# Adaptive Estimation of HMM Transition Probabilities

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

***Abstract*— This paper presents new schemes for recursive estimation of the state transition probabilities for hidden Markov models (HMM's) via extended least squares (ELS) and recursive state prediction error (RSPE) methods.**

**Local convergence analysis for the proposed RSPE algorithm is shown using the ordinary differential equation (ODE) approach developed for the more familiar recursive output prediction error (RPE) methods. The presented scheme converges and is relatively well conditioned compared with the previously proposed RPE scheme for estimating transition probabilities that perform poorly in low noise.**

**The ELS algorithm presented in this paper is computationally of order $N^2$, which is less than the computational effort of order $N^4$ required to implement the RSPE (and previous RPE) scheme, where $N$ is the number of Markov states.**

**Building on earlier work, an algorithm for simultaneous estimation of the state output mappings and the state transition probabilities that requires less computational effort than earlier schemes is also presented and discussed.**

**Implementation aspects of the proposed algorithms are discussed, and simulation studies are presented to illustrate convergence and convergence rates.**

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

## I. Introduction

HIDDEN Markov models (HMM's) are a powerful tool in the field of signal processing [1], [2] with applications to speech processing [6], digital communication systems [3], [4], and biological signal processing [12]. The major limitations of schemes for estimating HMM parameters in applications concern computational complexity 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 that, without loss of generality, is denoted as $S = \{e_1, e_2, \cdots, e_N\}$, where $N$ is the number of Markov states, and $e_i$ is a vector that is zero everywhere except for the $i$th element, which is 1. There are transitions between states described by fixed probabilities that form a stochastic matrix $A = (A^{ij})$, where $A^{ij}$ is the probability of transferring from state $e_j$ to state $e_i$. The state process is measured indirectly via measurements $y_k$, which are linear functions of the state denoted $CX_k$ in additive noise.

The Baum–Welch, or so called EM algorithm, for off-line estimation of the transition probabilities, given a sequence of observations $y_0, y_1, \cdots, y_T$, is well known and with multiple passes converges locally to maximum likelihood estimates (see [6]). However, this linearly convergent, multipass, forward–backward algorithm has computational effort and memory requirements of O$(N^2T)$ for each pass. Elliott has shown that the backward pass through the data can be eliminated at the expense of increasing the computational effort of the forward pass to being of O$(N^4T)$ (see [2, ch. 2]). One avenue for improving the computational and memory requirements is through the investigation of on-line adaptive schemes, which update parameter estimates at each iteration rather than after each pass through the data.

Recently, on-line identification of HMM's exploiting conventional identification theory has been studied [5], [13]. In [5], an algorithm designed to minimize the Kullback–Leibler information measure is proposed. This algorithm requires computational effort of only O$(N^2)$ per time instant, but convergence is less than asymptotically optimal. Alternatively, the recursive prediction error (RPE) algorithm of [13] seeks to minimize the observation prediction error cost and is asymptotically optimal but requires computational effort of O$(N^4)$ per time instant. The RPE algorithm of [13] appears attractive, due to its asymptotic optimality and its mature theoretical basis; however, it is actually ill conditioned in low noise and is computationally prohibitive for large $N$.

In [11], new algorithms are proposed for estimating the state output mapping $C$, via extended least squares (ELS) and RPE techniques. These algorithms exploit the discrete state structure of HMM's in ways for which there is no parallel in standard state space model identifications. The computational effort of the algorithms presented in [11] is also less than that for the algorithm presented in [13]. In this paper, we exploit and build on the ideas of [11] to produce algorithms for estimating the stochastic matrix $A$ with similar improvements in computational requirements and without computational difficulties as the noise level decreases.

The key contribution of this paper is a new recursive algorithm based on a state prediction error cost function, rather than that based on the output prediction error cost function

Manuscript received July 8, 1996; revised August 6, 1997. This work was supported in part by the Australian Government under the Cooperative Research Centres Program for the activities of the Cooperative Research Centre for Robust and Adaptive Systems. The associate editor coordinating the review of this paper and approving it for publication was Dr. A. Lee Swindlehurst.

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, Australia (e-mail: Jason.Ford@anu.edu.au).

Publisher Item Identifier S 1053-587X(98)03251-6.

used in [13]. The recursive state prediction error (RSPE) algorithm proposed here is shown to minimize the state prediction error cost and has fewer computational requirements than the scheme presented in [13]. An ELS algorithm is also proposed that requires computational effort of only O($N^2$) each time instant, compared with the O($N^4$) required for the RSPE and RPE schemes. Complete ordinary differential equation (ODE) convergence analysis is presented for the RSPE algorithm, but convergence analysis for the proposed ELS algorithm has not been completed. We also show that the proposed RSPE algorithm evanesces to the ELS algorithm and, indeed, to the least squares (LS) algorithm as the signal-to-noise ratio increases.

A second contribution of this paper is a scheme that allows simultaneous estimation of the state output mappings $C$ and the state transition probability matrix $A$. The proposed scheme requires less computational effort than the simultaneous estimation scheme presented in [13] but still requires O($N^4$) calculations per time instant.

This paper is organized as follows. In Section II, the signal model, conditional state estimates, and a parameterized information state model are introduced. In Section III, we initially focus on a simplified estimation problem, namely, when the state sequence is measured directly, and apply the familiar least squares approach. Some convergence results are presented. When the state sequence is not measured directly, the least squares approach leads to the proposal of an ELS algorithm. We then generalize the ELS algorithm by introducing a RSPE scheme and demonstrate convergence via ODE analysis. In Section IV, an algorithm for the simultaneous estimation of transition probabilities $A$ and state output mappings $C$ is presented. In Section V, some simulation studies that show relative performance of these algorithms are presented. Finally, conclusions are presented in Section VI.

## II. Problem Formulation

In this section, we introduce the HMM in state space form. Conditional state estimates and a parameterized information state model are also introduced.

### 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 $S = \{e_1, e_2, \cdots, e_N\}$, $e_i = (0, \cdots, 0, 1, 0, \cdots, 0)' \in \mathbb{R}^N$ with 1 in the $i$th position. We consider this process to be defined on the probability space $(\Omega, \mathcal{F}, P)$, with $\mathcal{F}_k^0 = \sigma\{X_0, \cdots, X_k\}$ and with complete filtration $\{\mathcal{F}_k\}$. The state space model is then defined, for $k \geq 0$, by

$$X_{k+1} = AX_k + M_{k+1} \tag{2.1}$$

$$y_k = CX_k + w_k \tag{2.2}$$

where $M_{k+1}$ is a sequence of $\mathcal{F}_k$ martingales; hence, $E[M_{k+1}|\mathcal{F}_k] = 0$. In addition, the $y_k$ are continuous valued belonging to $\mathbb{R}$ (although generalization to $\mathbb{R}^N$ is straightforward), and $w_k \in \mathbb{R}$ is i.i.d. with zero mean and of known density, such as when $w_k$ is Gaussian, i.e., $w_k \sim N[0, \sigma_w^2]$, or a mixture of Gaussians. In addition, $C \in \mathbb{R}^{1\times N}$ is a vector of state values termed the *state output mappings* of the Markov chain. The term *state levels* is commonly used for the vector $C$ when the observations are scalar. We also define $Y_k \triangleq (y_0, \cdots, y_k)$ and $\mathcal{Y}_k$ as the complete filtration generated by $y_\ell$, $\ell \leq k$. As a consequence

$$E[w_{k+1}|\mathcal{F}_k \vee \mathcal{Y}_k] = 0. \tag{2.3}$$

Due to the Markov nature of $X_k$, we can write

$$E[X_{k+1}|\mathcal{F}_k] = E[X_{k+1}|X_k] = AX_k$$

where $A = (A^{ij})$ and $A^{ij} \triangleq P(X_{k+1} = e_i|X_k = e_j)$. Obviously, $A^{ij} \geq 0$, and $\sum_{i=1}^N A^{ij} = 1$ for all $j$. We also assume that $X_0$ or its distribution is known.

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 when $w_k \sim N[0, \sigma_w^2]$, we can write

$$b_k(i) = \frac{1}{\sqrt{2\pi\sigma_w^2}} \exp\left[\frac{-(y_k - Ce_i)^2}{2\sigma_w^2}\right]. \tag{2.4}$$

We also write the initial state probability vector for the Markov chain $\pi = (\pi_i)$ with $\pi_i \triangleq P(X_0 = e_i)$. The HMM is denoted $\lambda = (A, C, \pi, \sigma_w^2)$.

### B. Conditional State Estimates and Information State Model

Let $\hat{X}_{k|k, A}$ denote the conditional filtered state estimate of $X_k$, given measurements $Y_k$ and $A$. In addition, let $\hat{X}_{k|k-1, A}$ denote the one-step-ahead prediction of $X_k$, given measurements $Y_{k-1}$ and $A$. That is

$$\hat{X}_{k|k, A} \triangleq E[X_k|\mathcal{Y}_k, A], \; \hat{X}_{k|k-1, A} \triangleq E[X_k|\mathcal{Y}_{k-1}, A]. \tag{2.5}$$

The forward recursion for obtaining conditional filtered state estimates $\hat{X}_{k|k, A}$ for an HMM is given in [2]

$$\hat{X}_{k|k, A} = N_k(y_k, A)\mathbf{B}(y_k)\mathbf{A}\hat{X}_{k-1|k-1, A} \tag{2.6}$$

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

We now proceed to introduce an information state model. An *information state* tells us all the information we know about the state from the observations and is here simply the state estimate $\hat{X}_{k|k, A}$. Consider the following lemmas.

*Lemma 1:* The one-step-ahead predictions $E[X_k|\mathcal{Y}_{k-1}, A]$ are given by

$$\hat{X}_{k|k-1, A} = A\hat{X}_{k-1|k-1, A}.$$

*Proof:*

$$\begin{aligned} \hat{X}_{k|k-1, A} &= E[X_k|\mathcal{Y}_{k-1}, A] = E[AX_{k-1} + M_k|\mathcal{Y}_{k-1}, A] \\ &= A\hat{X}_{k-1|k-1, A}. \end{aligned}$$ □

Hence, the following lemma now holds.

*Lemma 2:* The error term $(\hat{X}_{k+1|k+1,A} - A\hat{X}_{k|k,A})$ is orthogonal to $\hat{X}_{k|k,A}$, and the error term $(y_k - C\hat{X}_{k|k,A})$ is orthogonal to $\hat{X}_{k|k,A}$. Moreover, the HMM can be reorganized as an information state model; see [22, p. 79]. The state estimates can be written as

$$\hat{X}_{k+1|k+1,A} = A\hat{X}_{k|k,A} + n_k \tag{2.7}$$

$$y_k = C\hat{X}_{k|k,A} + v_k \tag{2.8}$$

where $n_k$ and $v_k$ are orthogonal to $\hat{X}_{k|k,A}$.

*Proof:* From (2.1), we see that

$$\begin{aligned}\hat{X}_{k+1|k+1,A} &- A\hat{X}_{k|k,A}\\ &= A(X_k - \hat{X}_{k|k,A}) - (X_{k+1} - \hat{X}_{k+1|k+1,A}) + M_{k+1}.\end{aligned}$$

We next show that each term on the right is orthogonal to $\hat{X}_{k|k,A}$.

From optimality, the estimation error $(X_k - \hat{X}_{k|k,A})$ is orthogonal to $\mathcal{Y}_k$, and since $\hat{X}_{k|k,A} \subset \mathcal{Y}_k$, then $\hat{X}_{k|k,A}$ is orthogonal to $(X_k - \hat{X}_{k|k,A})$. Similarly, $(X_{k+1} - \hat{X}_{k+1|k+1,A})$ is orthogonal to $\mathcal{Y}_{k+1}$ and because $\hat{X}_{k|k,A} \subset \mathcal{Y}_k \subset \mathcal{Y}_{k+1}$, then $\hat{X}_{k|k,A}$ is orthogonal to $(X_{k+1} - \hat{X}_{k+1|k+1,A})$. Finally, $\hat{X}_{k|k,A}$ is orthogonal to $M_{k+1}$ from (2.2) and because $\hat{X}_{k|k,A}$ is orthogonal to $\mathcal{Y}_k$. The result (2.7) follows, and (2.8) follows likewise by noting that $(y_k - C\hat{X}_{k|k,A}) = C(X_k - \hat{X}_{k|k,A}) + w_k$. □

Lemma 2 shows that the orthogonality property required for convergence of standard recursive identification is satisfied; see [14].

## III. Estimation of Transition Probabilities

In this section, we develop algorithms for estimating the HMM transition probability matrix $A$ from observations $Y_k$. Initially, we investigate the simplified problem of estimating $A$ from a known state sequence $\{X_k\}$ using a least squares (LS) algorithm. In the following subsection, we use conditional state estimates in an extended least squares (ELS) algorithm to produce estimates of $A$ when the state sequence is not measured directly. Finally, we introduce a state prediction error cost and propose a RSPE algorithm.

### A. Least Squares

In this subsection, we consider the signal model (2.1) and (2.2) and the simplified estimation problem. Estimate the state transition probability matrix $A$ from the state sequence $X_1, X_2, \cdots, X_k$. Subsequently, we will consider the more difficult estimation problem where the state sequence $X_1, X_2, \cdots, X_k$ must be estimated from $\mathcal{Y}_k$.

*Lemma 3:* Once each state has been active at least once, that is $(\sum_{k=1}^m X_k X_k')^{-1}$ exists, the optimal off-line least squares estimate of the transition probability matrix $A$, given $X_1, X_2, \cdots, X_m$, is

$$\hat{A}_m \triangleq \left(\sum_{k=1}^m X_{k+1}X_k'\right)\left(\sum_{k=1}^m X_k X_k'\right)^{-1}. \tag{3.1}$$

Moreover

$$\lim_{m\to\infty} \hat{A}_m \quad \text{exists a.s.} \tag{3.2}$$

*Furthermore, under the excitation condition assumption* $\lim_{m\to\infty}(\sum_{i=1}^m X_i X_i')^{-1} = 0$, then

$$\lim_{m\to\infty} \hat{A}_m = A \quad \text{a.s.} \tag{3.2}$$

*Proof:* Standard least squares algorithms are concerned with minimization with respect to $A$ of the following cost:

$$\sum_{k=1}^m \|X_{k+1} - AX_k\|^2. \tag{3.3}$$

Standard manipulations give (3.1). Now, since $(\sum_{k=1}^m X_k X_k') = [\sum_{k=1}^m \text{diag}(X_k)]$, where $\text{diag}(X)$ is the diagonal matrix with $X$ on its diagonal when $X$ is a vector, then

$$\hat{A}_m^{ij} = \sum_{k=1}^m X_{k+1}^{(i)} X_k^{(j)} \left[\sum_{k=1}^m X_k^{(j)}\right]^{-1}.$$

In addition, since $X_k^{(j)} \in \{0, 1\}$, then on the subsequence of $[1, m]$ for which $X_k^{(j)} = 1$, which is denoted $\{\ell_j(1), \ell_j(2), \cdots, \ell_j(\overline{m}_j)\}$ with $\overline{m}_j$ integers (where $\overline{m}_j \triangleq \sum_{k=1}^m X_k^{(j)}$), then

$$\hat{A}_m^{ij} = \overline{m}_j^{-1} \sum_{k=1}^{\overline{m}_j} X_{\ell_j(k+1)}^{(i)} \le 1,$$

where

$$\overline{m}_j \ge 1, \sum_{j=1}^N \overline{m}_j = m.$$

First, we prove the second lemma result (3.2), where the excitation condition that $\lim_{m\to\infty} \overline{m}_j^{-1} = 0$ for all $j$ holds. Consider the error term, which follows from algebraic manipulation of (3.1) and (2.1)

$$\hat{A}_m^{ij} - A^{ij} = \overline{m}_j^{-1}\left[\sum_{k=1}^{\overline{m}_j} M_{\ell_j(k+1)}^{(i)}\right].$$

Now, we define $W_m^{(i,j)} \triangleq \sum_{k=1}^{\overline{m}_j} 1/k M_{\ell_j(k+1)}^{(i)}$, whose elements are scalar martingales adapted to $\mathcal{F}_k$ since $E[W_{m+1}^{(i,j)}|\mathcal{F}_m] = W_m^{(i,j)}$ for all $i, j$. In addition, $W_m^{(i,j)}$

is bounded in $L_2$ for each $i, j$ since

$$\begin{aligned} &E\{[W_m^{(i,j)}]^2\} \\ &= E\left\{\left[\sum_{n=1}^{\overline{m}_j} \frac{1}{n} M^{(i)}_{\ell_j(n+1)}\right]\left[\sum_{k=1}^{\overline{m}_j} \frac{1}{k} M^{(i)}_{\ell_j(k+1)}\right]\right\} \\ &= E\left(E\left\{\sum_{n=1}^{\overline{m}_j}\sum_{k=1}^{\overline{m}_j} \frac{1}{n} M^{(i)}_{\ell_j(i+1)} \frac{1}{k} M^{(i)}_{\ell_j(k+1)} \middle| \mathcal{F}_{\max[\ell_j(n),\, \ell_j(k)]}\right\}\right) \\ &= E\left(\sum_{k=1}^{\overline{m}_j} \frac{1}{k^2} E\{[M^{(i)}_{\ell_j(k+1)}]^2 | \mathcal{F}_{\ell_j(k)}\}\right) \\ &\le B_\infty \sum_{k=1}^{\overline{m}_j} \frac{1}{k^2} \\ &< \infty \qquad \text{for all } i, j. \end{aligned}$$

Here, we have used that $E[M^{(i)}_{k+1} M^{(i)}_n | \mathcal{F}_k] = 0$ for all $n \le k$ and $E\{[M^{(i)}_{k+1}]^2 | \mathcal{F}_k\} \le B_\infty$ for some $B_\infty < \infty$.

Now, under the excitation condition $\overline{m}_j \to \infty$ for all $j$ and martingale convergence results [7], [8], we have that $W_m^{(i,j)}$ converges almost surely. Hence, by the Kronecker Lemma [8], [9], we have that

$$\lim_{m\to\infty} (\hat{A}^{ij}_m - A^{ij}) = \lim_{\overline{m}_j\to\infty} \frac{1}{\overline{m}_j} \sum_{k=1}^{\overline{m}_j} M^{(i)}_{\ell_j(k+1)} = 0 \quad \text{a.s. for all } i, j$$

and the lemma result (3.2) follows.

To obtain the first lemma result, we note that if $\lim_{m\to\infty} \overline{m}_j$ is finite, then $\sum_{k=1}^{\overline{m}_j} X^{(i)}_{\ell_j(k+1)}$ is also finite, and hence, clearly, $\lim_{m\to\infty} \hat{A}_m$ is finite. The existence of $\lim_{m\to\infty} \hat{A}_m$ when $\lim_{m\to\infty} \overline{m}_j^{-1} = 0$ is proven by the second lemma result; hence, the first lemma results follow as claimed. □

Consider now on-line estimation via recursive least squares (RLS) algorithms. Simple manipulations of (3.1) give the on-line recursions

$$\begin{aligned} \hat{A}_{k+1} &= \hat{A}_k + (X_{k+1} - \hat{A}_k X_k) X_k' P_k \\ P_k^{-1} &= P_{k-1}^{-1} + X_k X_k', \quad \text{or} \\ P_k &= P_{k-1} - P_{k-1} X_k (1 + X_k' P_{k-1} X_k)^{-1} X_k' P_{k-1} \end{aligned} \tag{3.4}$$

where $P_k$ can be thought of as related to the energy of the input sequence.

The indicator vectors $X_k$ have the property that nonlinear functions of an indicator vector $F(X_k)$ are linear functions $[F(e_1), \cdots, F(e_N)] X_k$ of the indicator vector $X_k$. Exploiting this property, it is possible to rewrite (3.4) so that the right-hand sides are linear in $X_k$.

We now proceed to consider the more realistic case when $X_k$ is not measured directly but must be estimated from observations. We first examine extended least squares (ELS) algorithms.

### *B. Extended Least Squares*

This subsection proposes an ELS algorithm for estimating HMM transition probabilities. Extended least squares algorithms are *ad hoc* algorithms in which conditional state estimates are used in lieu of actual states $X_k$ in an LS implementation; see [23] for more details.

Consider the ELS version of the LS recursion (3.4) obtained by replacing the state $X_k$ by conditional state estimates, that is

$$\begin{aligned} \hat{A}_{k+1} &= \hat{A}_k + (\hat{X}_{k+1|k+1,\, \hat{\mathcal{A}}_k} - \hat{A}_k \hat{X}_{k|k,\, \hat{\mathcal{A}}_{k-1}}) \hat{X}'_{k|k,\, \hat{\mathcal{A}}_{k-1}} \overline{P}_k \\ \overline{P}_k^{-1} &= \overline{P}_{k-1}^{-1} + \text{diag}(\hat{X}_{k|k,\, \hat{\mathcal{A}}_{k-1}}) \end{aligned} \tag{3.5}$$

where $\hat{\mathcal{A}}_k = [\hat{A}_0, \cdots, \hat{A}_k]$, diag$(X)$ is the diagonal matrix with $X$ on its diagonal when $X$ is a vector, and the recursion below is used to generate state estimates $\hat{X}_{k|k,\, \hat{\mathcal{A}}_{k-1}}$

$$\hat{X}_{k+1|k+1,\, \hat{\mathcal{A}}_k} = N_{k+1}(y_k, \hat{A}_k) B(y_{k+1}) \hat{A}_k \hat{X}_{k|k,\, \hat{\mathcal{A}}_{k-1}} \tag{3.6}$$

where $N_{k+1}(y_k, \hat{A}_k)$ is a scalar normalization factor as in (2.6).

*Remarks:*

1) Note that $(\hat{X}_{k+1|k+1,\, \hat{\mathcal{A}}_k} - \hat{A}_k \hat{X}_{k|k,\, \hat{\mathcal{A}}_{k-1}})$ is not orthogonal to $\hat{X}_{k|k,\, \hat{\mathcal{A}}_{k-1}}$ unless $\hat{A}_k = A$ for all $k$. Hence, standard theory no longer applies.
2) The computational cost of the ELS recursion (3.5) at each iteration is O$(N^2)$.

Since we are unable to proceed with further analysis of the convergence properties of this ELS algorithm, we proceed in the next subsection by taking the ELS concepts one step further. The RSPE algorithm that follows appears to naturally generalize this ELS algorithm. These RSPE algorithms are developed with the view of achieving asymptotic efficient convergence (in the sense of almost surely to a local minimum of the appropriate cost function) with rate of order $1/k^{1/2}$.

### *C. RSPE Method*

There exists mature theory for recursive identification of discrete-time models with states in $\mathbb{R}^N$ based on the minimization of the observation prediction error cost; see [14]. This RPE theory provides asymptotic quadratic convergent algorithms (admittedly to a local minimum) for linear and certain nonlinear models.

In this section, we proceed by applying this theory to obtain asymptotic convergent algorithms (in a local sense) for HMM identification that generalize the ELS scheme of the previous subsection.

Lemma 2 motivates the use of a state prediction error cost [see (3.3)], rather than the observation prediction error cost that is used in the standard RPE theory. Consider the cost function

$$V_k(\theta) \triangleq \frac{1}{2} \sum_{i=2}^{k} \|\hat{X}_{i|i,\,\theta} - A(\theta) \hat{X}_{i-1|i-1,\,\theta}\|^2 \tag{3.7}$$

where $\theta$ is used to parameterize the unknown transition probability matrix such that $\theta = [\mathbf{a_1}, \cdots, \mathbf{a_N}]'$, where $\mathbf{a_i} \triangleq [A^{i1}, \cdots, A^{iN}]$.

Thus, the RSPE recursions that seek to minimize the cost (3.7) are

$$\begin{aligned}\hat{\theta}_k &= \hat{\theta}_{k-1} + \hat{P}_k \kappa_{k|\hat{\theta}_{k-1}} \\ \hat{P}_k^{-1} &= \hat{P}_{k-1}^{-1} + \mathrm{diag}(\underline{1}' \otimes \hat{X}_{k-1|k-1,\, \hat{A}_{k-2}}) \\ \hat{P}_0^{-1} &= \Delta I \end{aligned} \tag{3.8}$$

where $\kappa_{k|\hat{\theta}_{k-1}} = (\kappa^{(i)}_{k|\hat{\theta}_{k-1}})$ for $\kappa^{(i)}_{k|\hat{\theta}_{k-1}} = d/d\theta^{(i)} V_k(\theta)|_{\theta=\hat{\theta}_{k-1}}$, and

$\underline{1}$ column vector of all ones;
$\otimes$ Kronecker product;
$\Delta$ some large constant.

Here, $\hat{P}_{k-1}^{-1}$ is an approximation to the second derivative of $V_k(\theta)$. Note that a projection operation can be implemented at each time step to ensure that $A(\hat{\theta}_k)$ is a valid stochastic matrix, and the convergence results presented in the following discussion still hold.

The recursion (3.8) can also be written as the scalar recursion

$$\begin{aligned}\hat{\theta}_k^{(i)} &= \hat{\theta}_{k-1}^{(i)} + \hat{P}_k^{(i)} \kappa^{(i)}_{k|\hat{\theta}_{k-1}} \\ [\hat{P}_k^{(i)}]^{-1} &= [\hat{P}_{k-1}^{(i)}]^{-1} + \hat{X}^{(\xi)}_{k-1|k-1,\, \hat{A}_{k-2}} \end{aligned} \tag{3.9}$$

where $\xi = i\ modn\ N$. Here, $modn$ is the usual modulo operation, except that $i\ modn\ i = i$.

*Gradient Calculations:*

$$\kappa^{(i)}_{k|\hat{\theta}_{k-1}} = \left.\frac{dV_k(\theta)}{dA^{mn}}\right|_{\theta=\hat{\theta}_{k-1}}$$

where $m = i\ modn\ N$, and $n = (i - m)/N + 1$. Now

$$\begin{aligned}\frac{dV_k(\theta)}{dA^{mn}} = &- \left[\hat{X}^{(m)}_{k|k,\, \theta} - \mathbf{a_m}(\theta)\hat{X}_{k-1|k-1,\, \theta}\right]\hat{X}^{(n)}_{k-1|k-1,\, \theta} \\ &+ \sum_{j=1}^{N} \left\{ [\hat{X}^{(j)}_{k|k,\, \theta} - \mathbf{a_j}(\theta)\hat{X}_{k-1|k-1,\, \theta}] \right. \\ &\left. \times \left[\frac{d\hat{X}^{(j)}_{k|k,\theta}}{dA^{mn}} - \mathbf{a_j}(\theta)\,\frac{d\hat{X}_{k-1|k-1,\theta}}{dA^{mn}}\right]\right\}. \end{aligned} \tag{3.10}$$

Here, $\mathbf{a_i}$ is defined in Section II-A, and

$$\begin{aligned}\frac{d\hat{X}^{(j)}_{k|k,\theta}}{dA^{mn}} = &N_k^2 \mathrm{diag}(B)\left[\hat{X}^{(n)}_{k-1|k-1,\theta}\, e_m + \mathbf{A}(\theta)\,\frac{d\hat{X}_{k-1|k-1,\theta}}{dA^{mn}}\right] \\ &\times B(j,\, j)\mathbf{a_j}(\theta)\hat{X}_{k-1|k-1,\theta} \\ &+ N_k B(j,\, j)\mathbf{a_j}(\theta)\,\frac{d\hat{X}_{k-1|k-1,\theta}}{dA^{mn}} \qquad \text{if } j \neq m \\ \frac{d\hat{X}^{(j)}_{k|k,\theta}}{dA^{mn}} = &N_k^2 \mathrm{diag}(B)\left[\hat{X}^{(n)}_{k-1|k-1,\theta}\, e_m + \mathbf{A}(\theta)\,\frac{d\hat{X}_{k-1|k-1,\theta}}{dA^{mn}}\right] \\ &\times B(j,\, j)\mathbf{a_j}(\theta)\hat{X}_{k-1|k-1,\theta} \\ &+ N_k B(j,\, j)\left[\hat{X}^{(n)}_{k-1|k-1,\theta} + \mathbf{a_j}(\theta)\,\frac{d\hat{X}_{k-1|k-1,\theta}}{dA^{mn}}\right] \\ &\text{if } j = m. \end{aligned} \tag{3.11}$$

*Convergence Proof:* Convergence of (3.8) and (3.9) is shown by considering the ordinary differential equation (ODE) associated with (3.8) and (3.9). That is

$$\begin{aligned}\frac{d}{d\tau}\,\theta(\tau,\, k) &= R^{-1}(\tau,\, k) f[\theta(\tau,\, k),\, k] \\ \frac{d}{d\tau}\, R(\tau,\, k) &= G[\theta(\tau,\, k),\, k],\ R(0,\, k) \geq \delta I. \end{aligned} \tag{3.12}$$

Here, $k$ is fixed, and $\delta$ is a small constant. Let us define for (3.8) and (3.9) with $\theta(\tau,\, k)$ abbreviated as $\theta_\tau$

$$f(\theta_\tau,\, k) = E[\kappa_{k|\theta_\tau}] \tag{3.13}$$

and

$$G(\theta_\tau,\, k) = E[\mathrm{diag}(\underline{1}' \otimes \hat{X}_{k-1|k-1,\, \theta_\tau})]. \tag{3.14}$$

The following lemma now holds.

*Lemma 4:* The recursions (3.8) and (3.9) will converge a.s. to the set $\overline{D} = \{\theta_\infty | \lim_{k\to\infty} E[f(\theta_\infty,\, k)] = 0\} \supset \theta$ (or possibly the boundary of the valid A region if a projection step is performed). Moreover, under the excitation condition $\hat{P}_k \to 0$ as $1/k$, then convergence of $\hat{\theta}_k$ is at the rate $1/k^{1/2}$.

*Proof:* The ODE associated with (3.8) and (3.9) for fixed $k$, under (3.13) and (3.14), is (3.12).

Now, a Lyapunov function for (3.12) under (3.13) and (3.14) is

$$\hat{W}(\theta_\tau,\, k) = E[\|\hat{X}_{k|k,\, \theta_\tau} - A(\theta_\tau)\hat{X}_{k-1|k-1,\, \theta_\tau}\|^2] \tag{3.15}$$

so that

$$\begin{aligned}\frac{d}{d\tau}\,\hat{W}(\theta_\tau,\, k) &= \frac{d\hat{W}(\theta_\tau, k)}{d\theta_\tau}\,\frac{d\theta_\tau}{d\tau} \\ &= - f'[\theta(\tau,\, k),\, k] R(\tau,\, k)^{-1} f[\theta(\tau,\, k),\, k]. \end{aligned} \tag{3.16}$$

Thus, $\hat{W}(\theta_\tau,\, k)$ converges for all $k$ and $\tau \to \infty$, and $\theta(\tau,\, k)$ converges to the set $\{\theta | E[f(\theta,\, k)] = 0\}$ (for discussion of convergence when a projection is performed, see Ljung [14]).

Here, the recursions (3.8) and (3.9) and intermediate steps are stable; hence, together with the results of [18]–[20], the various regularity conditions required by the ODE theory of Ljung [14] are satisfied, and the first result claimed follows. Note that the conditions given in [18]–[20] ensure that HMM filters forget initial conditions exponentially.

Observe from (3.16) 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.

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

*Remarks:*

1) The theory is not a global convergence theory. It is not excluded that the set $\overline{D}$ may contain locally optimal, but not globally optimal, parameterizations to which the recursions can converge. Simulation studies suggest that with reasonable initializations, $\hat{\theta}_k$ converges to $\theta$, as desired.

2) The lemma excitation condition $\hat{P}_k \rightarrow 0$ as $1/k$ is not particularly restrictive. It can be interpreted as an ergodicity requirement on the state sequence. That is, the Markov state sequence must visit each state (uniformly) as $k \rightarrow \infty$.
3) The existence of parameter estimates and/or convergence of these estimates (possibly only for a subset of the parameters) can be shown when the lemma excitation condition is relaxed, but this is not done here.
4) To reduce the number of calculations, the second half of (3.8) and (3.9) can be replaced by a stochastic approximation given by
$$\hat{P}_k^{-1} \sim k\text{diag}(E[X_k]).$$
Convergence can still be proven with a slight modification of Lemma 4.
5) The concept of using a cost function (3.7) that measures the state prediction error has been introduced previously in other contexts by Bryson; see [10, p. 349]. However, we believe this concept has not been used previously for HMM identification.
6) The state prediction error cannot be driven to zero for all $k$ by a particular choice of $\theta$ due to the nature of Markov sequences; however, the expected value of the error will tend to zero as $\hat{\theta}_k \rightarrow \theta$.
7) The number of calculations required to estimate $\theta$ in (3.8) and (3.9) is of O($N^4$).

In [13], the observation prediction cost function is used to identify transition probabilities, that is

$$\hat{\theta} = \arg\min_{\theta} \{\overline{V}_k(\theta) = E[(y_k - CX_k)^2 | Y_{k-1}]\}.$$

To understand the difficulty in using this type of cost function to estimate the transition probabilities of an HMM, consider the following lemma.

*Lemma 5:* As the measurement noise approaches zero in variance, that is, $\sigma_w^2 \rightarrow 0$, then

$$\frac{d\hat{X}^{(j)}_{k|k,\hat{A}_{k-1}}}{dA^{mn}} \rightarrow 0.$$

*Proof:* From (2.6) we see that

$$\hat{X}^{(j)}_{k+1|k+1,\, \hat{A}_k} = N_k b_k(j) \mathbf{a_j} \hat{X}_{k|k,\, \hat{A}_{k-1}}$$

where $N_k = [\sum_{j=1}^N b_k(j) \mathbf{a_j} \hat{X}_{k|k,\, \hat{A}_{k-1}}]^{-1}$, and $b_k(i)$ is defined in (2.4).

As $\sigma_w^2 \rightarrow 0$, then $b_k(i) \rightarrow 0$ for all $i$ that $X_k \neq e_i$ and $b_k(i) \neq 0$ for the $i$ that $X_k = e_i$. Hence, $N_k \rightarrow b_k(i) \mathbf{a_i} \hat{X}_{k|k,\, \hat{A}_{k-1}}$ a.s. for the $i$ that $X_k = e_i$.

Therefore, $\hat{X}^{(i)}_{k+1|k+1,\, \hat{A}_k} \rightarrow 0$ for all $i$ that $X_k \neq e_i$, and $\hat{X}^{(i)}_{k+1|k+1,\, \hat{A}_k} \rightarrow 1$ for the $i$ that $X_k = e_i$, i.e., $\hat{X}_{k+1|k+1,\, \hat{A}_k} \rightarrow X_{k+1}$ The lemma result follows. □

Lemma 5 implies that $d/d\theta \overline{V}_k(\theta) \rightarrow 0$ as $\sigma_w^2 \rightarrow 0$. That is, as $\sigma_w^2 \rightarrow 0$, the cost function $\overline{V}_k(\theta)$ becomes invariant of $\theta$. Hence, it is clear that $\overline{V}_k(\theta)$ is not a good criterion for identifying $\theta$. Lemma 5 correctly predicts that the performance of the RPE algorithm presented in [13] will deteriorate as $\sigma_w^2 \rightarrow 0$.

Our choice of cost function (3.7) does not suffer from the same difficulties as $\sigma_w^2 \rightarrow 0$. In fact, from (3.10), it is clear that as $\sigma_w^2 \rightarrow 0$, the RSPE algorithm reduces to the ELS algorithm (3.5). Similarly, as $\sigma_w^2 \rightarrow 0$, then $\hat{X}_{k+1|k+1,\, \hat{A}_k} \rightarrow X_k$, and hence, the ELS algorithm, and, likewise, the RSPE algorithm, simplifies to the LS algorithm (3.4).

*Remark:*

1) Even without $\sigma_w^2 \rightarrow 0$, it is possible to see the similarities between the ELS recursion (3.5) and the RSPE recursion (3.8). In fact, if we were to approximate the gradient $\kappa_{k|\hat{\theta}_{k-1}}$ by the first term in (3.10), then the RSPE recursions would reduce to the ELS recursions (3.5).

## IV. Estimation of Transition Probabilities and State Output Mappings

This section proposes an algorithm for simultaneous estimation of the state output mapping matrix $C$ and the transition probability matrix $A$, given a set of observations $Y_k$ and knowledge of the measurement noise variance $\sigma_w^2$. Local convergence results are presented. Stronger convergence results are neither shown nor excluded from our theory.

### A. Dual Cost Function Approach

To obtain simultaneous estimates for $A$ and $C$, we consider the coupled subproblems of estimating $C$, given an estimate of $A$ and estimating $A$, given an estimate of $C$. Each of these subproblems can be solved, respectively, via RPE and RSPE techniques after setting up appropriate cost functions. The estimates from the $C$ recursion and $A$ recursion can be fed back into the $A$ recursion and $C$ recursion, respectively, to couple the recursions.

Consider the minimization of the two separate cost functions (4.1) and (4.2).

$$\hat{\theta}_k^C = \arg\min_{\theta^C} \Bigg\{ V_k^1(\theta^C, \hat{\theta}_{k-1}^A) = \tfrac{1}{2} \sum_{i=1}^{k} \cdot \sum_{j=1}^{N} [y_i - C(\theta^C) e_j]^2 P(X_i = e_j | y_i) \Bigg\} \tag{4.1}$$

$$\hat{\theta}_k^A = \arg\min_{\theta^A} \Bigg\{ V_k^2(\theta^A, \hat{\theta}_{k-1}^C) = \tfrac{1}{2} \sum_{i=2}^{k} \cdot \|[\hat{X}_{i|\hat{\Theta}_{i-1}} - A(\theta^A) \hat{X}_{i-1|\hat{\Theta}_{i-2}}]\|^2 \Bigg\}. \tag{4.2}$$

Here, the two parameterizations $\theta^A \triangleq [\mathbf{a_1}, \cdots, \mathbf{a_N}]'$ and $\theta^C \triangleq [C^{(1)}, \cdots, C^{(N)}]'$ have been introduced, and $\hat{\Theta}_k \triangleq [\hat{\theta}_0^A, \hat{\theta}_0^C, \cdots \hat{\theta}_k^A, \hat{\theta}_k^C]$ denotes the history of estimation. The cost functions $V_k^1(\theta^C, \hat{\theta}_{k-1}^A)$ and $V_k^2(\theta^A, \hat{\theta}_{k-1}^C)$ are coupled through the $\hat{\theta}_{k-1}^A$ and $\hat{\theta}_{k-1}^C$ terms.

We proceed by introducing recursions in $\hat{\theta}_k^C$ and $\hat{\theta}_k^A$ before establishing convergence results.

$$\begin{aligned}\hat{\theta}_k^C &= \hat{\theta}_{k-1}^C + \dot{P}_k \phi_{k|\hat{\theta}_{k-1}^C, \hat{\theta}_{k-1}^A}\\ \dot{P}_k^{-1} &= \dot{P}_{k-1}^{-1} + \text{diag}\{[P(X_k = e_1|y_k), \cdots\\ &\qquad P(X_k = e_N|y_k)]\}\end{aligned} \tag{4.3}$$

where $\phi_{k|\hat{\theta}_{k-1}^C, \hat{\theta}_{k-1}^A} = d/d\theta^C V_k^1(\theta^C, \hat{\theta}_{k-1}^A)\Big|_{\theta^C = \hat{\theta}_{k-1}^C}$, and

$$\begin{aligned}\hat{\theta}_k^A &= \hat{\theta}_{k-1}^A + \check{P}_k \kappa_{k|\hat{\theta}_{k-1}^A, \hat{\theta}_{k-1}^C}\\ \check{P}_k^{-1} &= \check{P}_{k-1}^{-1} + \text{diag}(\underline{1}' \otimes \hat{X}_{k-1|k-1, \hat{\Theta}_{k-2}})\end{aligned} \tag{4.4}$$

where $\kappa_{k|\hat{\theta}_{k-1}^A, \hat{\theta}_{k-1}^C}$ is the same as $\kappa_{k|\hat{\theta}_{k-1}^S}$ defined in (3.10).

*Convergence Proof:* To demonstrate local convergence of the coupled algorithm, we first show that recursion (4.3) converges locally independently of $\hat{\theta}_k^A$ [or recursion (4.4)]. Next, we show local convergence of recursion (4.4).

*Lemma 6:* If the parameterized probability densities $b_k$ are independent of $\theta^A$, then the cost function (4.1) is independent of $\hat{\theta}_i^A$, $i = 0, \cdots, k$.

*Proof:* The lemma condition implies that $P(X_i = e_j|y_i)$ [as distinct from $P(X_i = e_j|Y_i)$] is independent of $\hat{\theta}_{k-1}^A$; hence, the cost $V_k^1(\theta^C, \hat{\theta}_{k-1}^A)$ in (4.1) is independent of $\hat{\theta}_{k-1}^A$. □

It follows from Lemma 6 that the recursions (4.3) are independent of $\hat{\theta}_{k-1}^A$; hence, convergence of (4.3) can be established as follows.

Consider the ODE (3.12) and with $\theta(\tau, k)$ abbreviated as $\theta_\tau^C$, and let us redefine for (4.3) the

$$f(\theta_\tau^C, k) = E[\phi_{k|\theta_\tau^C}] \tag{4.5}$$

and

$$G(\theta_\tau^C, k) = E[\text{diag}(\hat{X}_{k-1|k-1, \theta_\tau^C, \hat{\theta}_{k-2}^A})]. \tag{4.6}$$

The following lemma holds.

*Lemma 7:* If the parameterized probability densities $b_k$ are independent of $\theta^A$, then the recursion (4.3) will converge a.s. to the set $\overline{D}_C = \{\theta_\infty^C | \lim_{k\to\infty} E[f(\theta_\infty^C, k)] = 0\} \supset \theta^C$. Moreover, under the excitation condition $\dot{P}_k \to 0$ as $1/k$, then convergence of $\hat{\theta}_k^C$ is at the rate $1/k^{1/2}$.

*Proof:* A similar approach to Lemma 4 can be taken. See also [13]. □

Lemma 7 demonstrates local convergence results for the recursion (4.3). We now present convergence results for (4.4) under the assumption that (4.3) converges to the true value of $C$. Again, consider the ODE (3.12), and with $\theta(\tau, k)$ now abbreviated as $\theta_\tau^A$, let us redefine for (4.4)

$$f(\theta_\tau^A, k) = E[\kappa_{k|\theta_\tau^A}] \tag{4.7}$$

and

$$G(\theta_\tau^A, k) = E[\text{diag}(\hat{X}_{k-1|k-1, \theta_\tau^A, \theta^C})]. \tag{4.8}$$

The following lemma now holds.

*Lemma 8:* Given that $\hat{\theta}_k^C \to \theta^C$ converges a.s. as $k \to \infty$, then the recursion (4.4) will converge a.s. to the set $\overline{D}_A = \{\theta_\infty^A | \lim_{k\to\infty} E[f(\theta_\infty^A, k)] = 0\} \supset \theta^A$ (or possibly the boundary of the valid $A$ region if a projection step is performed). Moreover, under the excitation condition $\check{P}_k \to 0$ as $1/k$, then convergence of $\hat{\theta}_k^A$ is at the rate $1/k^{1/2}$.

*Proof:* Because $\hat{\theta}_k^C \to \theta^C$ as $k \to \infty$, then likewise, the cost $V_k^2(\theta^A, \hat{\theta}_k^C) \to V_k^2(\theta^A, \theta^C)$ as $k \to \infty$. Now, by inspection, it is clear that $V_k^2(\theta^A, \theta^C)$ is equivalent to $V_k(\theta)$ given in Section III. Hence, the rest of the proof follows Lemma 4. □

Together, Lemmas 6–8 imply local convergence of parameter estimates $\hat{\theta}_k^C$ and $\hat{\theta}_k^A$. However, note that Lemma 8 holds if and only if (4.3) has converged to the true value of $\theta^C$ rather than locally as Lemma 7 provides. In particular, for noise processes that are multimodal such as mixtures of Gaussian, this may not always occur.

*Remarks:*

1) Alternative cost functions for estimating $C$ have been proposed elsewhere; see [11] and [13].
2) The Lemma 6 conditions are not very restrictive. For example, Gaussian noise models and mixtures of Gaussians noise models both satisfy the lemma condition.
3) $P(X_i = e_j|y_i)$ can be replaced by $P(X_i = e_j|Y_i)$ in the cost function (4.1); however, convergence is no longer guaranteed. In simulations, it is found that a scheme with $P(X_i = e_j|Y_i)$ replacing $P(X_i = e_j|y_i)$ converges for all but the worst initial guesses. Note that if $\hat{\theta}_k^A = [1/N^2, \cdots, 1/N^2]'$, then $P(X_i = e_j|y_i) = P(X_i = e_j|Y_i)$, making $\hat{\theta}_0^A = [1/N^2, \cdots, 1/N^2]'$ a good initialization for the modified scheme if no other *a priori* information is available.
4) The dual cost function approach of this section has been found in simulations to converge more rapidly than a composite single cost function approach, e.g., minimization of $\check{V}_k(\theta) = V_k^1(\theta^C, \theta^A) + \lambda V_k^2(\theta^A, \theta^C)$, for some $\lambda$.
5) Implementation of recursions (4.3) and (4.4) requires $\text{O}(N^4) + \text{O}(N)$ calculations per time instant, which is less than the $\text{O}(N^4) + \text{O}(N^2)$ required using a composite single cost function approach. Further reduction in computational requirements can be achieved by implementing ELS versions of (4.3) and (4.4); however, convergence results are not yet established in this case.

## V. Simulations

### A. Implementation Considerations

In [11], several implementation issues are discussed, including the following:

- the use of step sequences and Polyak acceleration to improve transients performance;
- the modification of the parameter estimate recursions to include the variance of Markov state estimates and vice versa;
- modifications to allow tracking of slowly time-varying parameters.

The discussion in [11] equally applies to the algorithms presented in this paper.

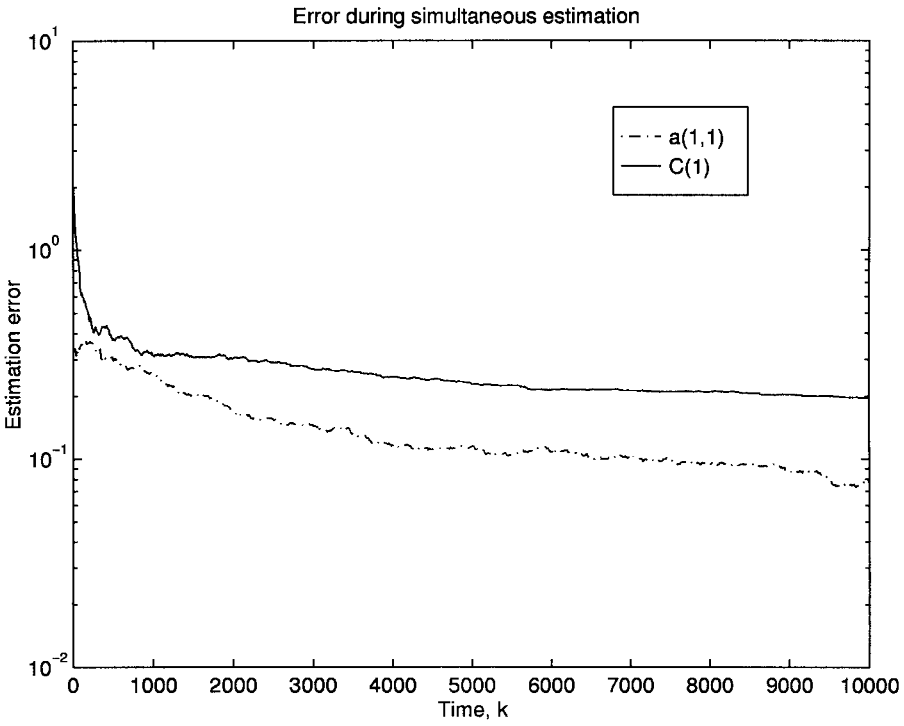


Fig. 1. Comparison of convergence rates.

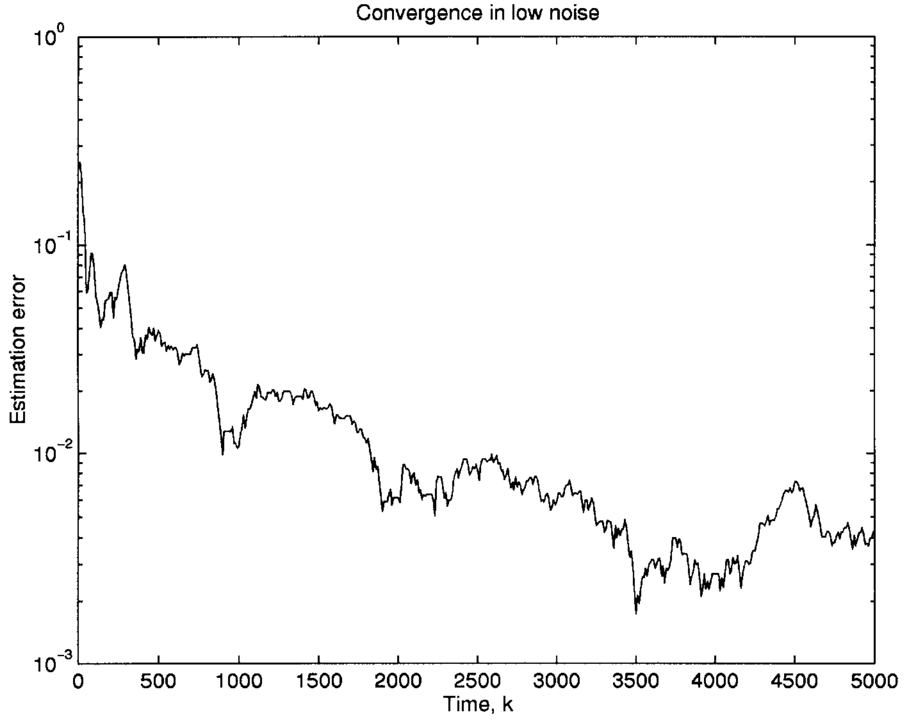


Fig. 2. Convergence in low noise.

### *B. Simulation Results*

We present results of simulation examples using computer-generated finite, discrete-state Markov chains to demonstrate features of the algorithms proposed in this paper.

*Estimation of Transition Probabilities:* A two-state Markov chain embedded in WGN is generated with parameter values $A^{ii} = 0.9,\ A^{ij} = (1 - A^{ii})$ for $i \neq j,\ C = [1,\ 3]',\ \sigma_w^2 = 1$, assuming $C$ and $\sigma_w^2$ known. The transition probability matrix is estimated using both the ELS and RSPE algorithms (3.5) and (3.8), respectfully. Fig. 1 shows a comparison of the estimation errors. This figure shows that convergence toward the true value occurs for both schemes and suggests that the RSPE scheme converges more rapidly that the ELS scheme.

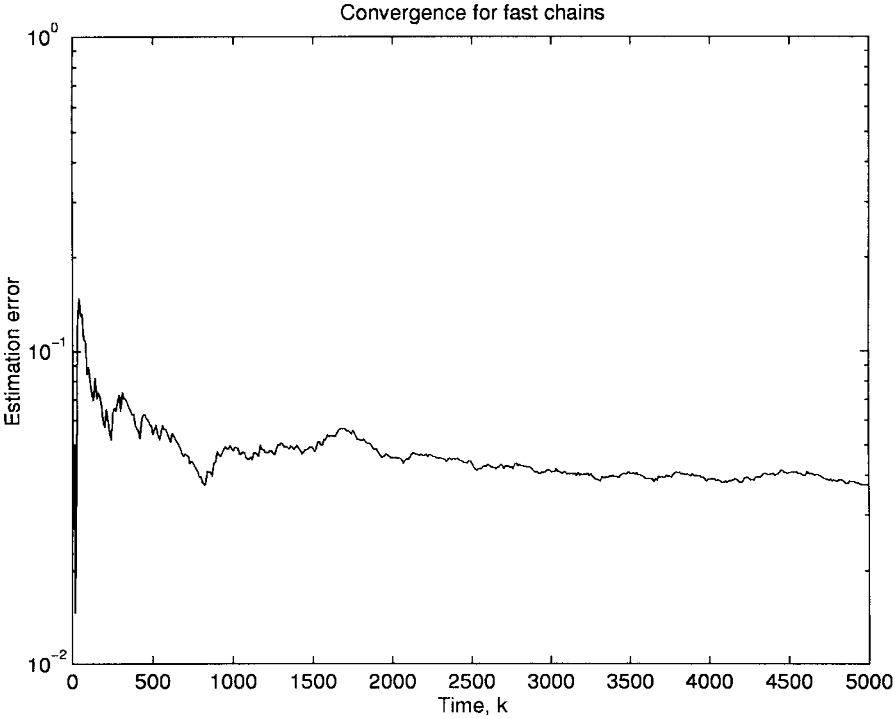


Fig. 3. Convergence of a fast chain.

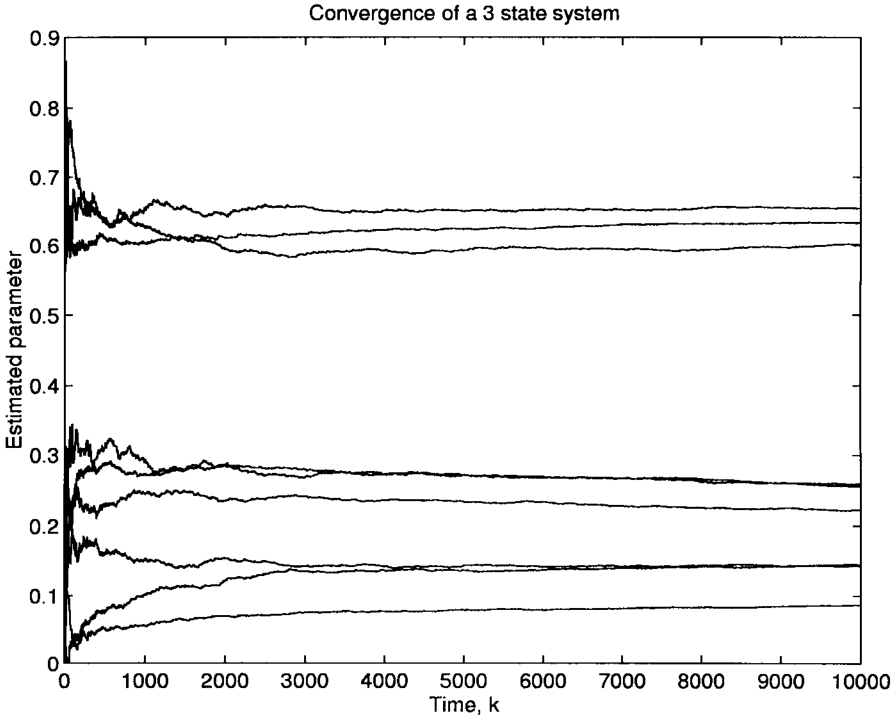


Fig. 4. Convergence of higher order chain.

*Estimation in Low Noise:* A two-state Markov chain embedded in WGN is generated with parameter values $A^{ii} = 0.9$, $A^{ij} = (1 - A^{ii})$ for $i \neq j$, $C = [1, 3]'$, $\sigma_w^2 = 0.0001$, assuming $C$ and $\sigma_w^2$ are known. The transition probabilities of the chain are estimated in low noise using the ELS algorithm, i.e., (3.5). For this noise level, the recursive schemes presented in [13] do not converge. Fig. 2 shows the error in estimation of (3.5) over time. This figure demonstrates that (3.5) convergence occurs in this low-noise environment.

*Estimation of Fast Markov Chains:* A two-state Markov chain embedded in WGN is generated with parameter values $A^{ii} = 0.6$, $A^{ij} = (1 - A^{ii})$ for $i \neq j$, $C = [1, 3]'$, $\sigma_w^2 = 1$,

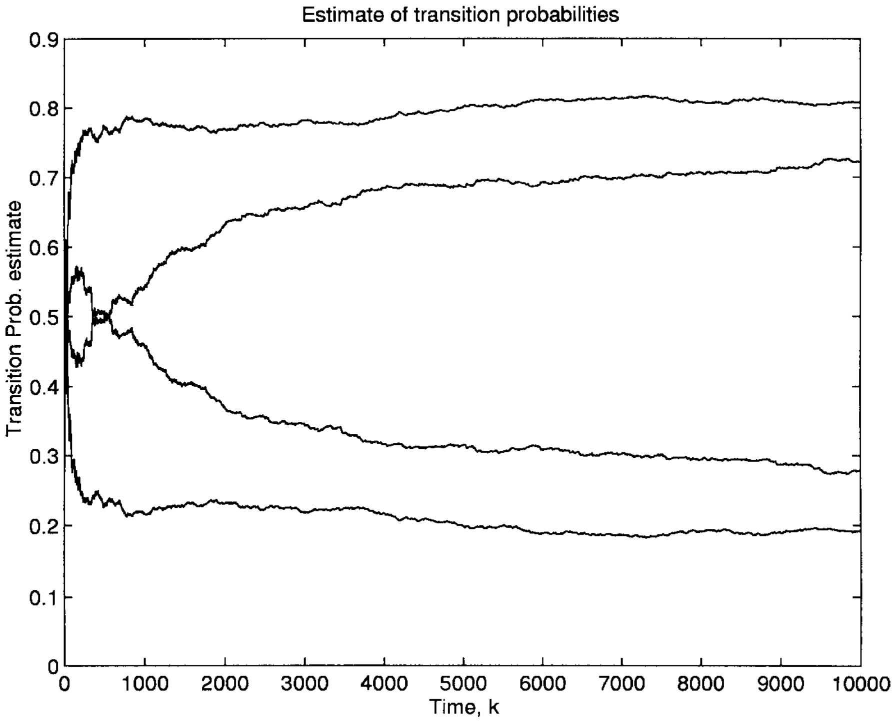


Fig. 5. Simultaneous estimation—Transition probabilities.

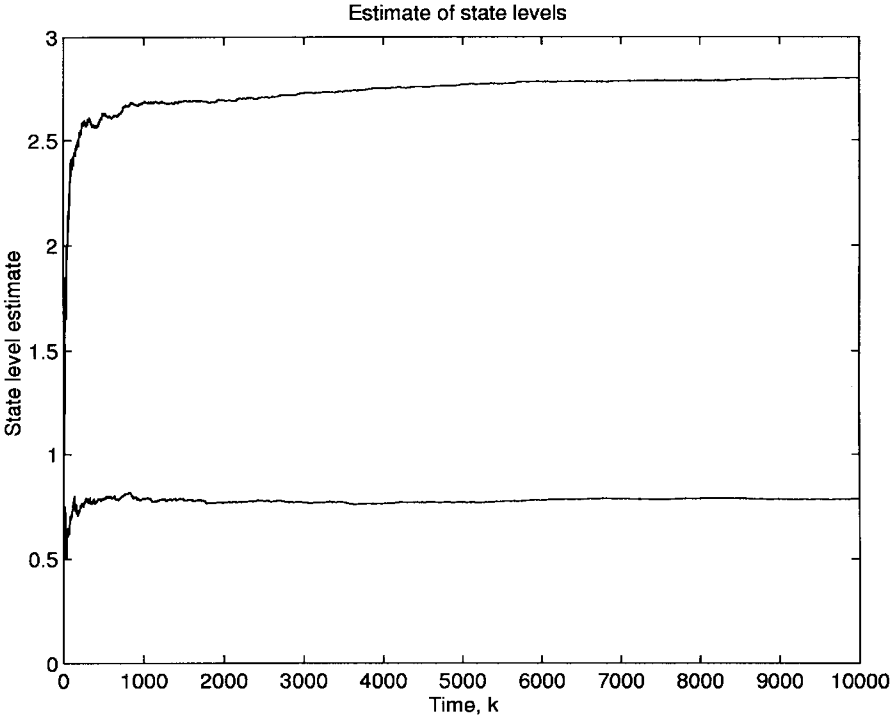


Fig. 6. Simultaneous estimation—State levels.

assuming $C$ and $\sigma_w^2$ are known. The transition probabilities of the chain are estimated using the RSPE algorithm, i.e., (3.8). Fig. 3 shows the size of the estimation error over time and demonstrates that convergence occurs.

*Higher Order Chain:* A three-state Markov chain embedded in WGN is generated with parameter values $A^{ii} = 0.9$, $A^{ij} = (1 - A^{ii})/2$ for $i \neq j$, $C = [1, 3, 5]'$, $\sigma_w^2 = 1$, assuming $C$ and $\sigma_w^2$ are known. The transition probabilities of the chain are estimated using the RSPE algorithms; see (3.8). Fig. 4 shows the time evolution of the transition probabilities estimates. This figure demonstrates that estimates converge to the correct values.

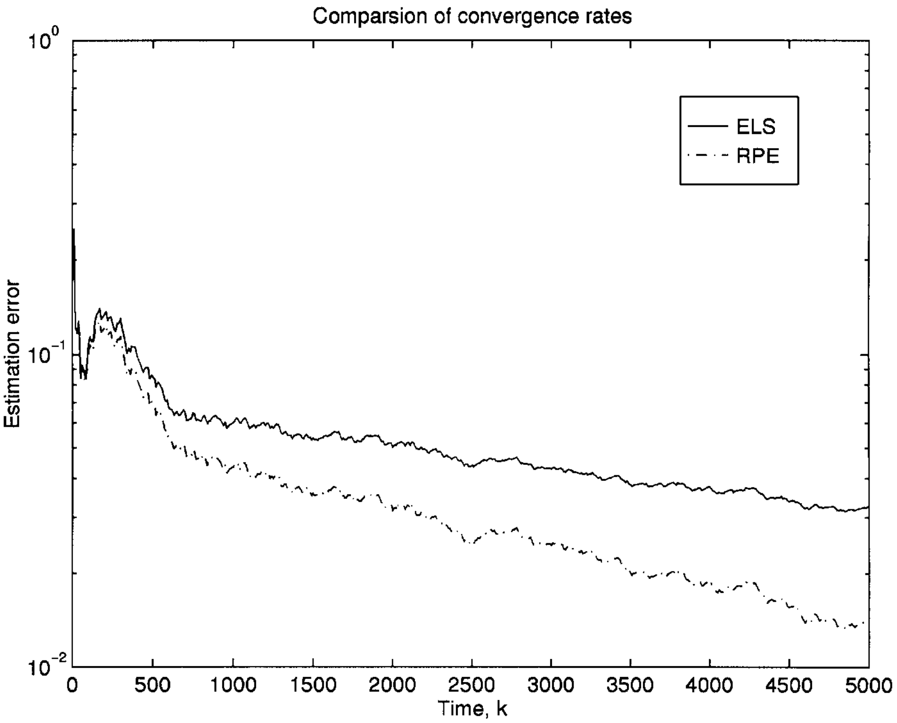


Fig. 7. Simultaneous estimation—Estimation error.

*Simultaneous Estimation:* A two-state Markov chain embedded in WGN is generated with parameter values $A^{ii} = 0.8$, $A^{ij} = (1 - A^{ii})$ for $i \neq j$, $C = [1, 3]'$, $\sigma_w^2 = 1$ with $\hat{\theta}_0^C = [0.5, 1]$ and $\hat{\theta}_0^A = 0.5 \;\forall\; i$. The transition probabilities and the state output mappings of the chain are estimated simultaneously using (4.3) and (4.4). Figs. 5 and 6 show the time evolution of the transition probabilities and state level estimates, respectively. Fig. 7 show the estimation error in $C^{(1)}$ and the transition probability $A^{11}$. These figures demonstrate that estimates converge to the correct values. Comparison with the results presented in [13] suggest that the convergence is considerably more rapid.

## VI. Conclusions

In this paper, we have proposed new algorithms for recursive estimation of the state transition probabilities for HMM's based on ELS and RSPE techniques. These algorithms avoid the ill conditioning in low noise of the schemes in [13]. Convergence analysis for the RSPE algorithm is provided via an ODE approach, but no convergence results are presented for the ELS algorithm. Despite the lack of convergence results, the ELS algorithm is attractive because it has computational complexity of only O($N^2$) per time instant, compared with the RPE scheme (of [13]) and the RSPE scheme of this paper which have computational complexity O($N^4$).

This paper also proposes a scheme for the simultaneous estimation of state output mapping levels and the state transition probabilities. Local convergence results are presented. The simulation studies presented demonstrate that the schemes proposed in this paper converge from reasonable initializations and are effective in low noise levels.

## 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] D. Clements and B. D. O. Anderson, "A nonlinear fixed-lag smoother for finite-state Markov processes," *IEEE Trans. Inform. Theory,* vol. IT-21, pp. 446–452, July 1975.

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

[5] 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, Aug. 1993.

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

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

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

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

[10] A. E. Bryson and Y. Ho, *Applied Optimal Control.* London, U.K.: Ginn, 1969.

[11] J. J. Ford and J. B. Moore, "On adaptive HMM state estimation," *IEEE Trans. Signal Processing,* to be published.

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

[13] 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.

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

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

[16] B. T. Polyak and A. B. Juditsky, "Acceleration of stochastic approximation by averaging," *SIAM J. Contr. Optim.,* vol. 30, pp. 838–855, July 1992.

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

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

[19] L. Shue, B. D. O. Anderson, and S. Dey, “ Exponential stability of filters and smoothers for hidden Markov models,” in *Proc. ECC*, Brussels, Belgium, July 1–4, 1997.

[20] F. Le Gland and L. Mevel, “Geometric ergodicity in hidden Markov models,” *IRISA,* Int. Pub. 1028, July 1996.

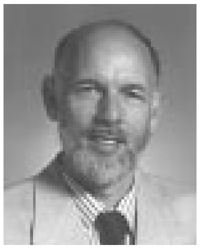
[21] I. B. Collings and J. B. Moore, “Multiple-prediction-horizon recursive identification of hidden Markov models,” in *Proc. ICASSP,* Atlanta, GA, 1996.

[22] P. R. Kumar and P. Varaiya, *Stochastic Systems.* Englewood Cliffs, NJ: Prentice-Hall, 1986.

[23] T. Söderström 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 the 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 Senior Lecturer at the Electrical Engineering Department, University of Newcastle, Callaghan, Australia, in 1967 and promoted to Associate Professor in 1968 and Full Professor (personal chair) in 1973. He was Department Head from 1975 to 1979. In 1982, he was appointed Professorial Fellow in the Department of Systems Engineering, Research School of Physical Sciences, Australian National University, Canberra, and promoted to Professor in 1990. He has been Head of the department since 1992 and is now with the Research School of Information Sciences and Engineering. 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 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 with T. T. Tay and I. Mareels entitled *High Performance Control* (Boston, MA: Birkhäsuer, 1997). He has held visiting academic appointments at the University of Santa Clara in 1968, the University of Maryland, College Park, in 1970 and 1994; Colorado State University, Fort Collins, and Imperial College, London, U.K., in 1974; the University of California, Davis, in 1977; the University of Washington, Seattle, in 1981; Cambridge University, Cambridge, U.K., and the National University of Singapore, Singapore, in 1985; the University of California, Berkeley, in 1987, 1989, and 1991; the University of Alberta, Edmonton, Alta., Canada, from 1992 to 1994; the University of Rengensburg, Rengensberg, Germany, in 1993; the Institute of Industrial Science, University of Tokyo (for six months from September 1993, where he held the Toshiba Chair); the Imperial College of Science, Technology, and Medicine, London, U.K., in 1995; the Technical University of Munich, Munich, Germany, in 1995; the University of Würzberg, Würzberg, Germany, in 1996 and 1997; and the Chinese University of Hong Kong, Hong Kong, in 1997). He has spent periods in industry as a Design Engineer and as a Consultant and currently has research grants from industry and government laboratories. He is a Team Leader in the Cooperative Research Centre for Robust and Adaptive Systems in the department.

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