Exponential Stability of Filters and Smoothers for Hidden Markov Models

Louis Shue, Brian D. O. Anderson, *Fellow, IEEE,* and Subhrakanti Dey

***Abstract*— In this paper, we address the problem of filtering and fixed-lag smoothing for discrete-time and discrete-state hidden Markov models (HMM's), with the intention of extending some important results in Kalman filtering, notably the property of exponential stability. By appealing to a generalized Perron–Frobenius result for non-negative matrices, we are able to demonstrate exponential forgetting for both the recursive filters and smoothers; furthermore, methods for deriving overbounds on the convergence rate are indicated. Simulation studies for a two-state and two-output HMM verify qualitatively some of the theoretical predictions, and the observed convergence rate is shown to be bounded in accordance with the theoretical predictions.**

***Index Terms*—Exponential forgetting, fixed-lag smoothing, hidden Markov model.**

I. Introduction

THE PURPOSE of this paper is to explain how some important results in Kalman filtering and smoothing can be carried over to contemporary problems involving hidden Markov models (HMM's). Formal definitions of the HMM's considered are given in Section II; suffice it to say here that we restrict attention in this paper to HMM's with finite state and observation sets. Relaxation of the results to cope with countably infinite states and continuous observations will be considered elsewhere, with the relaxation to continuous observations under some circumstances at least being completely straightforward. Presently, our attention is restricted to discrete-time models; tackling the problem for continuous-time models involve new tools that we are seeking to develop.

The problems we consider are of two kinds, but related. Most Kalman filters have an exponential stability property [4], even in nonstationary situations. This ensures that if the filter is initialized at some finite time, the initial conditions are forgotten exponentially fast. In addition, old measurements are forgotten exponentially fast; thus, the remote past cannot significantly influence the present. These properties of a filter, normally assured by making assumptions concerning complete stabilizability and detectability in the signal model, reflect common sense. The absence of this property in a given filter is likely to lead to inaccurate estimates, as round-off errors may overpower the measurements.

The first results of the paper extend the exponential forgetting result to HMM's. Similar results along these lines can be found in [1] and, of a more preliminary nature, in [6]. Our results are probably more comprehensive, appealing to more recent developments in properties of products of positive and non-negative matrices [12] and allowing the calculation of convergence rates. We include some derivations of how to compute the exponential forgetting rate and show that the forgetting rate of the HMM filter is, in a sense, at least as fast as that in the HMM itself, thus paralleling a similar result in Kalman filtering literature [4, p. 85].

The second main thrust of the paper deals with fixed-lag smoothing. In fixed-lag smoothing, measurements up to the present time are used to infer information about the state Δ time intervals in the past (Δ being a fixed quantity). Fixed-lag smoothing uses more measurements than does filtering to infer information about a certain quantity, which argues for using a large Δ. Moreover, there is necessarily a delay, relative to filtering, in inferring the information, as a result of the need to collect extra information. This second consideration argues for using a small Δ, and this leads to a key question: How should Δ be chosen?

In smoothing problems that can be tackled using Kalman filter ideas, it has been found that as Δ increases, there is increasingly less additional benefit gained in terms of the quality of the estimates. The fall off in the rate of increase of benefit is exponential, and with Δ equal to four or five times the dominant time constant of the Kalman filter, practically all the benefit derivable from smoothing (even with $\Delta \rightarrow \infty$) is achieved (see, for example, [2]). This important property can guide the selection of Δ.

In this paper, we establish equations for the HMM fixed-lag smoother (using two different approaches, having an eye on future extensions). Again, we are able to show that the rate of increase of benefit from smoothing falls off exponentially as Δ increases, with the same rate as the forgetting rate of the HMM filter, which is, furthermore, at least as fast as that of the original hidden Markov signal model. The results

Manuscript received August 22, 1996; revised December 19, 1997. This work was supported by the activities of the Cooperative Research Centre for Robust and Adaptive Systems by the Australian Commonwealth Government under the Cooperative Research Centres Program. The associate editor coordinating the review of this paper and approving it for publication was Prof. M. H. Er.

L. Shue and B. D. O. Anderson 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.

S. Dey was with the Department of Systems Engineering, Research School of Information Sciences and Engineering, Australian National University, Canberra, Australia. He is now with the Institute of Systems Research, University of Maryland, College Park, MD 20742 USA.

Publisher Item Identifier S 1053-587X(98)05227-1.

were predicted, as the subject of a simulation study, more than 20 years ago [7]. They constitute important guidelines for the use of fixed-lag smoothing, as opposed to filtering, of HMM's.

The outline of the paper is as follows. In Section II, we define the signal model and demonstrate exponential forgetting of initial conditions for recursive filters. In Section III, the filtering ideas are extended to smoothed estimation scheme, with further extensions to systems described with non-negative rather than positive matrices in Section IV. Section V contains some simulations results and discussions in which the performance of filtered and smoothed estimates for a two-state HMM are compared. In Section VI, we summarize the ideas presented and indicate possible areas of applications and for further investigations.

II. Filtering

In this section, we will consider the problem of filtering. Formally, a filtered estimate is a conditional estimate of the state of the HMM, given a series of measurements. In other words, the appropriate question to ask is the following: Given a series of measurements up to time k, what is the probability, at time k, of the HMM being found in a particular state?

A. Signal Model

Consider a first-order discrete-time and discrete-state Markov process X_k, where the subscript k denotes time. For simplicity, we shall define the states to be the values $1, 2, \cdots, N$. At each time instant k, a corresponding signal Y_k is observed, again having discrete values, in the range $1, 2, \cdots, M$. We will adopt the convention that a lowercase x_k denotes the actual state value and likewise for y_k. The probability vectors for X_k and Y_k are updated by the system matrices A (the state transition probability matrix) and C (observation matrix), with $A = \{a_{ij}\} = \{\Pr(X_{k+1} = i | X_k = j)\}$ and $C = \{c_{mn}\} = \{\Pr(Y_k = m | X_k = n)\}$. Further, unless otherwise stated, $a_{ij} > 0$ and $c_{mn} > 0$, $\forall i, j, n \in \{1, 2, \cdots, N\}$, $\forall m \in \{1, 2, \cdots, M\}$.

Remark 2.1: In the present definition, A and C are column-stochastic matrices, i.e., $\Sigma_{i=1}^{N} a_{ij} = 1$ and $\Sigma_{m=1}^{M} c_{mn} = 1$. This may be contrary to other notations (see e.g., [10]) and must be kept in mind to avoid possible confusion.

Remark 2.2: For the case of independent and identically distributed (iid) processes in which all state transitions are equally likely, A has identical entries in each row, and consequently, columns in A are identical. For a similar situation involving the C matrix, the interpretation is that the output process is independent of the state process.

B. Evolution of Filtered Distributions

Let $\Pi_{k|k}$ and $\Pi_{k+1|k}$ be the filtered and one-step prediction probability vector, with the ith entry being $\Pr(X_k = i | Y_0, Y_1, \cdots, Y_k)$ and $\Pr(X_{k+1} = i | Y_0, Y_1, \cdots, Y_k)$, respectively. By using a combination of Bayes' rule of conditional probability and the Markov property, the time evolution relations for the filtered probability vector are

$$\Pi_{k+1|k} = A\Pi_{k|k} \tag{1}$$

$$\Pi_{k+1|k+1} = \frac{1}{1'_N C_{y_{k+1}} \Pi_{k+1|k}} C_{y_{k+1}} \Pi_{k+1|k} \tag{2}$$

where $1'_N C_{y_{k+1}} \Pi_{k+1|k} = [1 \cdots 1] C_{y_{k+1}} \Pi_{k+1|k}$ is a scalar normalizing constant to ensure the entries of $\Pi_{k+1|k+1}$ sum to 1, and $C_{y_{k+1}} = \text{diag}(c_{l1} \quad c_{l2} \quad c_{l3} \quad \cdots \quad c_{lN})$ when $y_{k+1} = l$.

C. Exponential Forgetting

By iterating (1) and (2), the filtered probability vector at time k can be expressed in terms of an arbitrarily chosen initial distribution $\Pi_{0|0}$

$$\begin{aligned}\Pi_{k|k} &= \frac{1}{1'_N C_{y_k} A \Pi_{k-1|k-1}} C_{y_k} A \Pi_{k-1|k-1} \\ &= \frac{1}{1'_N C_{y_k} A C_{y_{k-1}} A \cdots C_{y_1} A \Pi_{0|0}} \times \\ &\quad (C_{y_k} A C_{y_{k-1}} A \cdots C_{y_1} A) \Pi_{0|0} \\ &= \frac{1}{1'_N U_{1,k} \Pi_{0|0}} U_{1,k} \Pi_{0|0}\end{aligned} \tag{3}$$

where

$$U_{1,k} = C_{y_k} A C_{y_{k-1}} A \cdots C_{y_1} A. \tag{4}$$

We now proceed to derive initial-condition forgetting of the filter by appealing to the generalized Perron–Frobenius result (see Appendix A, in particular Theorem A.1, as well as [12]) for an inhomogeneous product of matrices. Broadly, this theorem states tha under certain conditions, a product of positive matrices of the form $U_{1,r} = H_r H_{r-1} \cdots H_1$ may become dyadic[1] as $r \to \infty$.

As stated previously, $A > 0$ and $C > 0$; therefore, (4), which is a product of successive $C_y A$, is strictly positive, and the requirements of the aforementioned theorem are automatically satisfied. This means that as $k \to \infty$

$$U_{1,k} \to \boldsymbol{\alpha}(k) \boldsymbol{\beta}' \tag{5}$$

for some positive column vector $\boldsymbol{\alpha}(k)$ and positive row vector $\boldsymbol{\beta}'$. Without loss of generality, let $\beta_1 = 1$. Hence, the normalized filter probability vector becomes

$$\begin{aligned}\Pi_{k|k} &= \frac{1}{1'_N U_{1,k} \Pi_{0|0}} U_{1,k} \Pi_{0|0} \\ &\to \frac{1}{1'_N U_{1,k} \Pi_{0|0}} \begin{bmatrix} \alpha_1(k) \\ \alpha_2(k) \\ \vdots \\ \alpha_N(k) \end{bmatrix} [1 \ \beta_2 \ \beta_3 \cdots \beta_N] \Pi_{0|0} \\ &\to \frac{1}{\sum_{i=1}^{N} \alpha_i(k)} \boldsymbol{\alpha}(k)\end{aligned}$$

where $\boldsymbol{\alpha}(k)$ is a vector, which is a function of time k, but independent of the initial distribution $\Pi_{0|0}$.

[1] The term dyadic means that a matrix is of rank 1.

The forgetting property occurs as a direct consequence of the inhomogeneous product $U_{1,k}$ attaining rank of 1. In addition, each entry of $\Pi_{k|k}$ is a ratio of two quantities converging as in (5). Hence, the rate of convergence of $\Pi_{k|k}$ is identical to the rate of convergence of (5), which, furthermore, is exponential and is computable [12], as we shall now argue in detail.

D. Rate of Convergence

The Birkhoff coefficient $\tau_B(\cdot)$ (see Appendix A) is a measure of how close a given matrix is to having rank 1. Since the forgetting of initial conditions relies on this property, it is of benefit to be able to numerically evaluate this quantity. Furthermore, $\tau_B(U_{1,k})$ will provide information as to how fast $U_{1,k}$ converges to a rank 1 matrix.

1) Worst-Case Rates: To a first approximation, the overall $\tau_B(\cdot)$ for the filter is upper bounded by the $\tau_B(\cdot)$ of the original Markov state process

$$\begin{aligned}\tau_B(C_{y_k}AC_{y_{k-1}}\cdots C_{y_1}A)&\le \tau_B(C_{y_k}A)\tau_B(C_{y_{k-1}}A)\cdots\tau_B(C_{y_1}A)\\&\le \tau_B(A)^k\end{aligned}\tag{6}$$

using the fact that $\tau_B(\cdot)=1$ for a diagonal matrix with positive diagonal entries (see also Appendix A). This is analogous to a Kalman filtering result (see [4, p. 85]), which says that

$$|\det(\mathcal{F})|\le|\det(F)|$$

or

$$\left|\prod\lambda_i(\mathcal{F})\right|\le\left|\prod\lambda_i(F)\right|$$

where $\mathcal{F}$ is the Kalman filter state feedback matrix, and F is the state feedback matrix of the original signal model. That is, the product of the poles of the Kalman filter is less than or equal to the product of the poles of the original signal model.

A tighter overbound on the worst-case rate can be obtained as follows. Let $\mu_2=\sup_{y_i}\tau_B(AC_{yi}A)$; then

$$\begin{aligned}&\tau_B(C_{y_k}AC_{y_{k-1}}\cdots C_{y_1}A)\\&\quad\le\tau_B(C_{y_k})\tau_B(AC_{y_{k-1}}A)\tau_B(C_{y_{k-2}})\tau_B(AC_{y_{k-3}}A)\cdots\\&\quad\le\begin{cases}\mu_2^{(k/2)}, & \text{if } k \text{ even}\\ \tau_B(A)\mu_2^{(k-1)/2}, & \text{if } k \text{ odd.}\end{cases}\end{aligned}$$

More generally, let q be a fixed integer. Define

$$\mu_q=\sup_{y_1,y_2,\cdots,y_{q-1}}\tau_B(AC_{y_{q-1}}AC_{y_{q-2}}A\cdots C_{y_1}A).$$

Suppose that $k=rq+s$, with $s<q$; then

$$\tau_B(C_{y_k}AC_{y_{k-1}}\cdots C_{y_1}A)\le\tau_B^s(A)\mu_q^r.\tag{7}$$

Naturally the larger q is, the more the computations involved but, also the tighter the bound.

2) Average Rate: It is also sometimes more appropriate to use some sort of average $\tau_B(\cdot)$ for the recursive filter, given that each C_{y_k} is the result of a random variable Y_k. Consider the variation of (6) in

$$\begin{aligned}\ln\tau_B(C_{y_k}A\cdots C_{y_1}A)&\equiv\ln\tau_B(U_{1,k})\\&\le\sum_{i=1}^k\ln\tau_B(C_{y_i}A).\end{aligned}$$

Assuming that the Y_k process is ergodic (which is certainly true if all entries of A and C are positive), it follows that

$$\begin{aligned}\lim_{k\to\infty}\frac{1}{k}\ln\tau_B(U_{1,k})&\le\lim_{k\to\infty}\frac{1}{k}\sum_{i=1}^k\ln\tau_B(C_{y_i}A)\\&=E[\ln\tau_B(C_{y_i}A)]\\&\le\ln E[\tau_B(C_{y_i}A)]\end{aligned}\tag{8}$$

where Jensen's inequality has been used in the last line. If $E[\ln\tau_B(C_{y_i}A)]=\ln\tau_B(U_{1,k})_{\rm ob}$ with $[\tau_B(C_{y_i}A)_{\rm ob}]^k$ providing an overbound for $\tau_B(U_{1,k})$, then (8) implies that $E[\tau_B(C_{y_i}A)]=\tau_B(U_{1,k})_{\rm av}\ge\tau_B(U_{1,k})_{\rm ob}$. Since $\tau_B(C_{y_i}A)=\tau_B(A)$, (8) also leads to the conclusion that $\tau_B(U_{1,k})_{\rm ob}=\tau_B(A)$. This is in fact identical to (6) and therefore does not provide any extra information.

We can, however, achieve tighter bounds by paralleling the $q\ge2$ arguments above. Thus

$$\begin{aligned}\ln\tau_B(U_{1,k})&\le\ln\tau_B(C_{y_k}AC_{y_{k-1}}A)\\&\quad+\ln\tau_B(C_{y_{k-2}}AC_{y_{k-3}}A)+\cdots\\&=\ln\tau_B(AC_{y_{k-1}}A)+\ln\tau_B(AC_{y_{k-3}}A)+\cdots\end{aligned}$$

and as $k\to\infty$

$$\lim_{k\to\infty}\frac{1}{k}\ln\tau_B(U_{1,k})\le\frac{1}{2}E[\ln\tau_B(AC_{y_i}A)].$$

More generally

$$\begin{aligned}&\lim_{k\to\infty}\frac{1}{k}\ln\tau_B(U_{1,k})\\&\quad\le\frac{1}{q}E[\ln\tau_B(AC_{y_{q-1}}AC_{y_{q-2}}\cdots C_{y_1}A)].\end{aligned}\tag{9}$$

For the two-level output HMM studied in Section V, by grouping two successive C_yA's to form a three-term average, a tighter bound than (6) is found to be

$$\begin{aligned}\ln\tau_B(U_{1,k})_{\rm ob}&=\tfrac{1}{2}E[\ln\tau_B(AC_{y_k}A)]\\&=\tfrac{1}{2}[\Pr(Y=1)\ln\tau_B(AC_{y=1}A)\\&\quad+\Pr(Y=2)\ln\tau_B(AC_{y=2}A)].\end{aligned}\tag{10}$$

Stationarity is, of course, critical. This procedure can be executed by grouping more successive C_yA's and, thus, obtaining an even tighter bound on $\tau_B(\cdot)_{\rm ob}$. Naturally, as increasingly more terms are included, the complexity of the calculations increases. In Section V, both (6) and (10) have been used to assess how quickly the forgetting of initial conditions can be achieved.

III. SMOOTHING

Recall that a filtered probability vector is the probability of the state at a certain time j, given a series of measurements up to that time. Smoothing is an extension of this idea in the sense that the conditional probability vector for X_j uses more measurements: not just up to time j but beyond that time until some later time $k > j$. Since more measurements are used, better estimates should result. However, as more measurements are taken prior to the evaluation of the state estimates, there is an inherent delay before the smoothed estimates are available. Consequently, some practical questions that may be asked are as follows: What is a tolerable delay, and what is the improvement to be gained as more measurements are taken? We shall concentrate on the second question and provide only qualitative remarks relating to the first point.

Historically, three types of smoothing problems have been studied. Fixed-point smoothing is concerned with obtaining the smoothed probability vector $\Pi_{j|S}$ for some fixed time j, whereas the total number of measurements S remains variable. Fixed-lag smoothing deals with the problem of obtaining the probability vector $\Pi_{j|j+\Delta}$ for all j with Δ fixed. Last, fixed-interval smoothing is concerned with smoothing over a fixed interval. That is, the quantity of interest is $\Pi_{j|S}$, with S fixed, and $0 \leq j \leq S$. A fixed-interval smoothing scheme for HMM's can be found in [10]. Based on the backward recursions reviewed in [10], a numerical fixed-lag smoothing scheme for HMM's was developed in [9] that achieves smoothed state estimates while adaptively learning the unknown HMM parameters.

In this section, we will implement two fixed-lag smoothing schemes for an HMM by formulating the problem in terms of fixed-point smoothing and subsequently varying the fixed index j to obtain a fixed-lag smoothing scheme. The first methodology involves the construction of a fictitious augmented state model, as outlined in [4] and [7] and subsequently employing ordinary filtering for such an augmented state vector; the smoothed probability vector for the original HMM is embedded within the filtered probability vector for the augmented HMM. In the second method, the smoothed estimate consists of merging the filtered probabilities from forward and backward models[2] using a procedure outlined in [5].

A. Smoother from Augmented Signal Model

Our first approach to deriving equations for smoothing is based on carrying over an idea used in extending Kalman filtering results to smoothing results. This idea uses augmentation of a signal model, so that a filtered estimate for the state of the augmented model contains within it something equivalent to a smoothed estimate for the state of the original, unaugmented, signal model.

More precisely, in Kalman filtering (see [4]), if the original signal model is

$$x_{k+1} = Fx_k + Gw_k$$
$$z_k = H'x_k + v_k$$

where x_k is the state, z_k the measurement, and w_k, v_k (generally) white Gaussian processes, and if a smoothed estimate of x_j (j fixed) is desired, the augmented signal model has the following state vector for $k \geq j$:

$$\mathcal{X}_k = \begin{bmatrix} x_k \\ x_j \end{bmatrix}.$$

Thus

$$\mathcal{X}_{k+1} = \begin{bmatrix} F & 0 \\ 0 & I \end{bmatrix} \mathcal{X}_k + \begin{bmatrix} G \\ 0 \end{bmatrix} w_k$$
$$\mathcal{Z}_k = [H' \quad 0]\mathcal{X}_k + \begin{bmatrix} v_k \\ 0 \end{bmatrix}.$$

Obviously, a filtered estimate of $\mathcal{X}$ at time k given measurements up to k denoted by $\hat{\mathcal{X}}_{k|k}$ includes $\hat{x}_{k|k}$, which is the normal filtered estimate, and $\hat{x}_{j|k}$, which is the smoothed estimate. If $k = j + \Delta$ with j variable and Δ fixed, a fixed-lag smoothed estimate is available. The initial condition for the augmented filter and covariance equation is

$$X_{j|j-1} = \begin{bmatrix} \hat{x}_{j|j-1} \\ \hat{x}_{j|j-1} \end{bmatrix}$$

and

$$\begin{bmatrix} \Sigma_{j|j-1} & \Sigma_{j|j-1} \\ \Sigma_{j|j-1} & \Sigma_{j|j-1} \end{bmatrix}.$$

We carry this idea over to HMM's in the following way.

Definition 3.1: For each $k \geq j$, let $\mathcal{Z}_k = Z_{j,k} = (X_j X_k)'$ be an augmented state vector consisting of the states of the original Markov process, as defined in Section II-A, at a fixed time j and a variable time k.

We shall now argue that the output process (Y_k) of the original HMM (with state X_k) can also be regarded as the output process of an HMM with state $\mathcal{Z}_k$ for $k \geq j$.

From Definition 3.1, it can be seen that $\mathcal{Z}_k$ can only assume the values (1, 1), (1, 2), $\cdots, (1, N)$, (2, 1), $\cdots, (N, N)$. By denoting the probability vector that $\mathcal{Z}_k$ is in each of the N^2 possible states at time k as $\mathcal{P}_k$, it is seen that these hold

$$\mathcal{P}_{k+1} = \mathcal{A}\mathcal{P}_k$$

where

$$\mathcal{A} = \begin{bmatrix} A & 0 & \cdots & 0 \\ 0 & A & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \ldots & A \end{bmatrix} = I_N \otimes A.$$

Further

$$\mathcal{P}_j = \begin{bmatrix} \Pr(X_j = 1) \\ 0 \\ \vdots \\ 0 \\ \Pr(X_j = 2) \\ 0 \\ \vdots \\ 0 \\ \Pr(X_j = 3) \\ \vdots \\ \Pr(X_j = N) \end{bmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} N \text{ zeros} \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} N \text{ zeros.} \end{matrix}$$

[2] For proof of the equivalence between the two methods, see Appendix B.

Suppose that the output process associated with $\mathcal{Z}_k$ is the same as before, and consequently, $\Pr(Y_k = l|\mathcal{Z}_k) = \Pr(Y_k = l|X_k)$. This means that the corresponding observation matrix $\mathcal{C}$ is

$$\mathcal{C} = [C \quad C \quad \cdots \quad C] = 1'_N \otimes C$$

and

$$\mathcal{C}_{y_{k+1}} = \begin{bmatrix} C_{y_{k+1}} & 0 & \cdots & 0 \\ 0 & C_{y_{k+1}} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & C_{y_{k+1}} \end{bmatrix} = I_N \otimes C_{y_{k+1}}.$$

The filtered probability vector for $\mathcal{Z}_k$, which is denoted as $\hat{\Pi}_{j,k|k}$, evolve according to the following recursions

$$\hat{\Pi}_{j,k+1|k} = \mathcal{A}\hat{\Pi}_{j,k|k} \tag{11}$$

$$\hat{\Pi}_{j,k+1|k+1} = \frac{1}{1'_{N^2}\mathcal{C}_{y_{k+1}}\hat{\Pi}_{j,k+1|k}}\mathcal{C}_{y_{k+1}}\hat{\Pi}_{j,k+1|k} \tag{12}$$

where $1'_{N^2}\mathcal{C}_{y_{k+1}}\hat{\Pi}_{j,k+1|k}$ is a scalar normalizing constant.

Since, by definition, the ith entry of the smoothed probability vector $\Pi_{j|j+\Delta}$ at time j with lag Δ for the unaugmented HMM is just

$$\begin{aligned}&\Pr(X_j = i|Y_0, \cdots, Y_{j+\Delta}) \\ &\quad = \sum_{l=1}^{N} \Pr(X_j = i, X_{j+\Delta} = l|Y_0, \cdots, Y_{j+\Delta})\end{aligned}$$

the smoothed probability vector $\Pi_{j|j+\Delta}$ for the unaugmented HMM can be evaluated by summing appropriate terms in the filtered probability vector for the augmented model

$$\begin{aligned}\Pi_{j|j+\Delta} &= \begin{bmatrix} 1\cdots 1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & 1\cdots 1 \end{bmatrix}\hat{\Pi}_{j,j+\Delta|j+\Delta} \\ &= (I_N \otimes 1'_N)\hat{\Pi}_{j,j+\Delta|j+\Delta} \\ &= \frac{1}{\Theta_{j,j+\Delta}}[(I_N \otimes 1'_N)(\mathcal{C}_{y_{j+\Delta}}\mathcal{A}\mathcal{C}_{y_{j+\Delta-1}}\mathcal{A}\cdots \\ &\qquad \mathcal{C}_{y_{j+1}}\mathcal{A})]\hat{\Pi}_{j,j|j} \\ &= \frac{1}{\Theta_{j,j+\Delta}}[(I_N \otimes 1'_N)(I_N \otimes C_{y_{j+\Delta}})(I_N \otimes A)\cdots \\ &\qquad (I_N \otimes C_{y_{j+1}})(I_N \otimes A)]\hat{\Pi}_{j,j|j} \\ &= \frac{1}{\Theta_{j,j+\Delta}}[I_N \otimes (1'_N U_{j+1,j+\Delta})]\hat{\Pi}_{j,j|j} \end{aligned} \tag{13}$$

where $\Theta_{j,j+\Delta} = 1'_{N^2}[I_N \otimes (1'_N U_{j+1,j+\Delta})]\hat{\Pi}_{j,j|j}$. The above derivation uses the property $(A \otimes B)(C \otimes D) = AC \otimes BD$ with

$$U_{j+1,j+\Delta} = C_{y_{j+\Delta}}AC_{y_{j+\Delta-1}}A\cdots C_{y_{j+1}}A \tag{14}$$

and

$$\begin{aligned}\hat{\Pi}_{j,j|j} &= \begin{bmatrix} \Pr(X_j = 1, X_j = 1|Y_0, \cdots, Y_j) \\ \Pr(X_j = 1, X_j = 2|Y_0, \cdots, Y_j) \\ \vdots \\ \Pr(X_j = 2, X_j = 1|Y_0, \cdots, Y_j) \\ \vdots \\ \Pr(X_j = N, X_j = N-1|Y_0, \cdots, Y_j) \\ \Pr(X_j = N, X_j = N|Y_0, \cdots, Y_j) \end{bmatrix} \\ &= \begin{bmatrix} \Pi_{j|j}(1) \\ 0 \\ \vdots \\ 0 \\ \Pi_{j|j}(2) \\ 0 \\ \vdots \\ 0 \\ \Pi_{j|j}(3) \\ \vdots \\ \Pi_{j|j}(N) \end{bmatrix} \begin{matrix} \\ \\ \\ \left.\begin{matrix} \\ \\ \end{matrix}\right\} N \text{ zeros} \\ \\ \left.\begin{matrix} \\ \\ \end{matrix}\right\} N \text{ zeros} \\ \\ \\ \end{matrix}\end{aligned} \tag{15}$$

where $\Pi_{j|j}(i)$ denotes the ith entry in $\Pi_{j|j}$.

Remark 3.1: Although (13) has been obtained from an augmented system, which is of dimension N^2, it can be seen by some straightforward algebraic manipulations that each entry of the smoothed estimates can in fact be calculated from products of $N \times N$ matrices. That is, each component of $\Pi_{j|j+\Delta}$ can be written as

$$\Pi_{j|j+\Delta}(i) = 1'_N U_{j+1,j+\Delta}[e_i\Pi_{j|j}(i)]$$

where e_i is an $N \times 1$ vector of zero's, with a one in the ith position. This means that the additional computations required to obtain the smoothed estimates, once the filtered estimates have been evaluated (via the "initial" conditions of the smoother), is of the order of $O(\Delta N^2)$, taking into account the extra number of operations required. Note that for the purpose of this argument, we have regarded additions and multiplications to be equivalent operations.

As can be seen from the above, (14) has the same structure as (4). Consequently, with j fixed, as $\Delta \to \infty$

$$U_{j+1,j+\Delta} \to \boldsymbol{\alpha}(\Delta)\boldsymbol{\beta}' \tag{16}$$

for some real positive column vectors $\boldsymbol{\alpha}(\Delta)$ and $\boldsymbol{\beta}$. Furthermore, the same rate of convergence as for the filter determines how fast $U_{j+1,j+\Delta}$ becomes dyadic.

Remark 3.2: Disregarding the dependence on $j, \boldsymbol{\alpha}(\Delta) = \{\alpha_i(\Delta)\}$ is a Δ-dependent term, whereas $\boldsymbol{\beta} = \{\beta_i\}$ is a vector of constants $\forall i \in \{1, 2, \cdots, N\}$.

Since, by inspection of (13), as $\Delta \to \infty$

$$1'_N U_{j+1,j+\Delta} \to \left[\sum_{i=1}^{N} \alpha_i(\Delta)\right]\boldsymbol{\beta}'$$

and therefore, the unnormalized smoothed probability vector becomes

$$\begin{aligned}\tilde{\Pi}_{j|j+\Delta} &= \left[I_N \otimes \left(\sum_{i=1}^{N} \alpha_i(\Delta)\right)\boldsymbol{\beta}'\right]\hat{\Pi}_{j,j|j} \\ &= \begin{bmatrix}\beta_1 \cdots \beta_N & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \beta_1 \cdots \beta_N\end{bmatrix}\hat{\Pi}_{j,j|j} \\ &\quad \cdot \left[\sum_{i=1}^{N} \alpha_i(\Delta)\right] \\ &= \left[\sum_{i=1}^{N} \alpha_i(\Delta)\right]\begin{bmatrix}\beta_1 \Pi_{j|j}(1) \\ \beta_2 \Pi_{j|j}(2) \\ \vdots \\ \beta_N \Pi_{j|j}(N)\end{bmatrix}.\end{aligned}$$

Once normalized, it can be seen that the Δ-dependent terms are cancelled, thus establishing the Δ independence of the smoothing equation for large lags or Δ's. This means that all the improvement that can be gained by smoothing is attained after some finite Δ, and no significant gains can be made as the lag is extended further.

From (13), it is evident that the inhomogeneous product $U_{j+1,j+\Delta}$ occurs in both the numerator and the denominator of the smoothed estimates $\Pi_{j|j+\Delta}$. If $U_{j+1,j+\Delta}$ achieves a rank of 1 when $\Delta \to \infty$ at an exponential rate, then the ratio of two quantities that involve this product also converges at the same rate. Consequently, the smoothing lag independence property must occur at the same rate as the convergence rate of $U_{j+1,j+\Delta}$ and is computable as discussed earlier in Section II-D.

B. Smoother from Forward-Backward Filters

1) Backward Model: Following ideas in [3], an analogous backward Markov state process and HMM, which evolves backward in time, can be constructed from a forward Markov state process.

Definition 3.2: For a given forward Markov state process characterized by a state transition probability matrix A, the associated backward process has a system matrix A^b with elements a_{ij}^b such that

$$\begin{aligned}a_{ij}^b &= \Pr(X_k = i | X_{k+1} = j) \\ &= \Pr(X_k = i)\Pr(X_{k+1} = j | X_k = i) \\ &\quad \cdot [\Pr(X_{k+1} = j)]^{-1}.\end{aligned}$$

In other words, and further assuming stationarity so that $\Pr(X_k = i) = \Pr(X_{k+1} = i) = \Pr(X = i)$

$$A^b = \Lambda A' \Lambda^{-1}$$

where $\Lambda = \operatorname{diag}(\Pi)$, and Π is the steady-state distribution such that $A\Pi = \Pi$.

2) Smoother from Forward-Backward Model:

Definition 3.3: Denote a reverse filter probability and the associated analog of the one-step prediction probability vector by $\Pi_{k|k}^+$ and $\Pi_{k|k+1}^+$, respectively, with the ith entry being $\Pr(X_k = i | Y_k, Y_{k+1}, \cdots, Y_{k+\Delta})$ and $\Pr(X_k = i | Y_{k+1}, Y_{k+1}, \cdots, Y_{k+\Delta})$, where $k+\Delta$ is the maximum time at which measurements are available.

In accordance with Definition 3.3, $\Pi_{k+\Delta|k+\Delta}^+ = \{\Pr(X_{k+\Delta} = i | Y_{k+\Delta})\}$, and $\Pi_{k+\Delta|k+\Delta+1}^+ = \{\Pr(X_{k+\Delta} = i)\}$ since $k+\Delta$ is the maximum time at which measurements are available. By further assuming stationarity, it can be seen that

$$\Pi_{k+\Delta|k+\Delta+1}^+ = \{\Pr(X = i)\}.$$

The reverse-time filter obeys the updating equations

$$\Pi_{k|k+1}^+ = A^b \Pi_{k+1|k+1}^+ \tag{17}$$

$$\Pi_{k|k}^+ = \frac{1}{1_N' C_{y_k} \Pi_{k|k+1}^+} C_{y_k} \Pi_{k|k+1}^+. \tag{18}$$

By using Bayes' rule (and the fact that $(Y_0, Y_1, \cdots, Y_j)$ and $(Y_{j+1}, Y_{j+2}, \cdots, Y_{j+\Delta})$ are conditionally independent given X_j; see [5]), the smoothed probability vector can be expressed as a product of two terms

$$\begin{aligned}&\Pr(X_j = i | Y_0, \cdots, Y_{j+\Delta}) \\ &\quad = \frac{1}{\Theta \Pr(X_j = i)} \Pr(X_j = i | Y_0, Y_1 \cdots, Y_j) \\ &\qquad \cdot \Pr(X_j = i | Y_{j+1}, Y_{j+2}, \cdots, Y_{j+\Delta}) \\ &\quad = \frac{1}{\Theta \Pr(X = i)} \Pr(X_j = i | Y_0, Y_1, \cdots, Y_j) \\ &\qquad \cdot \Pr(X_j = i | Y_{j+1}, Y_{j+2}, \cdots, Y_{j+\Delta})\end{aligned} \tag{19}$$

where Θ is a normalizing constant utilizing stationarity in the last line.

Remark 3.3: Note, in particular, the structure of the smoothing equation (19). The first term in the numerator is a forward filter, where the second is a one-step prediction reverse-time filter. The denominator is the ith term of Π.

From (17)–(19), the i entry of the smoothed probability vector at time j becomes

$$\begin{aligned}&\Theta[\Pi_{j|j+\Delta}]_{i\text{th entry}} \\ &\quad = \frac{1}{\Pr(X = i)}[\Pi_{j|j}]_{i\text{th entry}}[\Pi_{j+|j+1}]_{i\text{th entry}} \\ &\quad = \frac{1}{\Pr(X = i)}[\Pi_{j|j}]_{i\text{th entry}}[A^b C_{y_{j+1}} A^b C_{y_{j+2}} \cdots \\ &\qquad A^b C_{y_{j+\Delta}} \Pi_{j+\Delta|j+\Delta+1}^+]_{i\text{th entry}} \\ &\quad = \frac{1}{\Pr(X = i)}[\Pi_{j|j}]_{i\text{th entry}} \\ &\qquad \cdot [T_{j+1,j+\Delta}^b \Pi_{j+\Delta|j+\Delta+1}^+]_{i\text{th entry}}.\end{aligned} \tag{20}$$

The Δ dependence lies entirely in $T_{j+1,j+\Delta}^b$; the apparent Δ dependence in $\Pi_{j+\Delta|j+\Delta+1}^+$ will be lost, as mentioned earlier, due to the initializing process of the reverse filter. Now, since A^b and C both are strictly positive, the previous exponential convergence results can be similarly applied. However, $T_{j+1,j+\Delta}^b$ involves the product of terms like $A^b C_y$

(and therefore $A'C_y$ by Definition 3.2); hence, as $\Delta \to \infty$

$$\begin{aligned} T^b_{j+1,j+\Delta} &= A^b C_{y_{j+1}} A^b C_{y_{j+2}} \cdots A^b C_{y_{j+\Delta}} \\ &= (\Lambda A' \Lambda^{-1}) C_{y_{j+1}} (\Lambda A' \Lambda^{-1}) C_{y_{j+2}} \cdots \\ &\qquad (\Lambda A' \Lambda^{-1}) C_{y_{j+\Delta}} \\ &= \Lambda (A' C_{y_{j+1}} A' C_{y_{j+2}} \cdots A' C_{y_{j+\Delta}}) \Lambda^{-1} \\ &= \Lambda (C_{y_{j+\Delta}} A C_{y_{j+\Delta-1}} A \cdots C_{y_{j+1}} A)' \Lambda^{-1} \\ &\to \Lambda \boldsymbol{\mu} \boldsymbol{\nu}' (\Delta) \Lambda^{-1} \\ &\equiv \boldsymbol{\mu}^b \boldsymbol{\nu}^{b\prime} (\Delta). \end{aligned}$$

The third equality follows from the commutativity of diagonal matrices. By the same arguments in Section III-A, the Δ-dependency in the smoothed estimate (20) can again be removed after normalization. Essentially, this can be seen by recognizing that $\boldsymbol{\nu}^{b\prime}(\Delta)\Pi^+_{j+\Delta|j+\Delta+1}$ is a scalar and will be cancelled after normalization.

Remark 3.4: In both Sections III-A and B, we have shown a Δ independence in the smoothed estimates as $\Delta \to \infty$. The precise equivalence of the two smoothed estimates for any Δ is deferred to Appendix B. However, although the two approaches are numerically equivalent, the augmented state model algorithm has some practical advantages over the method using combined forward and backward models. Due to the recursive nature of (11) and (12), a calculation of $\Pi_{j|j+\Delta}$ from $\hat{\Pi}_{j,j+\Delta|j+\Delta}$ requires the product $U_{j+1,j+\Delta} = C_{y_{j+\Delta}} A U_{j+1,j+\Delta-1} = C_{y_{j+\Delta}} A C_{y_{j+\Delta-1}} A U_{j+1,j+\Delta-2}$, etc., which means that $\Pi_{j|j+\Delta-1}, \Pi_{j|j+\Delta-2}, \cdots$ are readily available. This is not so easily achieved using (19), as each smoothed estimate is the product of two disjoint terms.

IV. Extensions to Systems with Non-Negative A, C

In this section, we will present some modifications to the theory for positive systems, as summarized in Appendix A, to incorporate the cases where A or C, but not both, may be non-negative. Such situations reflect natural systems in which some states may not be accessible to all others (zeros in A) or that certain classes of observations and states are mutually exclusive. The case of $A \geq 0$ and $C_{y_k} \geq 0$ with both $A > 0$ and $C > 0$ failing is more complicated and will be analyzed elsewhere.

A. $A \geq 0, C > 0$, and $A > 0$ Fails

For such combinations of A and C, it is obvious that $C_{y_k} A \geq 0 \ \forall k$. However, the previous convergence result is still applicable, provided a product of successive $C_y A$'s can result in a positive matrix in some finite time k, even though one or more individual inequalities $C_{y_k} A > 0$ fail. This is possible because the disposition of positive terms in a product of two non-negative matrices is independent of the magnitude of the nonzero elements in the components of the product. Further, $C_y A$ has the same disposition of nonzero elements as A. It follows that a product of r successive $C_y A$ is positive if and only if A^r is positive, which will be the case for some r if and only if A is primitive.[3]

[3] Primitivity means that $A^k > 0$ for some integer k.

B. $A > 0, C_{y_k} \geq 0$ for One or More y_k, and $C_{y_k} > 0$ Fails

To account for such cases, we need to modify the application of the metric in (21). In essence, our method consists of evaluating only the nonzero entries of $x'T$ and $y'T$ in $d(x'T, y'T)$ for $x, y > 0$, but $T \geq 0$. A contraction is then established based on the fact that the computations are over a subset of the original. For example, given an $(n \times n)$ matrix T, which has $n - m$ rows of zeros but the remaining entries positive, let A be the $(m \times n)$ matrix formed by excluding the zero rows. Using (22) and (23), $\phi(A)$ and, hence, $\tau_B(A)$, instead of $\tau_B(T)$, can be found.

Now, suppose T is $(n \times n)$ and strictly positive. Let Σ be a diagonal matrix with diagonal entries of $+1$ and zeros; then, ΣT is a matrix with some rows containing all zero entries, where the remaining positive entries are identical with those of T. By using (24), observe that

$$\phi(\Sigma T) \geq \phi(T)$$

since the minimization in computing $\phi(\Sigma T)$ is over a subset of values that are used in computing $\phi(T)$. Consequently, from (23)

$$\tau_B(\Sigma T) \leq \tau_B(T).$$

If we now replace Σ with a C_{y_k} with some zero diagonal entries and T by A, then the inequality in (6) again follows.

V. Simulations of a Two-State HMM

The basic system investigated is a two-state HMM with a symmetric A matrix (which is also known as random telegraph wave). Similar simulations had been carried out in [7] but as a discrete-time approximation to the continuous-time model discussed therein. In our case, the starting point is the discrete-time filtering and smoothing equations from Sections II and III.

The simulations aim to illustrate

- that noticeable improvement in estimates can be achieved by using smoothing as opposed to filtering;
- the dependence of the extent of improvement of smoothing over filtering on the smoothing lag;
- the existence of a finite lag, which if exceeded conveys no further practical benefit.

We demonstrate these points by varying the smoothing lag and performing the simulations in environments with different noise contents. Accordingly, we proceed by first keeping A constant and varying C between an almost zero-noise (C being diagonally dominant under row permutation) to totally noise-corrupted systems (C having identical entries). This is then repeated by varying A and keeping C constant. In both cases, the base and unmodified system consists of $A = \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix}, C = \begin{bmatrix} 0.1 & 0.8 \\ 0.9 & 0.2 \end{bmatrix}$.

Since we did not use continuous state or observation spaces, the conventional definition of noise does not apply here. We can regard an HMM as a generation of a random signal, namely, the state process X_k, and contaminating it by measurement noise to produce the output process Y_k. Observe that if a particular measurement Y_k is to yield valuable information regarding the state X_k, then there should be a high conditional

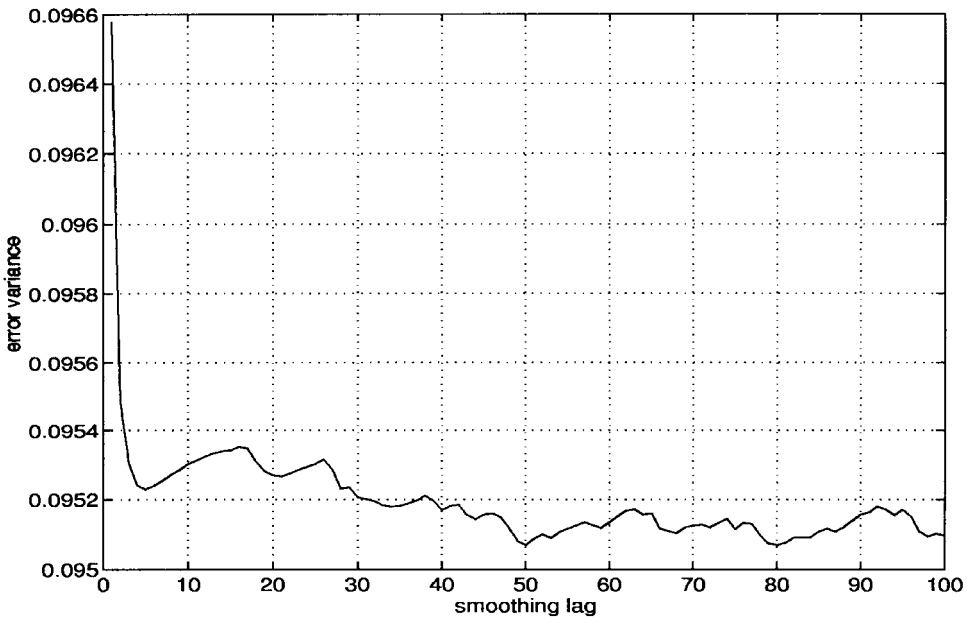

Fig. 1. Variation of smoothing error with Δ: A fixed, C_1. Filter error 0.1094.

probability relating the measurement to one particular value of the state, i.e., for each m there exists an n for which $\Pr(Y_k = m | X_k = n)$ is high. This corresponds to a relatively "noise-free" C. However, if C has identical entries in any one row, a particular value of the measurement is equiprobable no matter what the state of the HMM is; hence, little information can be inferred from it. This corresponds intuitively to the concept of noise-corrupted measurements.

On the other hand, there is quite a different interpretation for values in entries of A (as opposed to similar variations in entries of C) since A determines the correlation between the X process at different times. When A has a particular dominant element in a given row, say, element a_{ij}, then it can be seen that $\Pr(X_{k+1} = i)$ will be strongly correlated with $\Pr(X_k = j)$; hence, there is a strong dependence between X_{k+1} and X_k. This means that better state estimates are possible if more state information is available (e.g., smoothing).

However, when A has identical entries, at any given time k, all the state transitions are equally likely (see also Remark 2.2); hence, X_{k+1} is essentially independent of X_k, i.e., the X process is a white process. In this instance, smoothing offers no advantage overfiltering (see Fig. 9), and the significant factor in the quality of the estimates is the noise content in the measurements. Both filtering and smoothing are conditional estimation techniques given the measurements, which, in turn, are probabilistically related to the actual states. Consequently, if X_k is independent of $X_{k-i}, 1 \le i \le k$ (and similarly $X_{k+1}, X_{k+2}, \cdots$, etc.), then Y_{k-1} or Y_{k+1} contain no extra information regarding X_k; hence, any estimates of X_k will be based entirely on Y_k and, thus, dependent on the correlation between Y_k and X_k.

Just as for the filtering error variance, the smoothing error variance (see Appendix C for the definition of error variance) is based on the difference between the smoothed and the true state probability vectors. The maximum error given no measurements at all is $\frac{1}{2}(1 - \Pi'\Pi)$, where Π is the steady-state distribution, and the factor of 1/2 is introduced for normalization. For the special case of A being symmetric, the error variance with estimation based entirely on Π, and no measurements (hence no filtering or smoothing) is 0.25.

By using (25), the lag required for the smoothing error variance to reach steady state can be estimated, specializing to the case of $p, k_0 = 0$ and $g = 1$, so that

$$\tau_B(U_{j+1,j+\Delta}) \le \left(\frac{1-\epsilon}{1+\epsilon}\right)^{\Delta-1} = \tau_B(U_{j+1,j+\Delta})_{\text{ob}}^{\Delta-1}$$

where $[\tau_B(U_{j+1,j+\Delta})_{\text{ob}}]^{\Delta-1}$ provides an overbound on $\tau_B(U_{j+1,j+\Delta})$, and $\phi(\tau_B(U_{j+1,j+\Delta})_{\text{ob}}) = \epsilon^2$. For our simulation studies, in obtaining $\tau_B(U_{j+1,j+\Delta})_{\text{ob}}$, both the approximate bound (6) and the three-term average (10) have been used. The criteria for convergence, or $U_{j+1,j+\Delta}$ having rank 1, is to determine $\Delta = \Delta_{\text{crit}}$ such that $\tau_B(\cdot)_{\text{ob}}^{\Delta_{\text{crit}}/4} = 1/e$. The associated overbounds $\tau_B(\cdot)_{\text{ob}}$ and the expected lag before convergence have been listed in Table I. Note that for a filter, Δ_{crit} is also equivalent to the number of time steps before a given set of initial conditions can be forgotten.

A. A Constant, C Variable

In this series of simulations, the various C matrices are

$$C_1 = \begin{bmatrix} 0.1 & 0.8 \\ 0.9 & 0.2 \end{bmatrix}, \quad C_2 = \begin{bmatrix} 0.3 & 0.6 \\ 0.7 & 0.4 \end{bmatrix}$$
$$C_3 = \begin{bmatrix} 0.45 & 0.4 \\ 0.55 & 0.6 \end{bmatrix}, \quad C_4 = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}.$$

In each of Figs. 1–3, there is rapid initial improvement, compared with filtering, in the smoothing error, which then reaches some steady-state value. As seen in Figs. 1–3, this change usually occurs at $\Delta \approx 5$. The rate of convergence with Δ of the smoothing error[4] is relatively insensitive to

[4]The same arguments apply equally to the rate of forgetting of initial conditions for filters.

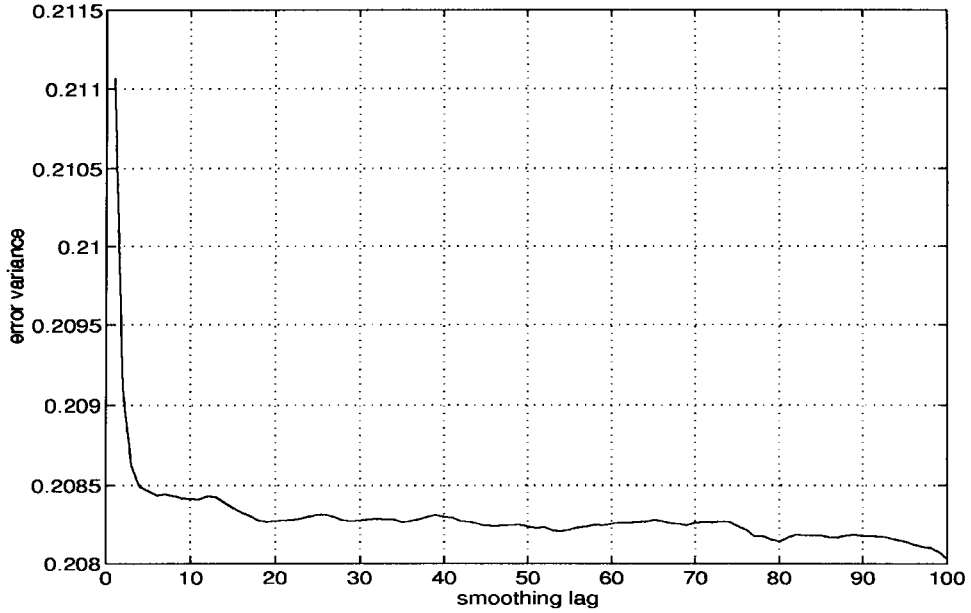

Fig. 2. Variation of smoothing error with Δ: A fixed, C_2. Filter error 0.2182.

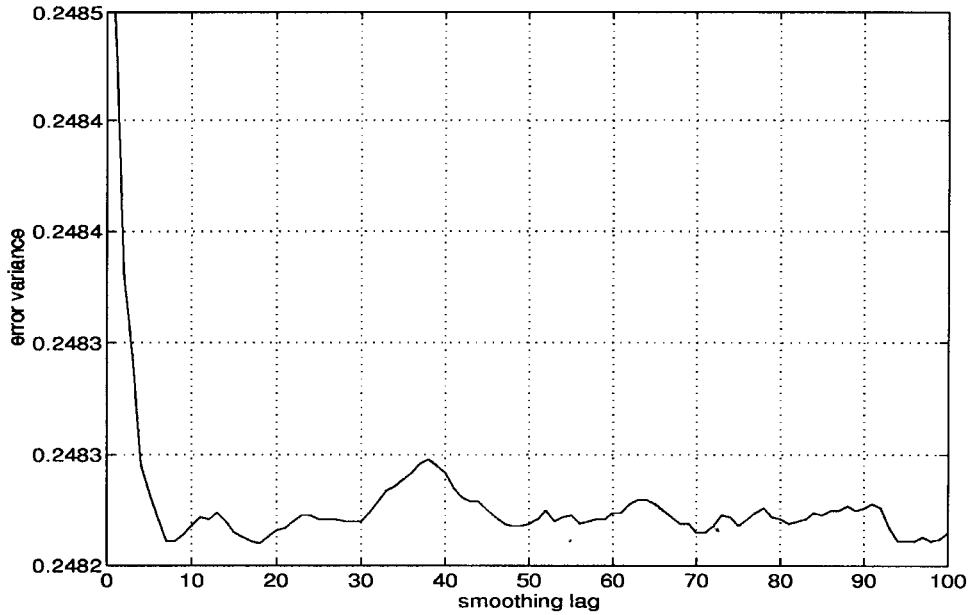

Fig. 3. Variation of smoothing error with Δ: A fixed, C_3. Filter error 0.2487.

variations in C, in accordance with the fact that $\tau_B(\cdot)$ of a diagonal matrix is 1. From Table I, it can be seen that Δ_{crit} provides an adequate, albeit slightly pessimistic, bound to the maximum Δ before convergence. However, since for the systems studied, convergence occurs within relatively small lags, it is probable that a particular output sequence might not exhibit the average behavior and thus converge to a steady-state value at much larger Δ's, although this was not observed in our simulations.

As C becomes "whiter," the magnitude of the errors for both the filtered and smoothed estimates increases and approaches the maximum value of 0.25 (Fig. 4). This means that in environments with high measurement noise, it is difficult to obtain better estimates by either filtering or smoothing than estimating via steady-state distributions alone.

TABLE I
OVERBOUNDS ON $\tau_B(T_{j+1,j+\Delta})$ AND THE EXPECTED SMOOTHING LAG BEFORE CONVERGENCE

		$\tau_B(T_{j+1,j+\Delta})_{\text{ob}}$		Δ_{crit}	
		approx. bound	3-term av.	approx. bound	3-term av.
A fixed,	C_1	0.8	0.729	17.9	12.7
	C_2	0.8	0.792	17.9	17.1
	C_3	0.8	0.800	17.9	17.9
	C_4	0.8	0.8	17.9	17.9
C fixed,	A_1	0.9	0.861	38.0	26.6
	A_2	0.8	0.729	17.9	12.7
	A_3	0.4	0.304	4.37	3.36
	A_4	0.2	0.144	2.49	2.06
	A_5	0	0	0	0

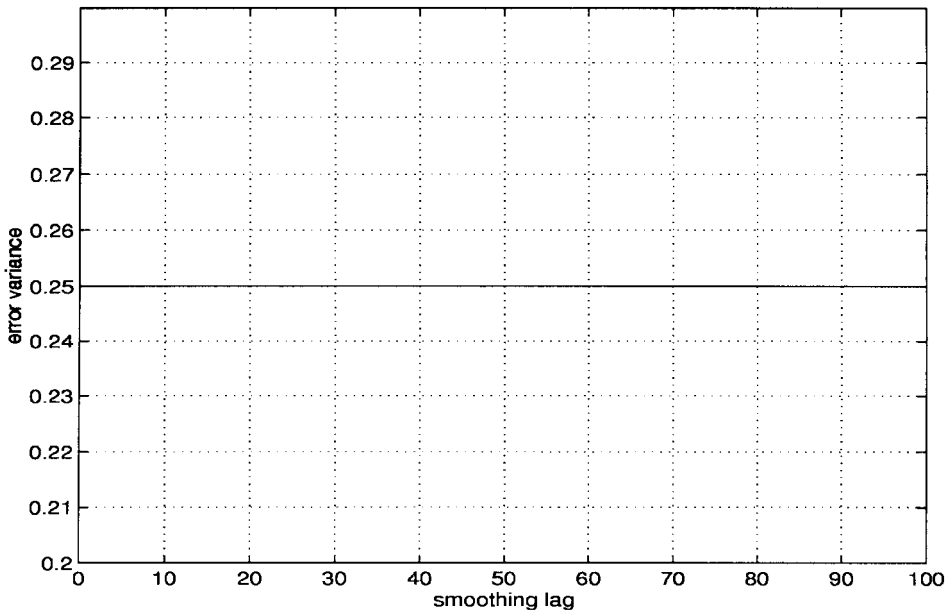

Fig. 4. Variation of smoothing error with Δ: A fixed, C_4. Filter error 0.25.

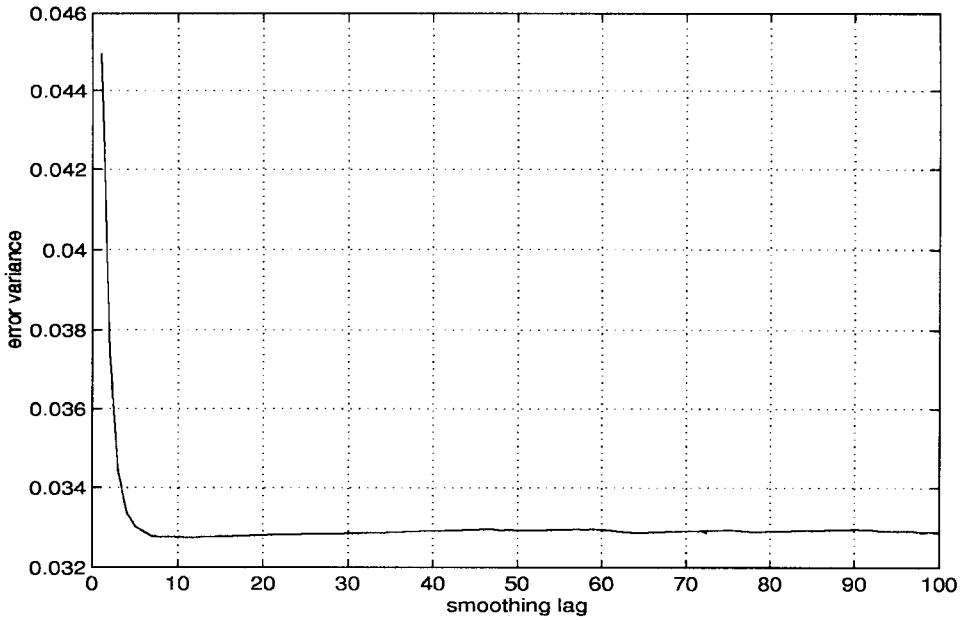

Fig. 5. Variation of smoothing error with Δ: C fixed, A_1; filter error: 0.0644.

B. A Variable, C Fixed

For this series of simulations, we intend to illustrate that as the entries in A become more alike (i.e., close to defining a white process, with all entries being 0.5), there is less improvement to be gained from smoothing over filtering. This is related to the degree of dependency between successive X_k's alluded to previously. The A matrices

$$A_1 = \begin{bmatrix} 0.95 & 0.05 \\ 0.05 & 0.95 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix}$$

$$A_3 = \begin{bmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{bmatrix}, \quad A_4 = \begin{bmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{bmatrix}$$

$$A_5 = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$$

have been used.

Similar to the previous set of simulations, there is noticeable improvement by smoothing over filtering, as is evident from the reduced error variance for smoothing; see Figs. 5 and 6. However, as A is varied from A_1 (an almost deterministic system) through to A_5 (a white process), it can be seen from Figs. 5–9 that the minimum Δ required for steady-state actually decreases until any gain/loss by smoothing is almost instantaneous (Fig. 9). Note that an increased smoothing lag has been shown in Fig. 9 to emphasize the fluctuating behavior of the smoothing error. This is also confirmed by the decrease in Δ_{crit} (see Table I) for the case of A_5, which is essentially an indication of how readily previous states are forgotten.

VI. Conclusions

The exponential stability property present in most Kalman filters is desirable for not only academic interest but also

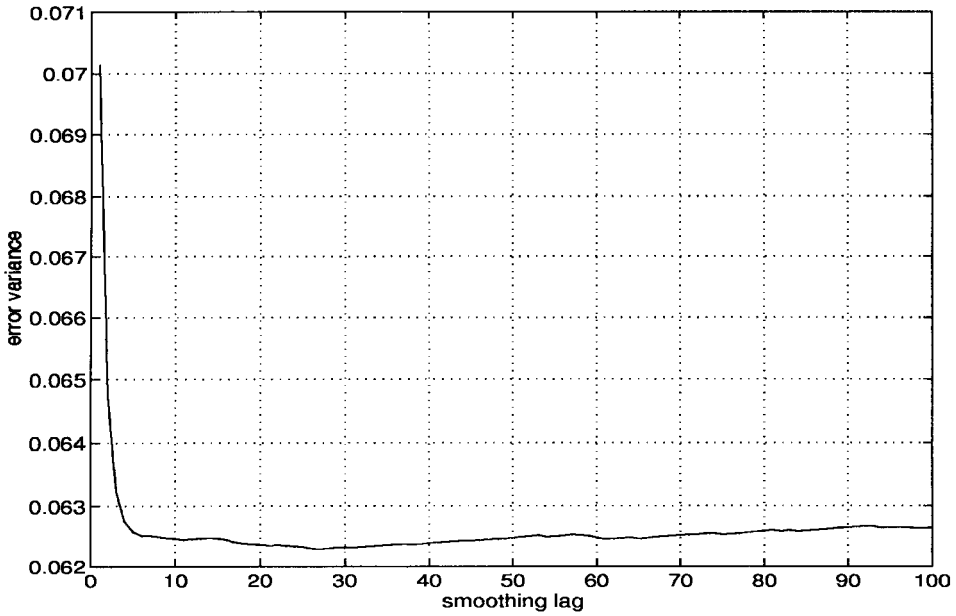

Fig. 6. Variation of smoothing error with Δ: C fixed, A_2. Filter error 0.0906.

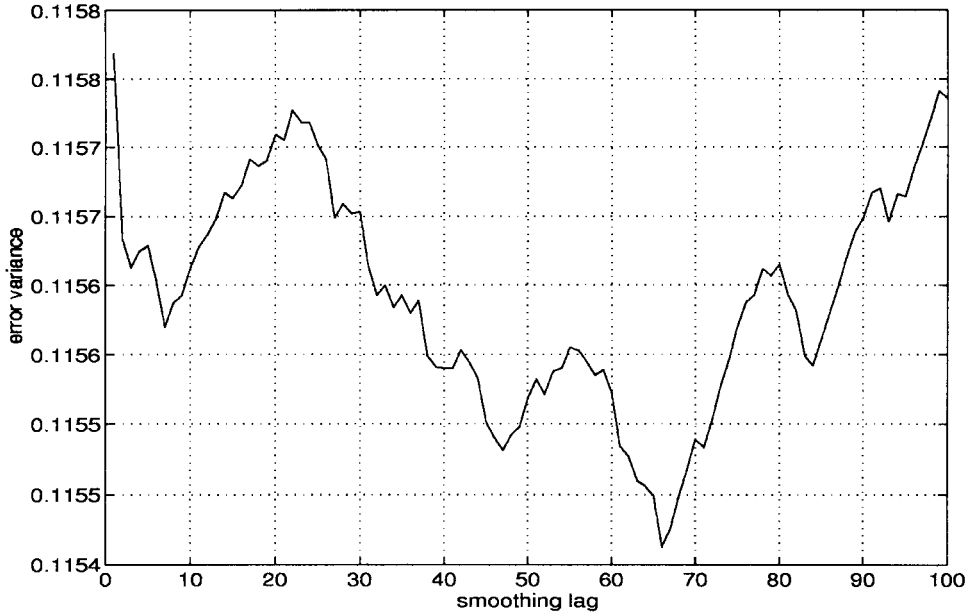

Fig. 7. Variation of smoothing error with Δ: C fixed, A_3. Filter error 0.1214.

for practical reasons. By extending the exponential forgetting results to HMM's, we have shown that similar properties exist for the recursive filters and fixed-lag smoothers for the class of HMM's studied in this paper. In particular, by appealing to results for the product of non-negative matrices [12], it is seen that the forgetting of initial conditions in filters is mirrored in the Δ independence of the fixed-lag smoothers; with suitable modifications, similar results apply to systems with either, but not both, A or C non-negative.

By judicious interpretation of the parameters in the system matrices A and C for an HMM, we are able to qualitatively account for the behavior of smoothing errors under different noise environments. As indicated in Section V and confirmed by simulations (Figs. 1–4), the improvement of smoothing over filtering is greatest when C is almost noise free, where the convergence rate with Δ is relatively independent of the C matrix. When A is varied so that long-term dependence between state values is diminished, it is seen that the advantage of smoothing over filtering diminishes significantly. In both cases, the theoretical overbound $\tau_B(\cdot)_{\rm ob}$ seems to provide bounds that, on occasion, are rather conservative on the minimum Δ required for convergence. Although the present method of estimating $\Delta_{\rm crit}$ has not been sufficiently tested, it is envisaged that such theoretical bounds can assist in the choice of certain parameters, such as the smoothing lag, in the implementation of fixed-lag smoothers.

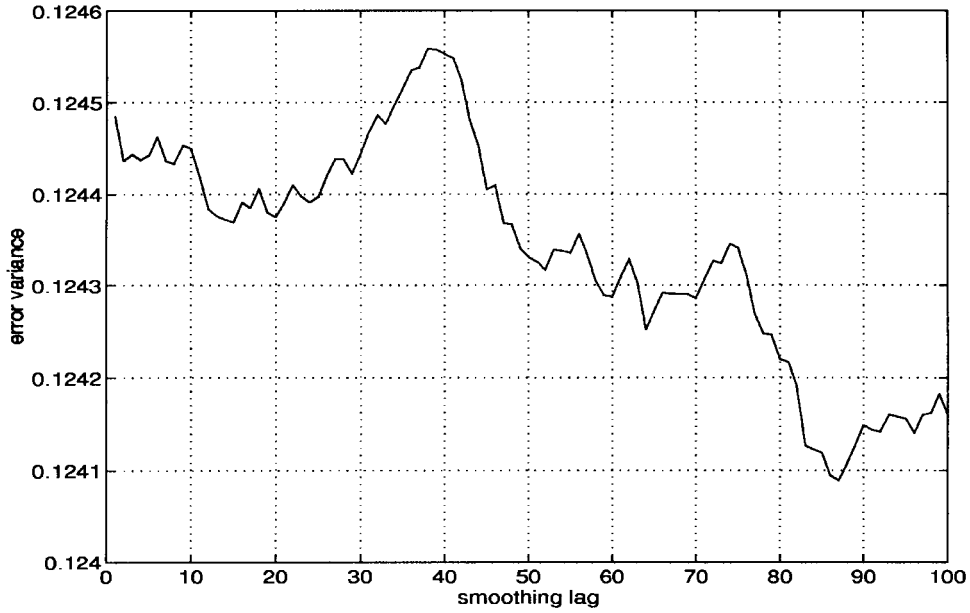

Fig. 8. Variation of smoothing error with Δ: C fixed, A_4. Filter error 0.1258.

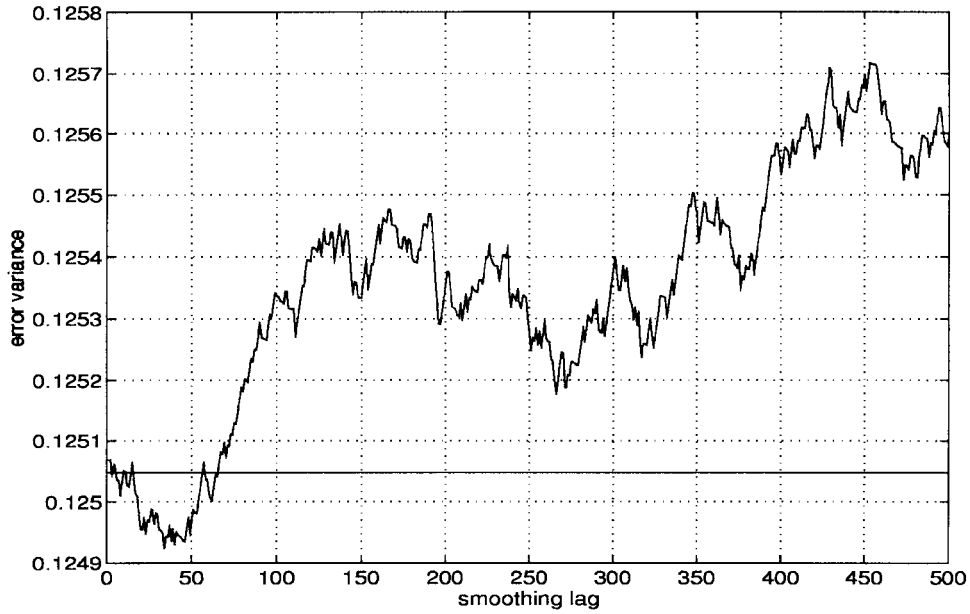

Fig. 9. Variation of smoothing error with Δ: C fixed, A_5. Filter error: 0.1250. Note that the filtering error has also been included.

Systems with both A and C both being non-negative remain incompletely treated and will be tackled in future. Suffice it to say that for such systems, in some situations, the initial conditions are not forgotten, even though the probability of such occurrences may be low. Natural extensions of the present work include investigations of continuous-time models and continuous-state/output HMM's, which makes possible direct comparisons between the performance of the filters/smoothers with SNR of the system. However, for continuous-time systems, an analog of the limiting theorem for a product of non-negative matrices is required. Other indirect applications include the Viterbi algorithm, where the concept of a Δ_{crit} may help to guide the choice of truncation point, and hybrid systems, which involve HMM's such as those arising from, for example, multitarget tracking.

APPENDIX A
GENERALIZATIONS OF THE PERRON–FROBENIUS THEOREM FOR INHOMOGENEOUS PRODUCT OF MATRICES

In this Appendix, we summarize certain key ideas relating to products of non-negative matrices and positive matrices from [12].

Definition A.1: An $(n \times n)$ matrix $T \geq 0$ is said to be row allowable if it has at least one positive entry in each row. It is said to be column allowable if T' is row allowable. It is said to be allowable if it is both column and row allowable.

Remark A.1: A column allowable T has the property that $x>0$ implies $x'T>0$.

Definition A.2: For two vectors $x' = (x_1 \cdots x_n) > 0$ and $y' = (y_1 \cdots y_n) > 0$, a pseudo metric[5] can be defined as

$$d(x', y) = \ln \left[\frac{\max_i (x_i/y_i)}{\min_i (x_i/y_i)} \right] = \max_{i,j} \ln \left[\frac{x_i y_j}{x_j y_i}. \right] \tag{21}$$

This is a measure of the alignment between two given vectors x and y and $d(x', y') = 0$ iff $x = \lambda y$ for some scalar $\lambda > 0$.

For $x, y > 0$ and column allowable T (to ensure $x'T > 0$ and $y'T > 0$) the function $d(\cdot, \cdot)$ has the following contraction properties:

1) $d(x'T, y'T) \leq d(x', y')$.
2) Given a T that also has at least one positive element in a coincident position in any two rows, $d(x'T, y'T) < d(x', y')$. This guarantees that for $x, y > 0$, multiplication by a strictly positive T always tends to align the two vectors x and y.

Remark A.2: A direct consequence of above properties is that under certain conditions, as $k \to \infty$

$$d(x'T_1T_2 \cdots T_k, y'T_1T_2 \cdots T_k) \to 0$$

where each T_i is allowable, $1 \leq i \leq k$, provided $x, y > 0$.

Definition A.3: The Birkhoff's contraction coefficient is defined as

$$\tau_B(T) = \sup_{\substack{x,y>0 \\ x \neq \lambda y}} \frac{d(x'T, y'T)}{d(x', y')}. \tag{22}$$

Birkhoff's contraction coefficient places an upper bound on the rate of contraction due to the multiplication with an allowable matrix T by providing a ratio between the pre and postmultiplied metric of two vectors by T. As will be shown below, the case when $\tau_B(\cdot) = 0$ is of special interest.

Remark A.3: For a column allowable matrix T, an explicit form (for the long derivations, see [12]) for $\tau_B(T)$ is

$$\tau_B(T) = \frac{1 - \sqrt{\phi(T)}}{1 + \sqrt{\phi(T)}} \tag{23}$$

where

$$\phi(T) = \begin{cases} \min_{i,j,k,l} \dfrac{t_{ik} t_{jl}}{t_{jk} t_{il}}, & \text{if } T > 0 \\ 0, & \text{if } T \not> 0. \end{cases} \tag{24}$$

If T is column allowable but not row allowable, let $A = \{a_{ij}\}$ be the matrix formed by deleting any row of zeros so that A is row allowable, then $\phi(A) = \phi(T)$.

For column allowable T and U, $\tau_B(\cdot)$ has the following properties.

1) $0 \leq \tau_B(T) \leq 1$.
2) $\tau_B(TU) \leq \tau_B(T)\tau_B(U)$; hence, $\tau_B(TU) \leq \tau_B(T)$, and $\tau_B(TU) \leq \tau_B(U)$.
3) For a positive diagonal matrix U, $\tau_B(U) = 1$.
4) $\tau_B(U) = 0$ iff U is also rank 1.

[5] A pseudo metric has the properties of a metric except that $d(x', y') = 0$ can occur even if $x \neq y$.

Definition A.4: For the sequence of non-negative allowable matrices $H_k = \{h_{ij}(k)\}, k \geq 1$, we define the forward and reverse products as

$$\text{Forward product} \quad T_{p,r} = H_{p+1} H_{p+2} \cdots H_{p+r}$$
$$\text{Reverse product} \quad U_{p,r} = H_{p+r} \cdots H_{p+2} H_{p+1}.$$

In the subsequent discussion, we shall denote $H_{p,r}$ as either of the preceding products, where $p \geq 0, r \geq 1$.

Definition A.5: The products $H_{p,r}$ are said to be weakly ergodic if, as $r \to \infty$, $H_{p,r}$ approaches a rank 1 matrix or, equivalently, as $r \to \infty, \tau_B(H_{p,r}) \to 0$.

With the above definitions in mind, the next theorem follows as a straightforward combination of [12, Th. 3.3 and Lemma 3.4].

Theorem A.1: If for the sequence of non-negative allowable matrices $H_k = \{h_{ij}(k)\}, k \geq 1$

1) $H_{p,r_0} > 0$, for all $p \geq 0$, where $r_0 \geq 1$ is some fixed integer independent of p;
2) $\min^+_{i,j} h_{ij}(k) / \max_{i,j} h_{ij}(k) \geq \gamma > 0$

(where $\min^+$ refers to the minimum of the positive elements, and γ is independent of k), then if $H_{p,r} = T_{p,r}$ (or $= U_{p,r}$), $p \geq 0, r \geq 1$, weak ergodicity (at a geometric rate) is obtained.

Further, if $H_{p,r} = T_{p,r}$, then $t_{ik}^{(p,r)} / t_{jk}^{(p,r)} \to W_{ij}^{(p)} > 0$ for all i, j, p, k, where $t_{ik}^{(p,r)}$ denotes the (i,k)th element of the product $T_{p,r}$, and the limit is independent of k. That is, as $r \to \infty$, the rows of $T_{p,r}$ tend to proportionality. Equivalently, as $r \to \infty, T_{p,r} \to \boldsymbol{\mu}\boldsymbol{\nu}'(r)$, with $\boldsymbol{\mu}, \boldsymbol{\nu}(r)$ positive real column vectors, and $\mu_i/\mu_j = W_{ij}^{(p)}$.

If $H_{p,r} = U_{p,r}$, then $u_{ki}^{(p,r)} / u_{kj}^{(p,r)} \to V_{ij}^{(p)} > 0$ for all i, j, p, k, where $u_{ki}^{(p,r)}$ denotes the (k,i)th element of the product $U_{p,r}$, and the limit is independent of k. That is, as $r \to \infty$, the columns of $U_{p,r}$ tend to proportionality, or $U_{p,r} \to \boldsymbol{\alpha}(r)\boldsymbol{\beta}'$, with $\boldsymbol{\alpha}(r), \boldsymbol{\beta}$ positive real column vectors, and $\beta_i/\beta_j = V_{ij}^{(p)}$.

Remark A.4: The first requirement in the above theorem is an inhomogeneous analog of the notion of primitivity. The second requirement is automatically satisfied if the set of allowable matrices from which each H_k is drawn from is finite, and is essentially a uniformity condition.

Corollary A.1: Consider a sequence of positive integers $\{k_s\}, s \geq 0$ and $k_{s+1} - k_s = g$, where g is a constant. If we redefine the product $T_{p,r}$ by grouping g matrices $\{H_k\}$ into composite terms so that $T_{p,r} = H_{p+1} H_{p+2} \cdots H_{p+r} = T_{p,k_0-p} T_{k_0,k_1-k_0} T_{k_1,k_2-k_1} \cdots T_{k_{t-1},k_t-k_{t-1}} T^*$, for some allowable T^*, where k_t is the nearest member of $\{k_s\}$ not greater than r, and if, further

$$\phi(T_{k_s,k_{s+1}-k_s}) \geq \epsilon^2$$

then the rate at which $\tau_B(T_{p,r}) \to 0$ as $r \to \infty$ (see [12]) can be overbounded, in the case of $p = 0$, as

$$\tau_B(T_{0,r}) \leq \left(\frac{1-\epsilon}{1+\epsilon} \right)^{-(k_0/g)-1} \left(\frac{1-\epsilon}{1+\epsilon} \right)^{r/g}. \tag{25}$$

Remark A.5: By using the same arguments, it is straightforward to show a similar exponential rate of convergence for the reverse product $U_{p,r}$.

APPENDIX B
EQUIVALENCE OF THE TWO METHODS OF OBTAINING SMOOTHING EQUATIONS

In this section, we will show the equivalence between the two methods of obtaining the smoothing equations in Sections III-A and B.

Recall that by using the augmented state signal model, the smoothed probability vector is

$$\begin{aligned}\Pi_{j|j+\Delta} &= \begin{bmatrix} 1\cdots 1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & 1\cdots 1 \end{bmatrix} \hat{\Pi}_{j,j+\Delta|j+\Delta} \\ &= (I_N \otimes 1'_N)\hat{\Pi}_{j,j+\Delta|j+\Delta} \\ &= \frac{1}{\Theta_{j,j+\Delta}}[(I_N \otimes 1'_N)(\mathcal{C}_{y_{j+\Delta}}\mathcal{A}\mathcal{C}_{y_{j+\Delta-1}}\mathcal{A}\cdots \\ &\qquad \mathcal{C}_{y_{j+1}}\mathcal{A})]\hat{\Pi}_{j,j|j} \\ &= \frac{1}{\Theta_{j,j+\Delta}}[(I_N \otimes 1'_N)(I_N \otimes C_{y_{j+\Delta}})(I_N \otimes A)\cdots \\ &\qquad (I_N \otimes C_{y_{j+1}})(I_N \otimes A)]\hat{\Pi}_{j,j|j} \\ &= \frac{1}{\Theta_{j,j+\Delta}}[I_N \otimes (1'_N U_{j+1,j+\Delta})]\hat{\Pi}_{j,j|j}.\end{aligned}$$

Hence, each component of the smoothed probability vector can be expressed as

$$\begin{aligned}&\Pr(X_j = i|Y_0,\cdots,Y_{j+\Delta}) \\ &\quad = \frac{1}{\Theta_{j,j+\Delta}} 1'_N U_{j+1,j+\Delta} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \Pi_{j|j}(i) \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}0\\\vdots\\0\end{matrix}}\right\} i-1 \text{ zeros} \\ \left.\vphantom{\begin{matrix}0\\\vdots\\0\end{matrix}}\right\} N-i \text{ zeros} \end{matrix} \\ &\quad = \frac{1}{\Theta_{j,j+\Delta}} 1'_N C_{y_{j+\Delta}} A C_{y_{j+\Delta-1}} A \cdots C_{y_{j+1}} A e_i \Pi_{j|j}(i).\end{aligned} \tag{26}$$

By using the forward-backward signal model, it has previously been shown that for some normalizing constant Θ, the smoothed probability vector is

$$\begin{aligned}&\Theta \Pr(X = i) \Pr(X_j = i|Y_0,\cdots,Y_{j+\Delta}) \\ &\quad = [\Pi_{j|j}]_{i\text{th entry}} [\Pi_{j|j+1}]_{i\text{th entry}} \\ &\quad = \Pi_{j|j}(i)[e'_i A^b C_{y_{j+1}} A^b C_{y_{j+2}} \cdots A^b C_{y_{j+\Delta}} \\ &\qquad \cdot \Pi^+_{j+\Delta|j+\Delta+1}] \\ &\quad = \Pi_{j|j}(i)[e'_i(\Lambda A' \Lambda^{-1}) C_{y_{j+1}} (\Lambda A' \Lambda^{-1}) C_{y_{j+2}} \cdots \\ &\qquad (\Lambda A' \Lambda^{-1}) C_{y_{j+\Delta}} \Pi^+_{j+\Delta|j+\Delta+1}] \\ &\quad = \Pi_{j|j}(i)[e'_i \Lambda A' C_{y_{j+1}} A' C_{y_{j+2}} \cdots A' C_{y_{j+\Delta}} \Lambda^{-1} \\ &\qquad \cdot \Pi^+_{j+\Delta|j+\Delta+1}].\end{aligned} \tag{27}$$

The last line is due to the commutativity of diagonal matrices. Upon taking the transpose, (27) becomes

$$\begin{aligned}&\Theta \Pr(X = i) \Pr(X_j = i|Y_0,\cdots,Y_{j+\Delta}) \\ &\quad = [\Pi'^+_{j+\Delta|j+\Delta+1} \Lambda^{-1} (C_{y_{j+\Delta}} A C_{y_{j+\Delta-1}} A \cdots \\ &\qquad C_{y_{j+1}} A)\Lambda e_i]\Pi_{j|j}(i).\end{aligned}$$

Now, noting the fact that $\Pi^+_{j+\Delta|j+\Delta+1} = \{\Pr(X_j = i)\} = \{\Pr(X = i)\}$, it is evident that the product of the first two terms is just $1'_N$, and it is relatively straightforward to simplify the remaining terms. In particular

$$\begin{aligned}\Lambda e_i \Pi_{j|j}(i) &= \begin{bmatrix} \Pr(X=1) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \Pr(X=N) \end{bmatrix} e_i \Pi_{j|j}(i) \\ &= \Pr(X = i) e_i \Pi_{j|j}(i)\end{aligned}$$

and hence, (27) is just

$$\begin{aligned}&\Theta \Pr(X = i) \Pr(X_j = i|Y_0,\cdots,Y_{j+\Delta}) \\ &\quad = 1'_N C_{y_{j+\Delta}} A C_{y_{j+\Delta-1}} A \cdots C_{y_{j+1}} A \Pr(X = i) \\ &\qquad \cdot e_i \Pi_{j|j}(i).\end{aligned}$$

Cancelling $\Pr(X = i)$ on both sides and comparing with (26) establishes the assertion of equality between the two methods used to obtain smoothed estimates.

APPENDIX C
CALCULATION OF ERROR VARIANCE

In this section, we will outline the steps in obtaining the error variance given the the true state, filtered, and smoothed probability vectors.

First of all, without loss of generality, we will redefine the state variables X_k to be taken from the set of N unit vectors $e_i, 1 \le i \le N$. This formulation was introduced in [11] and has been used in [8].

Analogous to results for Kalman filtering and using the shorthand $\mathcal{Y}_k = \{Y_0, Y_1, \cdots Y_k\}$, we define our filtering error variance to be

$$\begin{aligned}&\tfrac{1}{2} E[(X_k - \Pi_{k|k})'(X_k - \Pi_{k|k})|\mathcal{Y}_k] \\ &\quad = \tfrac{1}{2}[1 - 2E[X'_k \Pi_{k|k}|\mathcal{Y}_k] + E[\Pi'_{k|k}\Pi_{k|k}|\mathcal{Y}_k]] \\ &\quad = \tfrac{1}{2}[1 - 2\Pi_{k|k} E[X'_k|\mathcal{Y}_k] + \Pi'_{k|k}\Pi_{k|k}] \\ &\quad = \tfrac{1}{2}[1 - \Pi'_{k|k}\Pi_{k|k}].\end{aligned}$$

The first term on the right follows because $X_k = [1 \quad 0]'$ or $[0 \quad 1]'$; the factor of 1/2 is to normalize the error variance in such a way that if every filtered estimate is completely the opposite of the true state, then the resulting error variance equals 1 after averaging. For example, given the true state vector at time k is $X_k = [1 \quad 0]$, if the filtered probability vector $\Pi_{k|k} = [0 \quad 1]'$, then $(X_k - \Pi_{k|k})'(X_k - \Pi_{k|k}) = 2$ without the factor of 1/2. This definition of error variance can be similarly extended to the smoothed probability vectors $\Pi_{j|k}$.

Remark C.1: Note that this definition of error variance is in analogy with the Kalman filtering case where the state process is continuous-valued. However, an alternative definition of an error variance for HMM's can also be introduced after making a kind of hard decision about the state of the HMM (e.g., a MAP estimate [8]) from the filtered probability distributions. In this case, the error variance would then correspond to the number of decision errors.

When no measurements are available, the best estimate of X_k is just the steady-state distribution Π, where $A\Pi = \Pi$. Hence, by letting $\Pi_{k|k} = E[X_k] = \Pi$

$$\tfrac{1}{2}\, E[(X_k - \Pi_{k|k})'(X_k - \Pi_{k|k})] = \tfrac{1}{2}\,[1 - \Pi'\Pi].$$

Since $\Pi = [0.5 \quad 0.5]'$ for a two-state HMM with a symmetric A matrix, the maximum error variance for such a system based on the steady-state distributions alone is 0.25.

REFERENCES

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

[2] B. D. O. Anderson, "Properties of optimal linear smoothing," *IEEE Trans. Automat. Contr.*, vol. AC-14, pp. 114–115, Feb 1969.

[3] B. D. O. Anderson and T. Kailath, "Forward and backward models for finite-state Markov processes," *Adv. Applied Prob.*, vol. 11, pp. 118–133, 1979.

[4] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*. Englewood Cliffs, NJ: Prentice-Hall, 1979.

[5] B. D. O. Anderson and I. Rhodes, "Smoothing algorithms for nonlinear finite-dimensional systems," *Stochastics*, vol. 9, pp. 139-165, 1983.

[6] R. K. Boel, J. B. Moore, and S. Dey, "Geometric convergence of filters for Hidden Markov Models," in *Proc. 34th Conf. Decision Contr.*, 1995, pp. 69–74.

[7] 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, 1975.

[8] R. J. Elliott, L. Aggoun, and J. B. Moore, *Hidden Markov Models: Estimation and Control*. New York: Springer-Verlag, 1994.

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

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

[11] A. Segall, "Recursive estimation from discrete-time point processes," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 422–431, 1976.

[12] E. Seneta, *Non-Negative Matrices and Markov Chains*, 2nd ed. New York: Springer-Verlag, 1981, ch. 3–4.

Louis Shue was born in Vietnam in 1971. He received the B.Sc. degree in 1994 and the B.E. degree in 1996 with first-class honors from Monash University, Melbourne, Australia, majoring in physics and electrical engineering, respectively. He is currently working towards the Ph.D. degree at the Australian National University, Canberra, Australia.

His research interests include signal processing, certain DES techniques, and applications in the field of communications.

Brian D. O. Anderson (S'62–M'66–SM'74–F'75) was born in Sydney, Australia, and received his undergraduate education at the University of Sydney, with majors in pure mathematics and electrical engineering. He subsequently received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA.

Following completion of his education, he worked in industry in Silicon Valley and served as a faculty member with the Department of Electrical Engineering, University of Newcastle, Callaghan, Australia, from 1967 to 1981 and is now Professor of Systems Engineering at the Australian National University, Canberra, and Director of the Research School of Information Sciences and Engineering. His interests are in control and signal processing.

Dr. Anderson is a Fellow of the Royal Society, the Australian Academy of Science, and the Australian Academy of Technological Sciences and Engineering, and an Honorary Fellow of the Institution of Engineers, Australia. He holds doctorates (honoris causa) from the Université Catholique de Louvain, Louvain, Belgium, the Swiss Federal Institute of Technology, Zürich, the University of Sydney, and the University of Melbourne, Parkville, Australia. He served a term as President of the International Federation of Automatic Control from 1990 to 1993.

Subhrakanti Dey was born in Calcutta, India, in 1968. He received the Ph.D. degree in systems engineering from the Australian National University, Canberra, in 1996. Previously, he received the Bachelor of Technology degree from the Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology (IIT), Kharagpur, India, in 1991 and the Master of Technology degree from IIT, specalizing in telecommunication systems engineering, in 1993.

He worked as a research assistant in an antenna array design project from January 1993 to July 1993 with the Radar Technology Centre, IIT. He was a post-doctoral Research Fellow with the Department of Systems Engineering, RSISE, Australian National University, from September 1995 to September 1997. Currently, he is a post-doctoral Research Associate at the Institute of Systems Research, University of Maryland, College Park. His current research interests are hybrid and discrete-event systems, statistical and adaptive signal processing, adaptive control, stochastic control, and robust nonlinear estimation.