infty, \qquad \text{w.p. } 1.$$

The first lemma result now also follows. □

2) On Line: Simple manipulations of (3.1) give the recursions

$$\hat{\theta}_k = \hat{\theta}_{k-1} + P_k X_k [y_k - X_k' \hat{\theta}_{k-1}] \tag{3.6}$$

and

$$P_k^{-1} = P_{k-1}^{-1} + X_k X_k' \tag{3.7}$$

or

$$P_k = P_{k-1} - P_{k-1} X_k [1 + X_k' P_{k-1} X_k]^{-1} X_k' P_{k-1} \tag{3.8}$$

where $\hat{\theta}_k$ is an estimate of θ after k data points. Manipulations show that $\hat{\theta}_k$ minimizes a squares sum index, that is

$$\hat{\theta}_k = \arg\min_{\theta} \sum_{i=1}^{k} (y_i - \theta' X_i)^2. \tag{3.9}$$

Now, we note that nonlinear functions on X_k are linear in X_k as $f(X_k) = \sum_{i=1}^{N} f(e_i) X_k^{(i)}$, where $X_k^{(i)}$ denotes the ith element of X_k. The above recursion can be rewritten in an alternative form as

$$\hat{\theta}_k = \hat{\theta}_{k-1} + \sum_{i=1}^{N} P_k e_i [y_k - \hat{\theta}_{k-1}' e_i] X_k^{(i)} \tag{3.10}$$

and

$$P_k^{-1} = P_{k-1}^{-1} + \sum_{i=1}^{N} e_i e_i' X_k^{(i)} \tag{3.11}$$

or

$$P_k = P_{k-1} - \sum_{i=1}^{N} P_{k-1} e_i [1 + e_i' P_{k-1} e_i]^{-1} e_i' P_{k-1} X_k^{(i)} \tag{3.12}$$

and likewise, it can be shown that $\hat{\theta}_k$ minimizes the linear index

$$\hat{\theta}_k = \arg\min_{\theta} \sum_{i=1}^{k} \sum_{j=1}^{N} (y_i - \theta' e_j)^2 X_i^{(j)}. \tag{3.13}$$

Remarks:

1) The condition that $[(1/k) \sum_{i=1}^{k} X_i^{(j)}]^{-1} \to 0$ as $k \to \infty$ in Lemma 4 is an excitation condition. In addition, it is possible to show, but is not done here, that the rate of convergence to zero of $\tilde{\theta}_k \tilde{\theta}_k'$ is as $1/k^{1/2}$. See also Sternby [15].
2) To reduce the number of calculations, (3.7) and (3.11) can be replaced by a stochastic approximation given by

$$P_k^{-1} = kE[X_k X_k'] = k \operatorname{diag}(E[X_k])$$

where $E[X_k]$ is a vector of the *a priori* probabilities of being in each state. $E[X_k]$ is given by the normalized eigenvector of A corresponding to the eigenvalue of value one.

We now proceed to consider the case when the state sequence is unknown.

B. Extended Least Squares

This subsection considers the estimation task when the state sequence is not known and is presented in a manner paralleling the previous section. To produce estimates when the state sequence is unknown, the *ad hoc* idea of extended least squares is to use estimates of states in lieu of actual states X_k in a least squares implementation; see [18].

In linear estimation, it is usual to use one-step-ahead predictions of the states so that the observation noise remains white, at least when the predictions are optimal in a least squares sense. However, in our case, the one-step-ahead predictions of Markov chains can be far from optimal, particularly when the active state changes. This property of HMM highlights that there are differences between standard linear estimation theory and the HMM parameter estimation problem. Initially, in this section, we proceed by using one-step-ahead predictions to ensure that the observation noise remains white, but this requirement is relaxed toward the end of this section.

Let $\hat{X}_{k|k-1,\hat{\Theta}_{k-1}}$ denote the conditional filter estimate based on observations and model estimates, i.e.,

$$\hat{X}_{k|k-1,\hat{\Theta}_{k-1}} \triangleq E[X_k|\mathcal{Y}_{k-1}, \hat{\Theta}_{k-1}]$$

where $\hat{\Theta}_{k-1} \triangleq \{\hat{\theta}_1, \cdots, \hat{\theta}_{k-1}\}$ and the recursion below is used to generate the one-step-ahead predictions

$$\hat{X}_{k+1|k,\hat{\Theta}_k^*} = N_k \mathbf{AB}(y_k, \hat{\theta}_k^*)\hat{X}_{k|k-1,\hat{\Theta}_{k-1}^*}. \tag{3.14}$$

1) Off Line: Using the *ad hoc* idea of replacing states by one-step-ahead predictions, the extended least squares version of the off-line least squares algorithms (3.3) is

$$\hat{\theta}_k^{(j)} = \left[\sum_{i=1}^{k} \hat{X}_{i|i-1,\hat{\theta}_{k-1}}^{(j)}\right]^{-1} \sum_{i=1}^{k} \hat{X}_{i|i-1,\hat{\theta}_{k-1}}^{(j)} y_i \quad \text{for } j = 1, 2, \cdots, N \tag{3.15}$$

where $\hat{\theta}_k$ is the estimate of θ on k points of data and $\hat{X}_{i|i-1,\hat{\theta}_{k-1}} = E[X_i|\mathcal{Y}_k, \hat{\theta}_{k-1}]$.

No convergence analysis is attempted here for the off-line ELS algorithm.

2) On Line: From the least squares recursion (3.6)–(3.8), substituting one-step-ahead predictions gives the recursions

$$\hat{\theta}_k^* = \hat{\theta}_{k-1}^* + \hat{P}_k^* \hat{X}_{k|k-1,\hat{\Theta}_{k-1}^*}\left[y_k - \hat{X}'_{k|k-1,\hat{\Theta}_{k-1}^*}\hat{\theta}_{k-1}^*\right] \tag{3.16}$$

$$\hat{P}_k^{*-1} = \hat{P}_{k-1}^{*-1} + \hat{X}_{k|k-1,\hat{\Theta}_{k-1}^*}\hat{X}'_{k|k-1,\hat{\Theta}_{k-1}^*} \tag{3.17}$$

or

$$\hat{P}_k^* = \hat{P}_{k-1}^* - \sum_{i=1}^{N} \hat{P}_{k-1}^* \hat{X}_{k|k-1,\hat{\Theta}_{k-1}^*} \cdot \left[1 + \hat{X}'_{k|k-1,\hat{\Theta}_{k-1}^*}\hat{P}_{k-1}^*\hat{X}_{k|k-1,\hat{\Theta}_{k-1}^*}\right]^{-1} \cdot \hat{X}'_{k|k-1,\hat{\Theta}_{k-1}^*}\hat{P}_{k-1}^*. \tag{3.18}$$

Likewise, from (3.10), we construct the ELS recursion

$$\hat{\theta}_k = \hat{\theta}_{k-1} + \hat{P}_k \hat{\mathcal{X}}_{k|k-1,\hat{\Theta}_{k-1}} \sum_{i=1}^{N} e_i\left[y_k - \hat{\theta}'_{k-1}e_i\right] \tag{3.19}$$

$$\hat{P}_k^{-1} = \hat{P}_{k-1}^{-1} + \hat{\mathcal{X}}_{k|k-1,\hat{\Theta}_{k-1}} \tag{3.20}$$

or

$$\hat{P}_k = \hat{P}_{k-1} - \sum_{i=1}^{N} \hat{P}_{k-1}e_i e_i'\hat{P}_{k-1} \cdot \left[1 + e_i'\hat{P}_{k-1}e_i\right]^{-1} \hat{X}^{(i)}_{k|k-1,\hat{\Theta}_{k-1}}. \tag{3.21}$$

Note that for (3.20) and (3.21), if $\hat{P}_0$ is diagonal, then $\hat{P}_k$ will be diagonal for all k. Hence, (3.20) and (3.21) can be explicitly written as scalar equations

$$\hat{p}_{i,k}^{-1} = \hat{p}_{i,k-1}^{-1} + \hat{X}^{(i)}_{k|k-1,\hat{\Theta}_{k-1}}$$

or

$$\hat{p}_{i,k} = \hat{p}_{i,k-1} - \frac{p_{i,k-1}^2}{1+\hat{p}_{i,k-1}} \hat{X}^{(i)}_{k|k-1,\hat{\Theta}_{k-1}} \tag{3.22}$$

where $\hat{P}_k = \text{diag}([\hat{p}_{1,k}, \cdots, \hat{p}_{i,k}, \cdots, \hat{p}_{N,k}])$.

From a computational point of view, the diagonal nature of $\hat{P}_k$ in (3.19) means that the (3.19)–(3.21) is computationally more attractive for obtaining the θ estimates then (3.16)–(3.18) because it is of order N rather than N^2 in complexity.

However, (3.16)–(3.18) are not completely satisfactory because the estimates produced are biased. To see this, we now examine the convergence properties of (3.16)–(3.18) for the idealized case when

$$\hat{X}_{k|k-1,\hat{\Theta}_{k-1}} = \hat{X}_{k|k-1,\theta} = E[\hat{X}_k|\mathcal{Y}_{k-1}, \theta]. \tag{3.23}$$

Lemma 5: Consider the ELS scheme (3.19)–(3.21) in the idealized case, when (3.23) holds. Then, as $k \to \infty$

$$\hat{\theta}_k \to \hat{P}_k\left[\sum_{j=1}^{k} \hat{X}_{j|j-1,\theta}X'_j\right]\theta \text{ a.s.} \to \hat{P}_k\hat{P}_k^{*-1}\theta \text{ a.s.} \tag{3.24}$$

Moreover, in the case of excitation such that $\hat{P}_k^* \to 0$ as $1/k$ as $k \to \infty$, then for (3.16)–(3.18)

$$\hat{\theta}_k^* = \hat{P}_k^*\hat{P}_k^{-1}\hat{\theta}_k \to \theta \text{ a.s.} \tag{3.25}$$

Proof: Simple manipulations from (3.19) give

$$\hat{\theta}_k = \hat{P}_k\left[\sum_{j=1}^{k} \hat{X}_{j|j-1,\theta}X'_j\right]\theta + \left[\hat{P}_k \sum_{j=1}^{k} \hat{X}_{j|j-1,\theta}w_j\right] \tag{3.26}$$

and now

$$\left[\frac{1}{k}\sum_{j=1}^{k} \hat{\mathcal{X}}_{j|j-1,\theta}\right]^{-1}\left[\frac{1}{k}\sum_{j=1}^{k} \hat{X}_{j|j-1,\theta}w_j\right] \to 0 \text{ a.s.} \tag{3.27}$$

since each element of the second term can be shown to go to zero using martingale convergence results and the Kronecker

Lemma, as in the proof of Lemma 4. In addition, observe that

$$\hat{P}_k \left[\sum_{j=1}^{k} \hat{\mathcal{X}}^2_{j|j-1,\theta} \right] \leq I \tag{3.28}$$

since $0 \leq \hat{X}^2 \leq \hat{X} \leq 1$. Clearly, (3.25)–(3.27) yield the first half of (3.24) as claimed.

Now, observe that since $E[X_j - \hat{X}_{j|j-1,\theta} | \mathcal{Y}_{j-1}, \theta] = 0$, then likewise

$$\hat{P}_k \left[\sum_{j=1}^{k} \hat{X}_{j|j-1,\theta} (X_j - \hat{X}_{j|j-1,\theta})' \right] \to 0 \text{ a.s.} \tag{3.29}$$

to yield the second half of (3.24).

Now, from (3.16) and (3.19), simple manipulations lead to

$$\begin{aligned} \hat{\theta}^*_k &= \hat{P}^*_k \left[\sum_{j=1}^{k} \hat{X}_{j|j-1,\theta} y_j \right] \\ &= \hat{P}^*_k \hat{P}_k^{-1} \hat{\theta}_k \end{aligned} \tag{3.30}$$

and the result (3.25) follows from (3.26) under the excitation assumption that $\hat{P}^*_k \to 0$ so that $\| \hat{P}^*_k \hat{P}_k^{-1} \|$ is bounded above, and $\hat{P}^*_k \sum_{j=1}^{k} \hat{X}_{j|j-1,\theta} w_j \to 0$ a.s. as $k \to \infty$. □

Remarks:

1) The lemma result (3.25) holds without the excitation condition on $\hat{P}^*_k$, which incidentally assures a convergence rate of $1/k^{1/2}$, but more advanced theory such as in [15] is required.
2) Lemma 5 demonstrates that the scheme (3.19)–(3.21) leads to biased estimates, whereas the scheme (3.16)–(3.18) does not. Hence, because both schemes are of similar complexity, it seems that the scheme (3.16)–(3.18) should be used in preference. See the simulations section for a demonstration of the bias.
3) The complete ELS convergence analysis of (3.16)–(3.18) when (3.23) does not hold is virtually identical to that given in [16] and is not repeated here. Suffice it to say, a key sufficient condition for (3.23)–(3.25) to hold asymptotically is that a certain passivity condition holds, that is, the system driven by $\tilde{\theta}'_k \hat{X}_k$ and with output $\frac{1}{2} \tilde{\theta}'_k \hat{X}_k + \theta' \tilde{X}_k$ must be strictly passive (here $\tilde{\theta} = \theta - \hat{\theta}$ and $\tilde{X} = X - \hat{X}$). This system in the HMM case is nonlinear and is sample path dependent; therefore, further explorations along this line seems pointless.
4) Consider a hybrid version of (3.16)–(3.18)

$$\hat{\theta}^\dagger_k = \hat{\theta}^\dagger_{k-1} + \hat{P}_k \hat{X}_{k|k-1,\hat{\Theta}^\dagger_{k-1}} \left[y_k - \hat{X}'_{k|k-1,\hat{\Theta}^\dagger_{k-1}} \hat{\theta}^\dagger_{k-1} \right] \tag{3.31}$$

with (3.20) and (3.21) holding. We do not study (3.31) further here.
5) To further reduce the number of calculations required to estimate θ, (3.20) can be replaced by an stochastic approximation given by

$$P_k^{-1} = kE[\hat{\mathcal{X}}_{k|k-1,\hat{\Theta}_{k-1}}] = k \operatorname{diag}(E[X_k]).$$

C. A Posteriori Extended Least Squares

It is the nature of HMM's that the one-step-ahead predictions of the state can be far from optimal, particularly when the active state changes.

Hence, here we consider a ELS algorithm based on filtered estimates rather than one-step-ahead predictions. Consider a modified version of the (3.19)–(3.21) scheme.

$$\hat{\theta}_k = \hat{\theta}_{k-1} + \hat{P}_k \hat{\mathcal{X}}_{k|k,\hat{\Theta}_{k-1}} \sum_{i=1}^{N} e_i \left[y_k - \hat{\theta}'_{k-1} e_i \right] \tag{3.32}$$

$$\hat{P}_k^{-1} = \hat{P}_{k-1}^{-1} + \hat{\mathcal{X}}_{k|k,\hat{\Theta}_{k-1}}, \tag{3.33}$$

or

$$\begin{aligned} \hat{P}_k = \hat{P}_{k-1} &- \sum_{i=1}^{N} \hat{P}_{k-1} e_i e_i' \hat{P}_{k-1} \\ &\cdot \left[1 + e_i' \hat{P}_{k-1} e_i \right]^{-1} \hat{X}^{(i)}_{k|k,\hat{\Theta}_{k-1}}. \end{aligned} \tag{3.34}$$

Remarks:

1) The recursion (3.32)–(3.34) is computationally efficient because $\hat{P}_k^{-1}$ can be forced to be diagonal.
2) Unlike the recursions (3.19)–(3.21), the recursion (3.32)–(3.34) produces consistent results in simulations; see Section V.

This ELS recursion is the most attractive of the algorithms presented in this section; however, no martingale convergence analysis is available for the (3.32)–(3.34) scheme at present since the error term $(X_k - \hat{X}_{k|k})$ is not a martingale increment. Rather than proceed with a further analysis of this *a posteriori* ELS scheme, we proceed to look at RPE algorithms that are mildly more complicated but have a more complete theory.

IV. Recursive Prediction Algorithms

There exists mature theory for recursive estimation or identification of continuous discrete-time models based on the minimization of the prediction error costs; see [9]. This theory provides asymptotic quadratic convergent algorithms (admittedly local) for linear and nonlinear models.

In this section, we proceed by applying this mature theory in order to obtain asymptotic convergence algorithms that generalize the ELS schemes of the last section. First, we reintroduce the two cost functions from the least squares discussion to replace the usual prediction error cost. These cost functions are our criteria for estimation of θ. We present the RPE algorithm that minimizes each of these cost functions.

Following from this, we present an RPE algorithm corresponding to the attractive *a posteriori* extended least squares algorithm (3.32)–(3.34).

A. Prediction Error Cost Functions

First, we present the RPE schemes corresponding to the computational efficient one-step-ahead prediction-based ELS schemes in (3.19)–(3.21) and (3.31).

Consider the error cost functions

$$V_k(\theta) \triangleq \tfrac{1}{2} \sum_{i=2}^{k} (y_i - \theta' \hat{X}_{i|i-1,\hat{\Theta}_{i-1}})^2 \tag{4.1}$$

and

$$\overline{V}_k(\theta) \triangleq \tfrac{1}{2} \sum_{i=2}^{k} \sum_{j=1}^{N} (y_i - \theta' e_j)^2 \hat{X}^{(j)}_{i|i-1,\hat{\Theta}_{i-1}}. \quad (4.2)$$

The following RPE recursion minimizes locally the index in (4.1) (see Lemma 6) and generalizes the ELS recursion (3.31)

$$\begin{aligned} \hat{\theta}^{\dagger}_k &= \hat{\theta}^{\dagger}_{k-1} + \hat{P}^{\dagger}_k \psi_{k|\hat{\theta}^{\dagger}_{k-1}} \\ \hat{P}^{\dagger -1}_k &= \hat{P}^{\dagger -1}_{k-1} + \hat{\mathcal{X}}_{k|k-1,\hat{\Theta}^{\dagger}_{k-1}} \end{aligned} \quad (4.3)$$

where

$$\psi_{k|k-1,\hat{\Theta}^{\dagger}_{k-1}} \triangleq \frac{d}{d\theta}(y_k - \hat{X}'_{k|k-1,\hat{\Theta}_{k-2},\theta})^2|_{\theta=\hat{\theta}^{\dagger}_{k-1}}$$

and where $\hat{P}^{\dagger -1}_k$ is an approximation for the second derivative of $V_k(\theta)$. From (3.14)

$$\hat{X}_{k+1|k,\hat{\Theta}^{\dagger}_k} = N_k A B(y_k, \hat{\theta}^{\dagger}_k) \hat{X}_{k|k-1,\hat{\Theta}^{\dagger}_{k-1}}.$$

The RPE recursion that minimizes locally the index (4.2) (see Lemma 6) and generalizes the ELS recursion (3.19)–(3.21) is

$$\begin{aligned} \hat{\theta}_k &= \hat{\theta}_{k-1} + \hat{P}_k \kappa_{k|k-1,\hat{\Theta}_{k-1}} \\ \hat{P}^{-1}_k &= \hat{P}^{-1}_{k-1} + \hat{\mathcal{X}}_{k|k-1,\hat{\Theta}_{k-1}} \end{aligned} \quad (4.4)$$

where, with $\kappa^{(i)}$ denoting the ith element of κ

$$\begin{aligned} \kappa^{(i)}_{k|k-1,\hat{\Theta}_{k-1}} &= \hat{X}^{(i)}_{k|k-1,\hat{\Theta}_{k-1}} [y_k - \hat{\theta}^{(i)}_{k-1}] \\ &\quad - \sum_{j=1}^{N} \left. \frac{d\hat{X}^{(j)}_{k|k-1,\hat{\Theta}_{k-2},\theta}}{d\theta^{(i)}} \right|_{\theta=\hat{\theta}_{k-1}} [y_k - \hat{\theta}^{(j)}_{k-1}]^2. \end{aligned}$$

Convergence of both these RPE algorithms can be established by a conventional ODE analysis [9]. Since the state estimates $\hat{\mathcal{X}}_{i|i-1,\hat{\Theta}_{i-1}}$ and $\hat{X}_{i|i-1,\hat{\Theta}_{i-1}}$ are of necessity bounded, a projection into a stability domain as required in [8] is implicit here.

Actually, the ODE analysis requires that the filter generating $\hat{X}_{i|i-1,\hat{\Theta}_{i-1}}$ be exponentially stable. This exponential stability, in the sense that initial conditions are forgotten exponentially, is established in [17] for the $N = 2$ case and is known to hold more generally under reasonable conditions not spelt out here.

To demonstrate convergence of (4.3) and (4.4), let us first define for (4.3) and (4.4), respectively, and arbitrary θ

$$f(\theta, k) = E[\psi_{k|\theta} \varepsilon_{k|k-1,\theta}] \text{ or } E[\kappa_{k|k-1,\theta}] \quad (4.5)$$

and

$$G(\theta, k) = E[\hat{\mathcal{X}}_{k|k-1,\theta}], \quad (4.6)$$

The following lemma now holds

Lemma 6: The recursions (4.3) and (4.4) will converge a.s. to the set $\overline{D}_c = \{\theta | \lim_{k\to\infty} E[f(\theta, k)] = 0\}$; moreover, under the excitation condition $\hat{P}_k$ (or $\hat{P}^{\dagger}_k$) $\to 0$ as $1/k$, then convergence of $\hat{\theta}_k$ (or $\hat{\theta}^{\dagger}_k$) is at the rate $1/k^{1/2}$.

Proof: The ODE's associated with (4.3) and (4.4) for fixed k under (4.5) and (4.6) are

$$\begin{aligned} \frac{d}{d\tau}\theta(\tau, k) &= R^{-1}(\tau, k) f(\theta(\tau, k), k) \\ \frac{d}{d\tau} R(\tau, k) &= G(\theta(\tau, k), k), \quad R_0(k) \geq \delta I. \end{aligned} \quad (4.7)$$

Now, Lyapunov functions for (4.7) under (4.5) and (4.6) are

$$\hat{W}(\tau, k) = E[(y_k - \theta' \hat{X}_{k|k-1,\theta})^2]$$

or

$$\hat{W}(\tau, k) = E\left[\sum_{j=1}^{N} (y_k - \theta' e_j)^2 \hat{X}^{(j)}_{k|k-1,\theta}\right] \quad (4.8)$$

so that

$$\begin{aligned} \frac{d}{d\tau}\hat{W}(\tau, k) &= \frac{d\hat{W}(\tau, k)}{d\theta'_\tau} \frac{d\theta_\tau}{d\tau} \\ &= -f'(\theta(\tau, k), k) R^{-1}_\tau(k) f(\theta(\tau, k), k). \end{aligned} \quad (4.9)$$

Thus, $\hat{W}(\tau, k)$ converges for all k and $\tau \to \infty$, and $\theta(\tau, k)$ converges to the set $\{\theta | E[f(\theta, k)] = 0\}$.

Applying the ODE theory of Ljung [9], the various regularity conditions are satisfied here, and the first result claimed follows.

Observe from (4.9) that if $R_\tau(k)$ is of the order $1/k$, as under suitable excitation, then $f(\theta(\tau, k), k)$ converges to zero as $1/k^{1/2}$. Since, asymptotically, the stochastic difference equation behaves as the ODE, then rates of convergence translate across under the scaling of the theory.

This leads to the convergence rate result of the lemma. □

Remarks:

1) The RPE schemes are mildly more sophisticated than the *ad hoc* one-step-ahead prediction-based ELS schemes of the previous section. For this reason, we have kept the same ELS notation to assist in seeing the similarities and differences.
2) For a RPE version of (3.16), see [8].
3) The (4.3) and (4.4) do not result from standard RPE theory. The search direction has been changed so that P^{-1}_k is diagonal, but the scheme still provides quadratic convergence.
4) A complete and precise theory on convergence rates is not given in the above results because it is beyond the scope of this paper.
5) To reduce the number of calculations, the second half of (4.3) and (4.4) can be replaced by an stochastic approximation given by

$$P^{-1}_k = kE[X_k X'_k] = k\,\mathrm{diag}(E[X_k]).$$

Convergence can still be proven with a slight modification of Lemma 6.

B. A Posteriori Weighted RPE Scheme

To generalize the ELS algorithm based on filtered estimates (3.32)–(3.34) rather than one-step-ahead predictions, we consider RPE schemes based on filtered estimates. To do so, consider the cost function

$$\check{V}_k(\theta) \triangleq \tfrac{1}{2} \sum_{i=2}^{k} \sum_{j=1}^{N} (y_i - \theta' e_j)^2 \hat{X}^{(j)}_{i|i,\hat{\Theta}_{i-1}} \tag{4.10}$$

which gives the update equations given by

$$\boxed{\begin{aligned} \hat{\theta}_k &= \hat{\theta}_{k-1} + \hat{P}_k \kappa_{k|k,\hat{\Theta}_{k-1}} \\ \hat{P}_k^{-1} &= \hat{P}_{k-1}^{-1} + \hat{\mathcal{X}}_{k|k,\hat{\Theta}_{k-1}} \end{aligned}} \tag{4.11}$$

with $\kappa^{(i)}$ the ith element of κ defined from

$$\boxed{\begin{aligned} \kappa^{(i)}_{k|k,\hat{\Theta}_{k-1}} &= -\hat{X}^{(i)}_{k|k,\hat{\Theta}_{k-1}} [y_k - \hat{\theta}^{(i)}_{k-1}] \\ &\quad + \sum_{j=1}^{N} \left. \frac{d\hat{X}^{(j)}_{k|k,\hat{\Theta}_{k-2},\theta}}{d\theta^{(i)}} \right|_{\theta=\hat{\theta}_{k-1}} [y_k - \hat{\theta}^{(j)}_{k-1}]^2 \end{aligned}} \tag{4.12}$$

and

$$\boxed{\hat{X}_{k|k,\hat{\Theta}_k} = N_k B(y_k, \hat{\theta}_k) A \hat{X}_{k-1|k-1,\hat{\Theta}_{k-1}}.} \tag{4.13}$$

Lemma 7: The recursion (4.11) will converge a.s. to the set $\overline{D}c = \{\theta | \lim_{k\to\infty} E[f(\theta, k)] = 0\}$; moreover, under the excitation condition $\hat{P}_k$ (or $\hat{P}_k^{\dagger}$) $\to$ 0 as $1/k$, then convergence of $\hat{\theta}_k$ (or $\hat{\theta}_k^{\dagger}$) is at the rate $1/k^{1/2}$.

Proof: The proof is the same as for Lemma 6 with $f(\theta, k) = E[\kappa_{k|k,\theta}]$ and $G(\theta, k) = E[\hat{\mathcal{X}}_{k|k,\theta}]$ and using the Lyapunov function

$$\hat{W}(\tau, k) = E\left[\sum_{j=1}^{N} (y_k - \theta' e_j)^2 \hat{X}^{(j)}_{k|k,\theta}\right].$$

□

Remarks:

1) The cost function (4.10) is the sum of the predicted error of being in each state, weighted by the estimated probability of being in each state, which from (2.6) is $\hat{X}_{k|k,\theta}$.
2) The similarity of form between (4.11) and (3.32) suggest that the recursions (3.32) are valid at least as approximations for (4.11), for which convergence has been shown.
3) Again, to reduce the number of calculations, the second half of (4.11) can be replaced by an stochastic approximation given by

$$P_k^{-1} = kE[X_k X_k'] = k\,\mathrm{diag}(E[X_k])$$

and the convergence proof holds.

V. Implementation Considerations and Simulations Results

This section has two parts. In the first part, issues concerning implementation of algorithms for estimating θ are presented. The discussion is general in nature and, in fact, applies to any of the algorithms presented in this paper and others in the literature of this field.

In the second part, simulation studies of the various algorithms present in this paper are presented. We attempt to demonstrate the various properties of these algorithms that have been highlighted in the previous sections. The highlighted properties include convergence, convergence rates, bias, and the importance of the issues introduced in the first part of this section.

A. Implementation Considerations

The following were considered when implementing the schemes presented in the preceding chapters.

1) Transients: One reason for studying both ELS and RPE schemes in the same paper is that it appears to be a good approach to use them in combination in an actual implementation. The extra gradient terms used in the RPE schemes do not help during the transient period, where the dominant error is due to initialization rather then the noise; however, these terms do aid convergence subsequently. Thus, it is a reasonable practice to use an ELS scheme initially and change to an RPE scheme once the transient has decayed significantly.

2) Step Sequence: Step-size adjustments can be made for improved transient performance for iterative schemes, and indeed, $\hat{P}_k = (1/k)\hat{R}_k^{-1}$ can be replaced by $\gamma_k \hat{R}_k^{-1}$ for arbitrary γ_k satisfying $\sum_{k=1}^{\infty} \gamma_k = \infty$, $\sum_{k=1}^{\infty} \gamma_k^2 < \infty$, and the ODE analysis still applies. Further details are omitted here; see Ljung [9].

3) Markov State Errors: The time-varying variance of the state estimate, which is given by

$$\begin{aligned} \Sigma_k &= E[(X_k - \hat{X}_{k|k-1,\hat{\Theta}_{k-1}})(X_k - \hat{X}_{k|k-1,\hat{\Theta}_{k-1}})'] \\ &= \hat{\mathcal{X}}_{k|k-1,\hat{\Theta}_{k-1}} - \hat{X}_{k|k-1,\hat{\Theta}_{k-1}} \hat{X}'_{k|k-1,\hat{\Theta}_{k-1}} \geq 0 \end{aligned} \tag{5.1}$$

can be used in the recursive equations to "discount" time instants for which the Markov state is known with less certainty. If the variance of the state estimate is denoted $\sigma_k^{(s)^2} = \hat{\theta}_k' \Sigma_k \hat{\theta}_k$, then the modified update equation, according to standard Kalman filter theory, becomes

$$\hat{P}_k^{-1} = \hat{P}_{k-1}^{-1} + \frac{1}{\sigma_w^2 + \sigma_{k-1}^{(s)^2}} \hat{\mathcal{X}}_{k|k-1,\hat{\Theta}_{k-1}}. \tag{5.2}$$

In addition, in (3.19), (4.3), and (4.4), $\hat{P}_k$ is replaced by $\{1/[\sigma_w^2 + \sigma_{k-1}^{(s)^2}]\}\hat{P}_k$. Corresponding modifications apply to (3.16), (3.17), and (3.31).

4) Parameter Estimation Errors: Similarly, the variance of the parameter estimates approximated by $\hat{P}_k$ can be used to modify the variance used in (2.5) to estimate the Markov states.

$$\sigma_w^{(m)^2} = \sigma_w^2 + \hat{X}_k' \hat{P}_k \hat{X}_k. \tag{5.3}$$

Actually, in practice, it makes sense to limit the magnitude of additive term to σ_w^2 to, say, σ_w^2 because of the approximations involved. That is

$$\sigma_w^{(m)^2} = \sigma_w^2 + \min\{\sigma_w^2, \hat{X}_k' \hat{P}_k \hat{X}_k\}. \tag{5.4}$$

5) Polyak Acceleration: The increased step size and averaging used by Collings [8] is suggested by Polyak [14] as a technique to speed convergence. The Polyak increased step size has been found, in some cases, to aid convergence from poor initial estimates and in high noise.

6) Time-Varying Tracking: It is possible to modify the estimation schemes presented in this paper to allowing tracking of time-varying parameters by introducing a forgetting factor; see Ljung [9]. A forgetting factor λ is introduced by modifying the second equations of (4.3) or (4.11) or the corresponding ELS schemes to give

$$\hat{P}_k^{-1} = \lambda \hat{P}_{k-1}^{-1} + \hat{\mathcal{X}}_{k|k-1,\hat{\Theta}_{k-1}} \tag{5.5}$$

where typically,$\lambda \leq 1$.

This modification was also found to improve convergence in very high noise simulations.

B. Simulations

We present results of simulation examples using computer-generated finite discrete state Markov chains. The results presented in the following simulation plots were found to be representative examples of hundreds of simulation runs. We concentrated our efforts on the new algorithms that appear interesting. For example, the most frequently tested algorithms are (3.16), (3.31), (3.32), and (4.11). The least frequently tested algorithms are (3.6) and (3.19).

1) Convergence Using $\hat{X}_{k|k-1,\theta}$: A two-state Markov chain embedded in white Gaussian noise (WGN) was generated with parameter values $a_{ii} = 0.85$, $a_{ij} = (1 - a_{ii})$ for $i \neq j$, $\theta = [3, 6]'$, $\sigma_w = 0.5$. The state sequence $\hat{X}_{k|k-1,\theta}$ was estimated [i.e., assuming knowledge of θ, so that (3.23) holds], and the state values θ were estimated from $\hat{X}_{k|k-1,\theta}$. Each of the following schemes were examined:

- least squares algorithm (3.6) with X_k known;
- original ELS algorithm (3.16) using $\hat{X}_{k|k-1,\theta}$;
- RPE from [6] using $\hat{X}_{k|k-1,\theta}$;
- *a posteriori* ELS algorithm (3.32) using $\hat{X}_{k|k-1,\theta}$;
- hybrid ELS algorithm (3.31) using $\hat{X}_{k|k-1,\theta}$.

Fig. 1 shows an empirical comparison of the rates of convergence. This figure shows the convergence of various schemes to one of the parameters. We conclude that when using $\hat{X}_{k|k-1,\theta}$ estimates that our $O(N)$ schemes converge at approximately the same rate as the existing $O(N^2)$ scheme.

2) Comparison of $\hat{X}_{k|k-1,\hat{\Theta}_{k-1}}$ with $\hat{X}_{k|k-1,\theta}$: A two-state Markov chain embedded in WGN was generate from which $\hat{X}_{k|k,\hat{\Theta}_{k-1}}$ and $\hat{X}_{k|k-1,\theta}$ were estimated using (3.32). Fig. 2 shows the difference between the estimates over time. We conclude from this simulation that asymptotically, (3.23) holds. Note that the average over 100 points was used in this figure to reduce the amount of information.

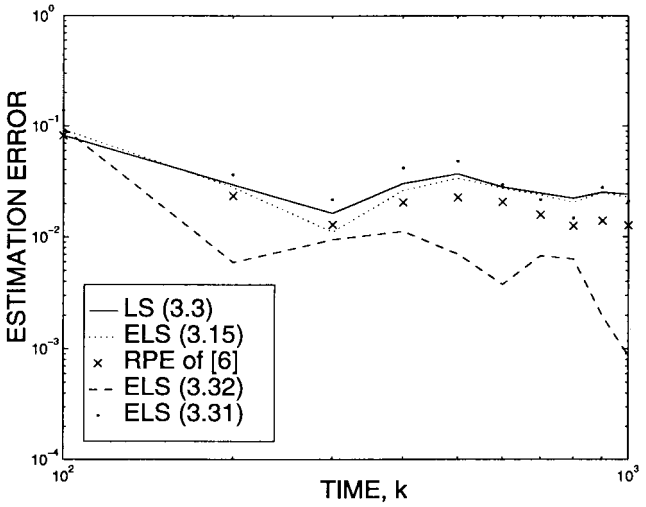

Fig. 1. Empirical comparison of proposed schemes under assumption (3.23).

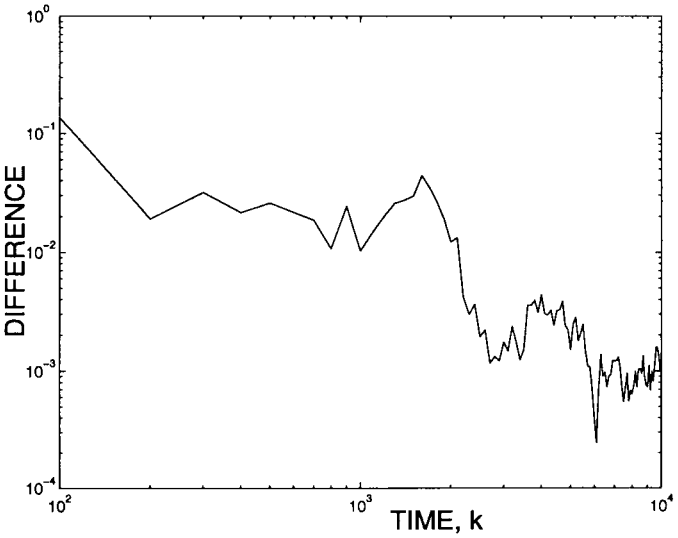

Fig. 2. Convergence to idealized, in the sense of (3.23), state estimates.

3) Bias of Estimation: To verify Remark 2 made in Section III-B, that a bias is indeed introduced by (3.19), a two-state Markov chain embedded in WGN was generated with parameter values $a_{ii} = 0.75$, $a_{ij} = (1 - a_{ii})$ for $i \neq j$, $\theta = [3, 6]'$, and $\sigma_w = 1$. The state values were estimated using (3.16) and (3.19) and noting that the state estimates $\hat{X}_{k|k-1,\theta}$ were used. The estimated parameters from the scheme were $\hat{\theta}_{3.16} \approx [3.01, 5.86]'$ and $\hat{\theta}_{3.19} \approx [4.23, 4.76]'$. The estimates obtained by (3.19) were indeed biased.

4) Convergence Rate Comparison: A two-state Markov chain embedded in WGN was generated with parameter values $a_{ii} = 0.70$, $a_{ij} = (1 - a_{ii})$ for $i \neq j$, $\theta = [2, 4]'$, and $\sigma_w = 1$. The state values were estimated using the following schemes:

- least squares algorithm (3.6) with X_k known;
- original ELS algorithm (3.16) using $\hat{X}_{k|k-1,\hat{\Theta}_{k-1}}$;
- RPE of [8] using $\hat{X}_{k|k-1,\hat{\Theta}_{k-1}}$;
- *a posteriori* ELS algorithm (3.32) using $\hat{X}_{k|k-1,\hat{\Theta}_{k-1}}$;
- RPE algorithm (4.11) using $\hat{X}_{k|k-1,\hat{\Theta}_{k-1}}$.

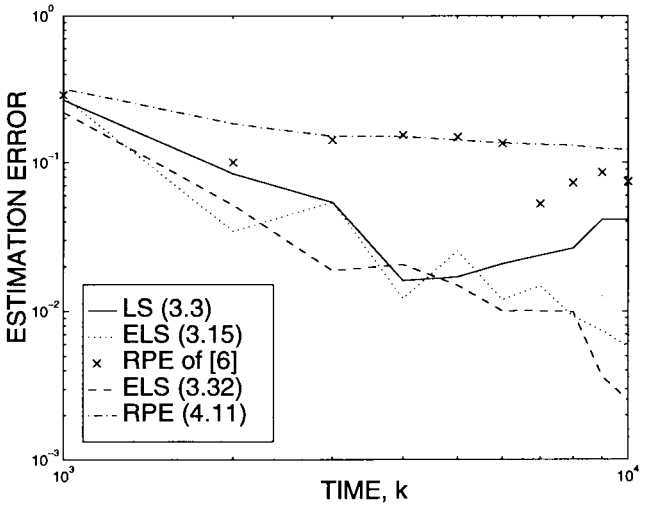

Fig. 3. Empirical comparison of proposed schemes.

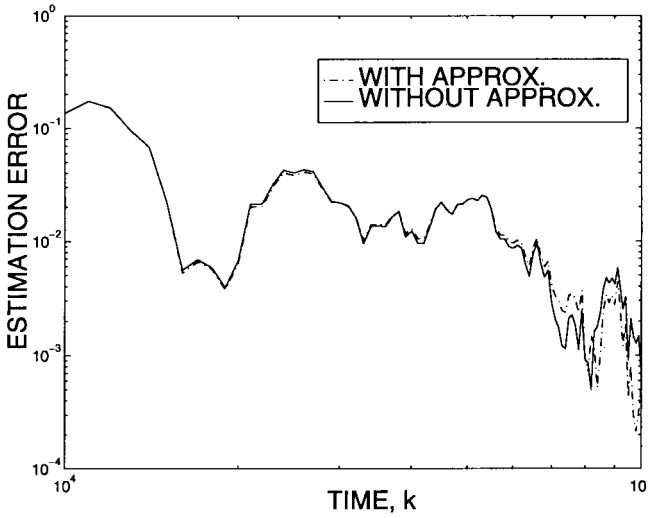

Fig. 5. Effect of stochastic approximation on convergence.

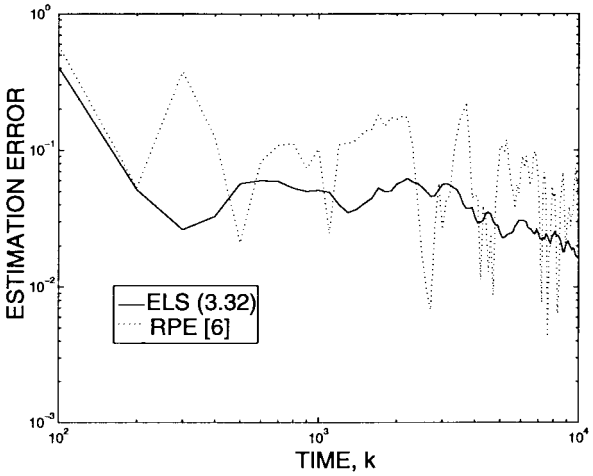

Fig. 4. Empirical comparison of the ELS scheme (3.32) and the RPE scheme of [6].

Fig. 3 shows an empirical comparison of the rates of convergence. This figure shows the convergence of these schemes to one of the parameters. We conclude that our $O(N)$ schemes converge at approximately the same rates.

5) Comparison with Existing the RPE Scheme: A two-state Markov chain embedded in WGN was generated with parameter values $a_{ii} = 0.75$, $a_{ij} = (1-a_{ii})$ for $i \neq j$, $\theta = [2, 4]'$, and $\sigma_w = 1$. The state values were estimated from the chain using (3.32) and the $O(N^2)$ RPE scheme presented in [8]. Fig. 4 shows an empirical comparison of the rates of convergence to one of the state values. This figure shows that similar rates of convergence are achieved by our new schemes with computational requirements of $O(N)$ and the existing RPE scheme [8] with requirements $O(N^2)$.

6) Stochastic Approximation: A two-state Markov chain embedded in WGN was generated with parameter values $a_{ii} = 0.80$, $a_{ij} = (1 - a_{ii})$ for $i \neq j$, $\theta = [2, 4]'$, and $\sigma_w = 1$. The state values were estimated from the chain using (3.32) with and without the approximation described in Remark 5. Fig. 5 shows that convergence of the scheme was not adversely affected by the approximation.

7) Fast Markov Chains: A three-state Markov chain embedded in WGN was generated with parameter values $a_{ii} = 0.65$, $a_{ij} = (1 - a_{ii})/(2)$ for $i \neq j$, $\theta = [0, 1, 2, 3]'$, and $\sigma_w = 1$. The state values were estimated using the recursive schemes [ELSM $\triangleq$ (3.32) recursion] and [RPE1 $\triangleq$ (4.3) recursion]. The ELSM has been found to converge to the correct values, whereas the RPE1 scheme was not. This simulation demonstrates that the RPE1 recursions do not estimate low inertia Markov chains well. In this simulation and others involving fast Markov chains, the ELSM recursions were found to perform better then the RPE1 recursions. Here, $a_{ii} = 0.65$ implies short times in each state.

This is a significant result. The (3.32) recursion is the only recursion we have studied that can handle fast chains effectively. These fast chains are known to appear often in actually applications such as the demodulation of coded QAM signals, which is under study in a companion paper.

8) A Six-State Example: A six-state Markov chain embedded in WGN was generated with parameter values $a_{ii} = 0.95$, $a_{ij} = (1 - a_{ii})/(2)$ for $i \neq j$, $\theta = [1, 2, 3, 4, 5, 6]'$, and $\sigma_w = 1$. The state values were estimated using the RPE1 recursions. Fig. 6 shows the parameter convergence of the RPE1 recursions. Note that in these simulations, the Polyak increased step size is used to allow convergence from poor initial estimates.

9) High Noise: A two-state Markov chain embedded in WGN was generated with parameter values $a_{ii} = 0.9$, $a_{ij} = (1 - a_{ii})$ for $i \neq j$, $\theta = [-1, 1]'$, and $\sigma_w = 40$. The states were estimated using two methods. First, Fig. 7 shows the convergence of parameters using the increased step sequence $1/\sqrt{n}$ and averaging over 1000 points. Second, Fig. 8 shows the convergence of parameters using a scheme modified to track time-varying parameters with a forgetting factor $\lambda = 0.995$. Both figures shows that it is possible to estimate state values in very nigh noise environments.

10) Variance Corrections: From (5.2) and (5.3), we can see that the correction factors are only going to have an effect when not dominated by σ_w^2. These variance correction factors

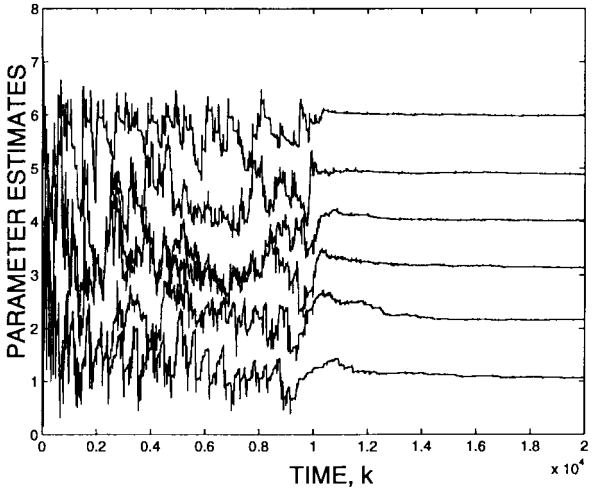

Fig. 6. Parameter convergence of a six-state chain example.

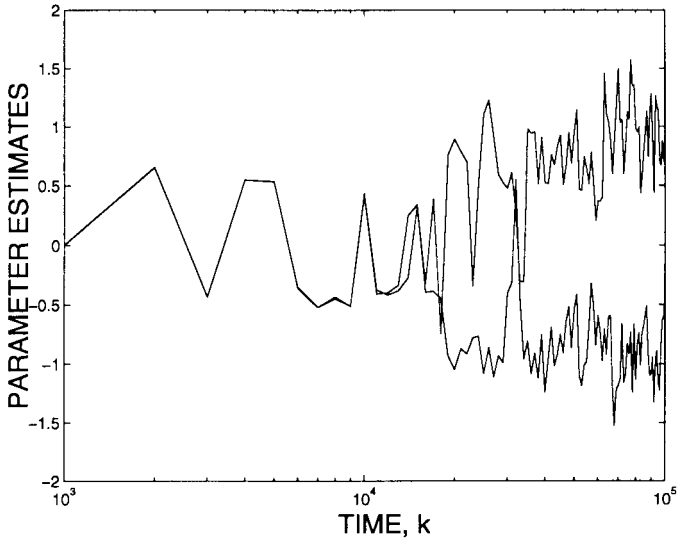

Fig. 7. Parameter convergence in very high noise using Polyak acceleration. Averaging is performed over 1000 points.

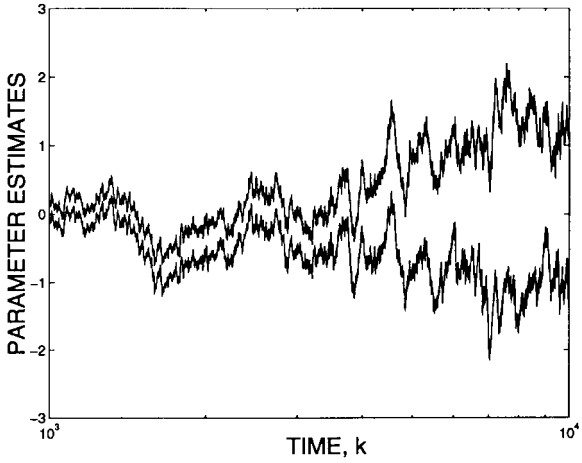

Fig. 8. Parameter convergence in very high noise using forgetting factors.

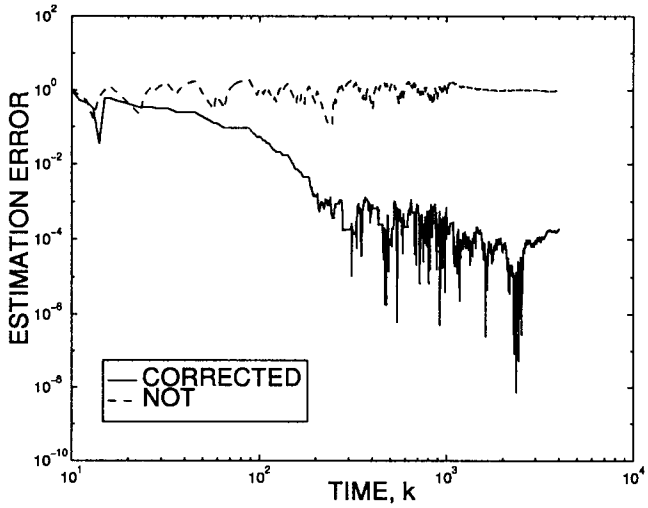

Fig. 9. Effect of variance correction factors.

only improve the performance of these schemes during initial transients or in low noise situations.

For example, a two-state Markov chain embedded in WGN was generated with parameter values $a_{ii} = 0.85$, $a_{ij} = (1 - a_{ii})$ for $i \neq j$, $\theta = [2, 4]'$, and $\sigma_w = 0.000\,01$. The states were estimated using the scheme in (3.32) both with and without the variance correction factors given by (5.2) and (5.3). It has been found that the scheme without correction factors will not converge, whereas the scheme using the correction factors does; see Fig. 9. In low measurement noise situations, the estimation noise dominates the measurements noise, and it must be included to achieve reasonable results. In high measurement noise situations, the relative contribution of the estimation noise is negligible and need not be considered.

C. Summary

A variety of schemes presented in this paper were demonstrated, at least in simulations, to provide competitive convergence performance in comparison with previous work present in [8]. In addition to this, several of the implementation issues raised in the previous subsection were examined.

A highlight of the simulation results was that the (3.32)–(3.34) scheme has been found to provide good convergence performance in situations involving high noise and/or fast Markov chains that exposed the limitations of the previous algorithms [8].

VI. Conclusions

In this paper, we have proposed new on-line parameter estimation schemes for HMM's based on extended least squares and recursive prediction error methods. The transition probabilities between states are assumed known, but the state values between which the noise-free measurements switch are learned in time. These new schemes exploit the idempotent property of the signal model states, noting that care must be taken for the ELS schemes to avoid bias. We present simulation studies of the schemes in a variety of conditions, highlight the similarity and the difference between the performance of these schemes, and compare them with the existing RPE scheme for parameter estimation. The algorithms presented

have computational complexity $O(N)$ yet perform as well asymptotically as earlier schemes proposed of $O(N^2)$.

The *a posteriori* ELS and *a posteriori* weighted RPE scheme, which exploit filtered state estimates rather than prediction estimates, appear to be the most attractive for application purposes. These *a posteriori* schemes have been also found to be consistent and, thus, attractive in signaling environments that include low-inertia HMM's that could not be handled well by earlier algorithms [8].

REFERENCES

[1] J. G. Kemeny and J. L. Snell, *Finite Markov Chains.* Princeton, NJ: Van Nostrand, 1960.

[2] R. J. Elliott, L. Aggoun, and J. B. Moore, *Hidden Markov Models, Estimation and Control.* New York: Springer-Verlag, 1995.

[3] I. B. Collings and J. B. Moore, "Adaptive HMM filters for signals in noisy fading channels," in *Proc. Int. Conf. Acoust., Speech, Signal Processing ICASSP,* Adelaide, Australia, 1994, vol. 3, pp. 305–308.

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

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

[6] S. H. Chung, V. Krishnamurthy, and J. B. Moore, "Adaptive processing techniques based on hidden Markov models for characterising very small channel currents buried in noise and deterministic interferences," *Philo. Trans. R. Soc. London B,* vol. 334, pp. 357–384, 1991.

[7] J. J. Ford and J. B. Moore, "Adaptive estimation of HMM transition probabilities," *IEEE Trans. on Signal Processing,* to be published.

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

[9] L. Ljung and T. Soderstrom, *Theory and Practice of Recursive Identification.* Cambridge, MA: MIT Press, 1983.

[10] L. Ljung, "Analysis of recursive stochastic algorithms," *IEEE Trans. Automat. Contr.,* vol. AC-22, pp. 551–575, Aug. 1977.

[11] J. Neveu, *Discrete Parameter Martingales.* Amsterdam, The Netherlands: North-Holland, 1975.

[12] P. Meyer, *Martingales and Stochastic Integrals–I.* New York: Springer-Verlag, 1972, Lecture Notes in Mathematics Series no. 284.

[13] M. Loeve, *Probability Theory,* 2nd ed. Princeton, NJ: Van Nostrand, 1960.

[14] B. T. Polyak and A. B. Juditsky, "Acceleration of stochastic approximation by averaging," *SIAM J. Contr. Optimization,* vol. 30, no. 4, pp. 838–855, July 1992.

[15] J. Sternby, "On consistency for the method of least squares using Martingale theory," *IEEE Trans. Automat. Contr.,* vol. AC-22, June 1977.

[16] R. Kumar and J. B. Moore, "Convergence of adaptive minimum variance algorithms via weighting coefficient selection," *IEEE Trans. Automat. Contr.,* vol. AC-27, Feb. 1982.

[17] R. K. Boel, J. B. Moore, and S. Dey, "Geometric convergence of filters for hidden Markov models," in *Proc. CDC,* New Orleans, LA, Dec. 1995, pp. 69–74.

[18] T. Soderstrom and P. Stoica, *System Identification.* Englewood Cliffs, NJ: Prentice-Hall, 1988.

Jason J. Ford was born in Canberra, Australia. He received the B.Sc. and B.E. degrees from Australian National University, Canberra, in 1995. He is currently working toward the Ph.D. degree at the Australian National University.

He also worked with the Cooperative Research Centre for Robust and Adaptive Systems at the beginning of 1996. His research interests include system identification, signal processing, and adaptive systems.

John B. Moore (F'79) was born in China in 1941. He received the bachelor and masters degrees in electrical engineering in 1963 and 1964, respectively, and the doctorate degree in electrical engineering from the University of Santa Clara, Santa Clara, CA, in 1967.

He was appointed a Senior Lecturer in the Electrical Engineering Department, University of Newcastle, Callaghan, Australia, in 1967, promoted to Associate Professor in 1968, and promoted to Full Professor (personal chair) in 1973. He was a Department Head from 1975 to 1979. In 1982, he was appointed Professorial Fellow in the Department of Systems Engineering, Australian National University, Canberra, and promoted to Professor in 1990. He has been Head of the Department since 1992. His current research is in control and communication systems and signal processing. He is co-author, with B. Anderson, of three books: *Linear Optimal Control* (Englewood Cliffs, NJ: Prentice-Hall, 1971), *Optimal Filtering* (Englewood Cliffs, NJ: Prentice-Hall, 1979), and *Optimal Control–Linear Quadratic Methods* (Englewood Cliffs, NJ: Prentice-Hall, 1989). He is also co-author of a book with U. Helmke entitled *Optimization and Dynamical Systems* (New York: Springer-Verlag, 1993), with R. Elliott and L. Aggoun entitled *Hidden Markov Model Estimation and Control Via Reference Methods* (New York: Springer-Verlag, 1995), and with T. T. Tay and I. Mareels entitled *High Performance Control* (Boston, MA: Birkhäuser, 1997).

Dr. Moore is a Fellow of the Australian Academy of Technological Sciences and Engineering and of the Australian Academy of Science.