Complexity Reduction in Fixed-Lag Smoothing for Hidden Markov Models

Louis Shue and Subhrakanti Dey, *Member, IEEE*

***Abstract*—In this paper, we investigate approximate smoothing schemes for a class of hidden Markov models (HMMs), namely, HMMs with underlying Markov chains that are nearly completely decomposable. The objective is to obtain substantial computational savings. Our algorithm can not only be used to obtain aggregate smoothed estimates but can be used also to obtain systematically approximate full-order smoothed estimates with computational savings and rigorous performance guarantees, unlike many of the aggregation methods proposed earlier.**

***Index Terms*—Hidden Markov model, nearly completely decomposable, reduced-complexity, slow–fast decomposition, state aggregation.**

I. Introduction

HIDDEN Markov models (HMMs) are extremely useful for modeling nonlinear physical phenomena. Although originally applied in speech recognition applications [1], signal processing methods based on HMMs have been successfully applied in equalization of communication channels [2], time-series applications such as in econometrics and seismic studies [3], biological signal processing, and many more areas. Most of these methods heavily depend on generic signal processing techniques such as state and parameter estimation algorithms. It is well known that "filtering" and "smoothing" are the two most important techniques for state estimation. In this paper, we address the problem of smoothing for a class of HMMs with underlying Markov chains that are "nearly completely decomposable." The results presented in this paper are an extension of the filtering results presented in [4].

Nearly completely decomposable Markov chains (NCDMCs) are Markov chains with a readily identifiable hierarchical structure of two or more levels. Although, in various applications, NCDMCs consist of a large number of states, the states of NCDMCs can be easily grouped together in what we shall term "super-states" [4], with strong interactions (i.e., high probability of transition) within these super-states and weak interactions between any two such super-states (i.e., small probability of transition). In [5], applications of NCDMCs were studied in queuing and computer systems. Further applications can be found in production planning of manufacturing systems [6], variable bit-rate video coding [7], multiple time-scale traffic modeling in communication networks [8], and many other biological and physical systems where dynamics of multiple rates are involved. Currently, investigations are underway to develop image-enhanced tracking algorithms for high-resolution radar applications with a large number of targets where NCDMCs are useful. While most previous work on NCDMCs concentrated on fully observed systems, very little work had been done (prior to [4]) on reduced-complexity state estimation for partially observed NCDMCs (or in other words, HMMs with underlying NCDMCs). Since, in many of the above applications, the underlying NCDMCs are only observed through noisy measurements, reduced-complexity filtering and smoothing results for such partially observed NCDMCs are indispensable. In this paper, we propose an approximate smoothing algorithm for partially observed NCDMCs that utilizes this hierarchical structure and provides estimates to the conditional smoothed state probabilities but with a reduced order of computations compared with if exact smoothing was carried out.

NCDMCs have been extensively studied in [5], which considered NCDMCs (for a two-level hierarchy) with transition probability matrices of the form $P = I_n + A + \epsilon B$, where n denotes the total number of states, and $\epsilon > 0$ is a small perturbation parameter. Here, I_n is the $(n \times n)$ identity matrix, and

$$A = \begin{bmatrix} A_{11} & 0 & \cdots & 0 \\ 0 & A_{22} & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & A_{NN} \end{bmatrix}$$

where $A_{ii} \in \mathbb{R}^{n_i \times n_i}\ \forall i$, $\sum_i n_i = n$, and 0 denotes zero matrices of appropriate dimensions. Note that $I_n + A$ is also row-stochastic, and rows of A and B sum to 0. We also make the following assumption.

Assumption 1.1: P and $I_{n_i} + A_{ii}, \forall i$ are irreducible.

Typically, n is much larger than N, and thus, if $\epsilon = 0$, the chain decomposes into N separate noninteracting Markov chains. For small ϵ, the states can be clustered into N groups such that there is strong interaction between the states in a given group but weak interaction between the groups. Following [4], we will term the N groups the "super-states." We denote the state of the full Markov chain as $X_k \in \{1, 2, \ldots, n\}$, and the lth superstate is denoted by S_l, $l \in \{1, 2, \ldots, N\}$. Without loss of generality, let $S_1 = \{1, 2, \ldots, n_1\}$, and $S_2 = \{n_1 + 1, n_1 + 2, \ldots, n_1 + n_2\}$, etc. Note that consequently, we have

$$\Pr(X_k \in S_l) = \sum_{i=1+n_1+\cdots+n_{l-1}}^{n_1+n_2+\cdots+n_l} \Pr(X_k = i).$$

Manuscript received February 28, 2001; revised February 4, 2002. The associate editor coordinating the review of this paper and approving it for publication was Prof. Randolph L. Moses.

L. Shue is with the Centre for Signal Processing, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore.

S. Dey is with the Department of Electrical and Electronic Engineering, University of Melbourne, Parkville, Australia.

Publisher Item Identifier S 1053-587X(02)03221-X.

This weak coupling is crucial in obtaining computational savings in many estimation and control algorithms for Markovian systems with underlying NCDMCs.

Previous research on NCDMCs has mostly concentrated on obtaining approximations to the steady-state distributions of these chains. Various approaches have been put forth to derive from P an aggregate version of smaller dimension, resulting in schemes of $O(\epsilon)$ [5] approximation or even the exact distribution using stochastic complementation methods [9]. An iterative scheme to obtain approximations of (potentially) arbitrary accuracy that avoids possible numerical ill-conditioning was given in [10]. There have been several other studies that contributed to the development of decomposition-aggregation methods for obtaining reduced-order approximations for uncontrolled [11], as well as controlled Markov chains [12], [13]. A singular perturbation interpretation to Courtois' aggregation is also given in [13]. The singular perturbation approach to study aggregation of finite-state Markov chains has also been studied in [14]–[16]. In [10], the aggregation method developed is also used to obtain aggregation of the policy iteration method in infinite-horizon optimal control of such Markov chains. In summary, these works developed various aggregation/decomposition schemes to obtain approximations to steady-state distributions and hierarchical aggregation of optimal control policies for such Markov chains. The problem of the infinite horizon average cost control problem for such Markov chains was also addressed in [17] and [18]. It was shown that the optimal solution can be approximated by an optimal solution to the so-called *limit Markov control problem* for a sufficiently small ϵ. Algorithms were also provided to achieve these control strategies.

Very few studies exist, however, on partially observed NCDMCs. The state estimation of HMMs where the underlying Markov chain is an NCDMC was first *systematically* investigated in [4], although some related studies can be found in [19] and, more recently, in [20]. In [4], apart from the structure inherent in NCDMCs, an additional assumption was made on the observation probabilities. This assumption was that the observation probabilities reflect the block structure of the Markov states. That is, for a given observation symbol, the state-to-observation transition probability is constant for all states within the same super state. To state formally, let the state-to-measurement mapping be given by the measurement matrix C, where $c_{ij} = \Pr(Y_k = i | X_k = j)$, $i \in \{1, 2, \ldots, M\}$ and $j \in \{1, 2, \ldots, n\}$, and Y_k denotes the discrete observation at time k. Following [4], we also make the following assumption on C to be used in this paper.

Assumption 1.2: $c_{ij} = \bar{c}_{il},\ \forall\, j \in S_l,\ \forall\, i.$

We make the following additional but standard assumption on C since otherwise, the measurements contain very little information of the Markov states.

Assumption 1.3: $\min_{ij} c_{ij} \geq \underline{c} > 0.$

The applicability of such block-structured observation probability (as required by Assumption 1.2) matrices not only lies in modeling of management systems (where top levels of management are only interested in macro-behavior rather than micro-behavior) but in real engineering applications like distributed control environments, particularly with communication constraints, as well. For example, in an environment where multiple sensors are sending information, it might not be possible to send fine information due to bit-rate constraints, and hence, it might just be practical to send coarser information (e.g., information about the macro states). This may also be of use in hierarchical control systems, where a controller at one of the top levels of the hierarchy may not want fine information since it may only want to control transitions from one macro-state to another (e.g., the controller may want to know that a failure has occurred and not what particular kind of failure it is). It was demonstrated in [4] that substantial computational savings can be obtained in calculating the aggregate filtered estimates (approximate) via a decoupling scheme for this class of HMMs. It was shown that one can obtain $O(\epsilon^2)$ approximation to the aggregate filtered estimates with substantial savings. It was also seen that some aggregation methods (including Courtois' method) may be adapted to obtain comparable results as far as aggregate filtered estimates are concerned. However, the algorithm proposed in [4] can be used to obtain $O(\epsilon^2)$ approximation to the full-order filtered state estimates, whereas none of the aggregation methods can be adapted to achieve that. The computational savings in calculating approximate full-order estimates are also substantial if the large-scale NCDMC has superstates with small individual dimensions. All the results of [4] are also valid for a state-to-observation transition probability matrix that is a polynomial in ϵ perturbation of the "slow" block-structured transition matrix. However, obtaining reduced-order computations for the state estimates for a general C matrix is still an unsolved problem.

In this paper, we present reduced-complexity smoothing algorithms for such partially observed NCDMCs. Following similar techniques in [4], we do the following.

1) We provide a systematic method to obtain an $O(\epsilon^2)$ approximation to aggregate and full-order conditional smoothed probabilities.
2) We show that using aggregation methods of [5] and [10], one can obtain comparable approximations to the aggregate estimates (in fact, Courtois' method results in the same $O(\epsilon^2)$ approximation, as indicated originally for the filtering results in [4]).
3) We perform comparative studies regarding computational savings obtained in calculating full-order and aggregate smoothed estimates.

The novelty of our contribution lies in the following.

1) Our method provides a *systematic* way to obtain $O(\epsilon^2)$ approximations to the full-order (not just aggregate) smoothed estimates, whereas no aggregation method can be adapted to achieve this.
2) Even though the aggregation methods of [5] and [10] can be adapted to achieve comparable (and in some cases better) approximations to the aggregate smoothed estimates, they can become *ad hoc* in certain cases, e.g., when the state-to-observation transition probability matrix is a small perturbation of the block-structured matrix (as discussed previously).
3) Unlike the filtering results of [4], the computational savings obtained in calculating the full-order approximate

smoothed estimates are quite substantial compared with exact calculations.

4) In a special case when the individual sub-Markov chains identified by the super states (when $\epsilon = 0$) are independent and identically distributed (in this case, the state-to-observation transition probability matrix has no restriction on itself), one can obtain $O(\epsilon^3)$ approximations to the full-order and aggregate smoothed estimates. Notice that no aggregation method can be adapted to achieve this. This result is not explicitly included here but follows immediately from similar filtering results in [4]. This observation is also a clear indication that our algorithm provides a *systematic* way of exploiting the system structure in obtaining computational savings while providing rigorous performance bounds on the order of these approximations.

In Section II, we describe the aggregate smoothing algorithm followed by the approximate scheme in Section III. We illustrate the performance of the smoothing algorithm by some simulations and comparative studies in Section IV. Finally, some concluding remarks are presented in Section V.

II. Aggregate Smoothed Estimates

A. Exact Smoothing Equations

To construct an HMM fixed-lag smoother, we proceed similarly to [21] (which extends Kalman filtering results to smoothing results) by constructing an augmented signal model, consisting of the original Markov chain and a state X_j of which the smoothed estimate is sought after. A filtered estimate of the augmented model at time $k(>j)$ will then contain within it a filtered estimate of the state X_k as well as something equivalent to a smoothed estimate of the original model at time j.

Definition 2.1: For each $k \geq j$, let $\mathcal{Z}_k = Z_{j,k} = [X_j \quad X_k]'$ be an augmented state vector consisting of the states of the original Markov chain at a fixed time j and a variable time k.

From Definition 2.1, it can be seen that $\mathcal{Z}_k$ can only assume the values $(1, 1), (1, 2), \ldots, (1, n), (2, 1), \ldots, (n, n)$. It follows (see [22] for details) then that the transition probability matrix for such a Markov chain is $\mathcal{P} = I_n \otimes P$, where $\otimes$ denotes Kronecker product. We will now argue that the output process Y_k of the original HMM can also be regarded as the output process of an HMM with state $\mathcal{Z}_k$ for $k \geq j$. That is, suppose that the output process associated with $\mathcal{Z}_k$ is the same as before and, consequently, that $\Pr(Y_k = \ell | \mathcal{Z}_k) = \Pr(Y_k = \ell | X_k)$; this means that the corresponding observation matrix $\mathcal{C}$ for the augmented Markov chain is $\mathcal{C} = [C \quad C \quad \cdots \quad C] = 1_n' \otimes C$, where 1_p denotes a column vector of length p with all entries equal to 1.

In the subsequent discussions, we will use the shorthand $\mathcal{Y}_k$ to denote $\{Y_0, Y_1, \ldots, Y_k\}$. Denote the $(1 \times n^2)$ filtered probability vector for $\mathcal{Z}_k$ as $\Pi_{j,\,k|k}$, with each component being

$$\Pi_{j,\,k|k}(I) = \Pr(\mathcal{Z}_k = I | \mathcal{Y}_k)$$

where $I \in \{(1, 1), (1, 2), \ldots, (n, n)\}$. The recursion for this probability vector is (see also [22])

$$\begin{aligned}\Pi_{j,k+1|k+1} &= \frac{1}{Z_{k+1}} \Pi_{j,k|k} \mathcal{P} \mathcal{C}_{Y_{k+1}} \\ &= \frac{1}{Z_{k+1}} \Pi_{j,k|k} (I_n \otimes P) \left(I_n \otimes C_{Y_{k+1}}\right) \quad (1)\end{aligned}$$

where $C_{Y_{k+1}} = \text{diag}(c_{mi})$, where $Y_{k+1} = m$, and $i = 1, 2, \ldots, n$. $Z_{k+1} = \Pi_{j,\,k|k} \mathcal{P} \mathcal{C}_{Y_{k+1}} 1_{n^2}$ is a scalar normalizing constant that ensures $\sum_{i=1}^{n^2} \Pi_{j,\,k+1|k+1}(i) = 1$.

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

$$\begin{aligned}\Pi_{j|j+\Delta}(i) &= \Pr(X_j = i | \mathcal{Y}_{j+\Delta}) \\ &= \sum_{\ell=1}^{n} \Pr(X_j = i,\, X_{j+\Delta} = \ell | \mathcal{Y}_{j+\Delta}).\end{aligned}$$

Hence, 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

$$\begin{aligned}\Pi_{j|j+\Delta} &= \Pi_{j,\,j+\Delta|j+\Delta} \begin{bmatrix} 1_n & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & 1_n \end{bmatrix} \\ &= \Pi_{j,\,j+\Delta|j+\Delta} (I_n \otimes 1_n) \\ &= \frac{1}{Z_{j,j+\Delta}} \Pi_{j,j|j} (I_n \otimes P) \left(I_n \otimes C_{Y_{j+1}}\right) \\ &\quad \cdots (I_n \otimes P) \left(I_n \otimes C_{Y_{j+\Delta}}\right) (I_n \otimes 1_n) \\ &= \frac{1}{Z_{j,j+\Delta}} \Pi_{j,\,j|j} (I_n \otimes U_{j+1,j+\Delta} 1_n) \quad (2)\end{aligned}$$

where $Z_{j,\,j+\Delta}$ is a normalizing constant, and

$$U_{j+1,\,j+\Delta} = P C_{Y_{j+1}} P C_{Y_{j+2}} \cdots P C_{Y_{j+\Delta}}.$$

We have also used the property $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$.

Remark 2.1: Although (2) has been obtained from an n^2-state augmented system, by straightforward algebraic manipulations, we can rewrite it in a more compact form involving products of $(n \times n)$ matrices. That is, each component of the $\Pi_{j|j+\Delta}$ can be written in terms of the corresponding filtered estimate at time j as

$$\Pi_{j|j+\Delta} = \frac{1}{Z_{j,j+\Delta}} \text{diag}(\Pi_{j|j}) U_{j+1,\,j+\Delta} 1_n.$$

B. Aggregate HMM Smoother

Let us consider, as in [10], the nonsingular transformation $\Gamma = [W_1 \quad W_2]$ such that $I_n = [W_1 \quad W_2] \left[\begin{smallmatrix} V_1 \\ V_2 \end{smallmatrix}\right]$, where $W_1 \in$

$\mathbb{R}^{n\times N}$ and the ith diagonal blocks in $W_2 \in \mathbb{R}^{n\times(n-N)}$, $V_1 \in \mathbb{R}^{N\times n}$, and $V_2 \in \mathbb{R}^{(n-N)\times n}$ are given as follows:

$$W_1 = \begin{bmatrix} 1_{n_1} & 0 & \cdots & 0 \\ 0 & 1_{n_2} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & \cdots & 0 & 1_{n_N} \end{bmatrix} \tag{3a}$$

$$W_2^{(i)} = \begin{bmatrix} 0\cdots 0 \\ I_{n_i-1} \end{bmatrix} \tag{3b}$$

$$V_1^{(i)} = [1 \quad 0 \quad \cdots \quad 0] \tag{3c}$$

$$V_2^{(i)} = [-1_{n_i-1} \quad I_{n_i-1}]. \tag{3d}$$

We note that the aggregate fixed-lag smoothed state estimates can be represented as

$$\zeta_{j|j+\Delta} = \Pi_{j|j+\Delta} W_1 \tag{4}$$

where $\zeta_{j|j+\Delta} \in \mathbb{R}^{1\times N}$, and $\zeta_{j|j+\Delta}(i) = \Pr(X_j \in S_i|\mathcal{Y}_{j+\Delta}) = \sum_{\ell\in S_i} \Pr(X_j = \ell|\mathcal{Y}_{j+\Delta})$.

We will now indicate the steps to obtain the aggregate smoothed estimates using the transformation matrices (3a)–(3d).

Step 1: Denote the product of $\Pi_{j,k|k}$ with the ith diagonal block of $I_n \otimes [W_1 \quad W_2]$ as $[\zeta_{j,k}^{(i)} \quad \eta_{j,k}^{(i)}]$, with $\zeta_{j,k}^{(i)} \in \mathbb{R}^{1\times N}$ and $\eta_{j,k}^{(i)} \in \mathbb{R}^{1\times(n-N)}$, where each $\zeta_{j,k}^{(i)}$ is

$$\zeta_{j,k}^{(i)} = [\Pr(X_j = i,\, X_k \in S_1|\mathcal{Y}_k) \quad \Pr(X_j = i,\, X_k \in S_2|\mathcal{Y}_k) \quad \cdots \quad \Pr(X_j = i,\, X_k \in S_N|\mathcal{Y}_k)].$$

We will now rewrite (1) in terms of $\zeta_{j,k}^{(i)}$ and $\eta_{j,k}^{(i)}$, $i = 1, 2, \ldots, n$. That is

$$\begin{aligned} &\Pi_{j,k+1|k+1}(I_n \otimes [W_1 \quad W_2]) \\ &\quad = \frac{1}{Z_{k+1}} \Pi_{j,k|k} \left(I_n \otimes [W_1 \quad W_2] \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \right) \\ &\qquad \cdot (I_n \otimes P)\left(I_n \otimes C_{Y_{k+1}}\right)(I_n \otimes [W_1 \quad W_2]) \end{aligned}$$

or

$$\begin{aligned} &\left[\zeta_{j,k+1}^{(1)} \;\; \eta_{j,k+1}^{(1)} \;\; \zeta_{j,k+1}^{(2)} \;\; \eta_{j,k+1}^{(2)} \;\; \cdots \;\; \zeta_{j,k+1}^{(n)} \;\; \eta_{j,k+1}^{(n)}\right] \\ &\quad = \frac{1}{Z_{k+1}} \left[\zeta_{j,k}^{(1)} \;\; \eta_{j,k}^{(1)} \;\; \zeta_{j,k}^{(2)} \;\; \eta_{j,k}^{(2)} \;\; \cdots \;\; \zeta_{j,k}^{(n)} \;\; \eta_{j,k}^{(n)}\right] \\ &\qquad \cdot \left(I_n \otimes \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} PC_{Y_{k+1}} [W_1 \quad W_2] \right). \end{aligned} \tag{5}$$

Note that this recursion can also be written in two steps:

1)

$$\begin{aligned} &\left[\zeta_{j,k+1}^{(1)u} \;\; \eta_{j,k+1}^{(1)u} \;\; \zeta_{j,k+1}^{(2)u} \;\; \eta_{j,k+1}^{(2)u} \;\; \cdots \;\; \zeta_{j,k+1}^{(n)u} \;\; \eta_{j,k+1}^{(n)u}\right] \\ &\quad = \left[\zeta_{j,k}^{(1)} \;\; \eta_{j,k}^{(1)} \;\; \zeta_{j,k}^{(2)} \;\; \eta_{j,k}^{(2)} \;\; \cdots \;\; \zeta_{j,k}^{(n)} \;\; \eta_{j,k}^{(n)}\right] \\ &\qquad \cdot \left(I_n \otimes \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} PC_{Y_{k+1}} [W_1 \quad W_2] \right) \end{aligned} \tag{6}$$

where $\zeta_{j,k+1}^{(i)u}$, $\eta_{j,k+1}^{(i)u}$ denote the unnormalized variables for each i, j.

2)

$$\begin{aligned} &\left[\zeta_{j,k+1}^{(1)} \;\; \eta_{j,k+1}^{(1)} \;\; \zeta_{j,k+1}^{(2)} \;\; \eta_{j,k+1}^{(2)} \;\; \cdots \;\; \zeta_{j,k+1}^{(n)} \;\; \eta_{j,k+1}^{(n)}\right] \\ &\quad = \frac{1}{Z_{k+1}} \left[\zeta_{j,k+1}^{(1)u} \;\; \eta_{j,k+1}^{(1)u} \;\; \zeta_{j,k+1}^{(2)u} \;\; \eta_{j,k+1}^{(2)u} \;\; \cdots \;\; \zeta_{j,k+1}^{(n)u} \;\; \eta_{j,k+1}^{(n)u}\right] \end{aligned} \tag{7}$$

where it is easy to show that Z_{k+1} of (1) can be also expressed as $Z_{k+1} = \sum_{i=1}^{n} \zeta_{j,k+1}^{(i)u} 1_N$.

Step 2: At each $k = j + \Delta$, where Δ denotes the smoothing lag, the full-order smoothed probability vector is computed by summing the contribution from each X_j

$$\begin{aligned} &\Pi_{j|j+\Delta} \\ &\quad = \left[\zeta_{j,j+\Delta}^{(1)} \;\; \eta_{j,j+\Delta}^{(1)} \;\; \zeta_{j,j+\Delta}^{(2)} \;\; \eta_{j,j+\Delta}^{(2)} \;\; \cdots \;\; \zeta_{j,j+\Delta}^{(n)} \;\; \eta_{j,j+\Delta}^{(n)}\right] \\ &\qquad \cdot (I_n \otimes 1_{sn}) \end{aligned} \tag{8}$$

where $1_{sn} = [1_N' \quad 0 \quad \cdots \quad 0]'(\in \mathbb{R}^{n\times 1})$, and $\Pi_{j|j+\Delta}(i) = \Pr(X_j = i|\mathcal{Y}_{j+\Delta})$.

Step 3: The aggregate smoothed probability vector is then computed using (4). We then have

$$[\zeta_{j|j+\Delta} \quad \eta_{j|j+\Delta}] = \Pi_{j|j+\Delta} [W_1 \quad W_2]. \tag{9}$$

Remark 2.2: Similar to Remark 2.1, (6) and (7) can also be written as

$$\begin{aligned} &\begin{bmatrix} \zeta_{j,k+1}^{(1)u} & \eta_{j,k+1}^{(1)u} \\ \zeta_{j,k+1}^{(2)u} & \eta_{j,k+1}^{(2)u} \\ \vdots & \\ \zeta_{j,k+1}^{(n)u} & \eta_{j,k+1}^{(n)u} \end{bmatrix} \\ &\quad = \begin{bmatrix} \zeta_{j,k}^{(1)} & \eta_{j,k}^{(1)} \\ \zeta_{j,k}^{(2)} & \eta_{j,k}^{(2)} \\ \vdots & \\ \zeta_{j,k}^{(n)} & \eta_{j,k}^{(n)} \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} PC_{Y_{k+1}} [W_1 \quad W_2] \end{aligned} \tag{10}$$

where

$$\left[\zeta_{j,k+1}^{(i)} \;\; \eta_{j,k+1}^{(i)}\right] = \frac{1}{Z_{k+1}} \left[\zeta_{j,k+1}^{(i)u} \;\; \eta_{j,k+1}^{(i)u}\right] \quad \forall i = 1, 2, \ldots, n.$$

In [4], it has been shown that such calculations (for a fixed i) can be performed approximately by decoupling $\zeta_{j,k}^{(i)}$ from $\eta_{j,k}^{(i)}$. The subsequent calculations (whether for aggregate or full-order smoothed estimates) require a reduced number of computations [as they only require recursive computations of $\zeta_{j,k+1}^{(i)}$, $i = 1, 2, \ldots, n$] and the estimates are of $O(\epsilon^2)$ in approximation. We will show a similar development in the next section.

III. Approximate $O(\epsilon^2)$ Aggregate Smoother

In this section, we will adopt the decoupling transformation technique as used in [4] to obtain approximate aggregate and full-order smoothed state estimates.

As indicated in (10), for fixed j and at each k, the computation of $[\zeta_{j,k}^{(i)} \quad \eta_{j,k}^{(i)}]$ involves only products of $(n \times n)$ matrices. We will now recall some results from [4] that are directly applicable in the present situation. Denote the transformed variables as $[\overline{\zeta}_{j,k}^{(i)} \quad \overline{\eta}_{j,k}^{(i)}]$ given by

$$[\overline{\zeta}_{j,k}^{(i)} \quad \overline{\eta}_{j,k}^{(i)}] = [\zeta_{j,k}^{(i)} \quad \eta_{j,k}^{(i)}] \begin{bmatrix} I_N & L_k \\ 0 & I_{n-N} \end{bmatrix} \tag{11}$$

where it is trivial to demonstrate that

$$\begin{bmatrix} I_N & L_k \\ 0 & I_{n-N} \end{bmatrix}^{-1} = \begin{bmatrix} I_N & -L_k \\ 0 & I_{n-N} \end{bmatrix}. \tag{12}$$

Here, $\{L_k \in \mathbb{R}^{N \times (n-N)}\}$ is assumed to be (for the time being) a sequence of uniformly bounded time-varying matrices to be solved for. More rigorous statements will be made regarding the uniform boundedness of L_k later in this section.

Now, to simplify the notation, let us introduce the shorthand

$$\begin{bmatrix} \tilde{A}_{11}^k & \tilde{A}_{12}^k \\ \tilde{A}_{21}^k & \tilde{A}_{22}^k \end{bmatrix} = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} PC_{Y_{k+1}} [W_1 \quad W_2] \tag{13}$$

where $\tilde{A}_{11}^k = \tilde{A}_1^k + \epsilon \tilde{B}_1^k$, $\tilde{A}_{12}^k = \tilde{A}_2^k + \epsilon \tilde{B}_2^k$, $\tilde{A}_{21}^k = \tilde{C}_1^k + \epsilon \tilde{D}_1^k$ and $\tilde{A}_{22}^k = \tilde{C}_2^k + \epsilon \tilde{D}_2^k$, and the individual terms are

$$\tilde{A}_1^k = V_1(I_n + A)C_{Y_{k+1}}W_1, \quad \tilde{B}_1^k = V_1 B C_{Y_{k+1}} W_1 \tag{14a}$$

$$\tilde{A}_2^k = V_1(I_n + A)C_{Y_{k+1}}W_2, \quad \tilde{B}_2^k = V_1 B C_{Y_{k+1}} W_2 \tag{14b}$$

$$\tilde{C}_1^k = V_2(I_n + A)C_{Y_{k+1}}W_1, \quad \tilde{D}_1^k = V_2 B C_{Y_{k+1}} W_1 \tag{14c}$$

$$\tilde{C}_2^k = V_2(I_n + A)C_{Y_{k+1}}W_2, \quad \tilde{D}_2^k = V_2 B C_{Y_{k+1}} W_2. \tag{14d}$$

Using (11)–(13) [and writing only the i-row of (10) for simplicity], we have the following recursion:

$$\begin{aligned} &[\overline{\zeta}_{j,k+1}^{(i)} \quad \overline{\eta}_{j,k+1}^{(i)}] \\ &= \frac{1}{Z_{k+1}} [\overline{\zeta}_{j,k}^{(i)} \quad \overline{\eta}_{j,k}^{(i)}] \begin{bmatrix} I_N & -L_k \\ 0 & I_{n-N} \end{bmatrix} \\ &\quad \cdot \begin{bmatrix} \tilde{A}_{11}^k & \tilde{A}_{12}^k \\ \tilde{A}_{21}^k & \tilde{A}_{22}^k \end{bmatrix} \begin{bmatrix} I_N & L_{k+1} \\ 0 & I_{n-N} \end{bmatrix} \\ &= \frac{1}{Z_{k+1}} [\overline{\zeta}_{j,k}^{(i)} \quad \overline{\eta}_{j,k}^{(i)}] \\ &\quad \cdot \begin{bmatrix} \tilde{A}_{11}^k - L_k \tilde{A}_{21}^k & 0 \\ \tilde{A}_{21}^k & \tilde{A}_{21}^k L_{k+1} + \tilde{A}_{22}^k \end{bmatrix} \end{aligned} \tag{15}$$

where L_k satisfies

$$\left(\tilde{A}_{11}^k - L_k \tilde{A}_{21}^k\right) L_{k+1} = L_k \tilde{A}_{22}^k - \tilde{A}_{12}^k, \qquad L_0 = 0. \tag{16}$$

Note that one can solve for L_k recursively using (16), provided $(\tilde{A}_{11}^k - L_k \tilde{A}_{21}^k)^{-1}$ exists for all k. However, this method results in substantial computational requirements and defeats the purpose of the original objective of this work. Under the uniform boundedness assumption on L_k, it is a common practice in singular perturbation literature [23] to expand L_k as a power series of ϵ: $L_k = L_k(0) + \epsilon L_k(1) + \cdots$, where $L_k(0)$ is the solution to (16) when $\epsilon = 0$. One can iteratively compute $L_k(0)$, $L_k(1)$, etc., and truncate this power series at some finite power of ϵ to obtain a desired order of accuracy for L_k.

The first step in order to obtain recursive computations for $L_k(0)$, $L_k(1)$, etc., is to notice the following fact. Under Assumption 1.2, the ith diagonal element in $\tilde{A}_1^k$ is $\overline{c}_{ji}$ if $Y_{k+1} = j$ and $\tilde{C}_1^k = 0$. Similarly, all the elements of the ith diagonal block in $\tilde{A}_2^k$, $\tilde{C}_2^k$ are scaled by $\overline{c}_{ji}$ if $Y_{k+1} = j$. Due to this simple scaling property, $(\tilde{A}_1^k)^{-1}\tilde{A}_2^k$ can be written as a time-invariant matrix $G_0 \in \mathbb{R}^{N \times (n-N)}$. One can now rewrite (16) as

$$L_{k+1} = \left(\tilde{A}_1^k\right)^{-1} L_k \tilde{C}_2^k - G_0 + \epsilon G_{k+1}, \quad L_0 = 0 \tag{17}$$

where $G_{k+1} = (\tilde{A}_1^k)^{-1}(L_k \tilde{D}_2^k - \tilde{B}_1^k L_{k+1} + L_k \tilde{D}_1^k L_{k+1} - \tilde{B}_2^k)$. As mentioned above, it is a usual practice in singular perturbation literature to express the solution to (17) as $L_k = L_k(0) + \epsilon \tilde{L}_k(\epsilon)$, where $\tilde{L}_k(\epsilon)$ can again be written (under the uniform boundedness assumption) as a power series in ϵ. It was shown in [4] that $L_k(0)$ satisfies the following recursion:

$$L_{k+1}(0) = \left(\tilde{A}_1^k\right)^{-1} L_k(0) \tilde{C}_2^k - G_0, \quad L_0(0) = 0. \tag{18}$$

It was also shown in [4] that as $k \to \infty$, $L_k(0) \to L(0)$, where $L(0) = V_1 A W_2 (V_2 A W_2)^{-1}$ (noting that $V_2 A W_2$ is invertible [10]). Since we are mainly interested in the solution to (17) as $k \to \infty$, we can obtain an $O(\epsilon^2)$ approximation to L_k by replacing L_k with $L(0) + \epsilon L_k(1)$. It was shown in [4] how one can obtain reduced-complexity $O(\epsilon^2)$ filtered state estimates using this approximation to L_k.

It would be appropriate here to make some comments about the uniform boundedness of L_k as a solution to (17). Denote by $\|.\|_2$ the Frobenius norm for a matrix (note that this is a matrix norm for a square matrix). Note that the uniform boundedness on L_k demands that $L_k \in \mathcal{D} \triangleq \{L\colon \|L\|_2 < (1 + \epsilon\overline{L})\|L(0)\|_2\}$, where $\|\tilde{L}_k(\epsilon)\|_2 < \epsilon\overline{L}\|L(0)\|_2$, and $\|L_k\|_2 < (1 + \epsilon\overline{L})\|L(0)\|_2, \ \forall k$. It was shown that under some sufficient conditions (which essentially indicate that ϵ should be sufficiently small), one can guarantee that $L_k \in \mathcal{D}$. For more details on these sufficient conditions and a rigorous proof of this result, see [4].

One can similarly show from (17) (under the uniform boundedness assumption on L_k) that the recursion for $L_k(1)$ is as follows:

$$L_{k+1}(1) = \left(\tilde{A}_1^k\right)^{-1} L_k(1) \tilde{C}_2^k + Q_k \tag{19}$$

where $Q_k = (\tilde{A}_1^k)^{-1}(L(0)\tilde{D}_2^k + L(0)\tilde{D}_1^k L(0) - \tilde{B}_1^k L(0) - \tilde{B}_2^k)$. For a remark on the uniform boundedness of $L_k(1)$, see [4].

It will be clear presently that in order to reduce the number of computations with an $O(\epsilon^2)$ approximation to the smoothed state estimates, we only need to consider solving for $L_k(1)$. Higher order approximations to L_k do not result in computational reductions. Later in this paper, we discuss how using $L_k \approx L(0) + \epsilon L_k(1)$ lets us obtain reduction in computations.

TABLE I
COMPARISON OF NUMBER OF COMPUTATIONS FOR EACH SMOOTHING SCHEME

		Additions	Multiplications + Divisions[2]
Full-state:	Exact	$\Delta n^2(n-1)+nN-1$	Δn^3+n
	Decoupling scheme	$(\Delta-1)N^2(N-1)+nN^2-1$	$(\Delta-1)N^3+nN^2+N$
Aggregate:	Exact	$\Delta n^2(n-1)+nN-1$	Δn^3+N
	Decoupling scheme	$(\Delta-1)N^2(N-1)+nN^2-1$	$(\Delta-1)N^3+nN^2+N$
	Courtois' method	$\Delta N^2(N-1)+(N^2-1)$	ΔN^3+N

First, however, notice that, from (15), one can rewrite the decoupled recursion for $\overline{\eta}^{(i)}_{j,k}$, $i=1,2,\ldots,n$ as

$$\overline{\eta}^{(i)}_{j,k+1}=\frac{1}{Z_{k+1}}\overline{\eta}^{(i)}_{j,k}\left(\tilde{A}^k_{21}L_{k+1}+\tilde{A}^k_{22}\right). \tag{20}$$

It is easy to show that under the following assumption (Assumption 3.1), $\overline{\eta}^{(i)}_{j,k}\to 0,\ \forall i$ as $k\to\infty$. For sufficient conditions under which this assumption holds, see [4].

Assumption 3.1: The evolution $z_{k+1}=z_k(\tilde{A}^k_{21}L_{k+1}+\tilde{A}^k_{22})$ [namely, the recursion for $\overline{\eta}^{(i)}_{j,k}$ in (15)], where $z_k\in\mathbb{R}^{1\times(n-N)}$ is exponentially stable.

The rate at which $\overline{\eta}^{(i)}_{j,k}\to 0$ is determined by the fast eigenvalues of $\tilde{A}^k_{21}L_{k+1}+\tilde{A}^k_{22}$ and how close they are to the origin. It follows from Assumption 3.1 that there exists a large enough but finite k_0 such that for $k\geq k_0$, $|\overline{\eta}^{(i)}_{j,k}|$ is of $O(\epsilon^2)$. Setting $\tilde{\zeta}^{(i)}_{j,k}=\zeta^{(i)}_{j,k}$, $\tilde{\eta}^{(i)}_{j,k}=\eta^{(i)}_{j,k}$ for $k<k_0$, consider the following approximate recursions for $k\geq k_0$:

$$\begin{aligned}
\tilde{\zeta}^{(i)u}_{j,k+1}&=\tilde{\zeta}^{(i)}_{j,k}\left(\tilde{A}^k_{11}-L(0)\tilde{A}^k_{21}\right)\\
\tilde{\zeta}^{(i)}_{j,k+1}&=\frac{1}{\tilde{Z}_{k+1}}\tilde{\zeta}^{(i)u}_{j,k+1},\qquad \tilde{Z}_{k+1}=\sum_{i=1}^{n}\tilde{\zeta}^{(i)u}_{j,k+1}1_N\\
\tilde{\eta}^{(i)}_{j,k+1}&=-\tilde{\zeta}^{(i)}_{j,k+1}\left(L(0)+\epsilon L_{k+1}(1)\right).
\end{aligned} \tag{21}$$

It was shown in [4] how these recursions result in $O(\epsilon^2)$ approximations to the exact recursions given by (7). Before we summarize our results in the following theorem (Theorem 3.1), we need to make one further assumption.

Assumption 3.2: ϵ is sufficiently small such that $1/Z_k=(1/\tilde{Z}_k)+O(\epsilon^2)$ uniformly in k.

Remark 3.1: Note that Assumption 3.2 guarantees that the normalization procedure does not alter the order of approximation of the unnormalized variables.

Without further ado, we present the main result of this paper in the following theorem. The proof is not included here simply because it is identical to that of a similar theorem (Theorem 1) in [4]. It also follows easily from the previous discussions. Note that in the statement of the theorem, it is implicitly assumed that there is a uniformly bounded solution to (17), i.e., $\{L_k\}$, $k>0$ is a sequence of uniformly bounded matrices. The sufficient conditions for this uniform boundedness to hold are stated in [4] and are not repeated here in the statement of the following theorem.

Theorem 3.1: Consider a hidden Markov model with the system matrices P, C as given in Section I. Suppose that Assumptions 1.1, 1.2, 1.3, 3.1, and 3.2 hold. Consider the exact smoothing recursions given by (6) and (7). Then, there exists a large enough but finite k_0 such that $\forall k\geq k_0$, an $O(\epsilon^2)$ approximation to $\zeta^{(i)}_{j,k}$, $\eta^{(i)}_{j,k}$ is given by $\tilde{\zeta}^{(i)}_{j,k}$ and $\tilde{\eta}^{(i)}_{j,k}$, respectively, via (21) $\forall i=1,2,\ldots,n$. Furthermore, for $j\geq k_0-\Delta$, an $O(\epsilon^2)$ approximation to the exact fixed-lag smoothed estimate $\Pi_{j|j+\Delta}$ (which is denoted as $\tilde{\Pi}_{j|j+\Delta}$) is given via (8) with $\zeta^{(i)}_{j,j+\Delta}$, $\eta^{(i)}_{j,j+\Delta}$ replaced by $\tilde{\zeta}^{(i)}_{j,j+\Delta}$, $\tilde{\eta}^{(i)}_{j,j+\Delta}$, respectively, $\forall i=1,2,\ldots,n$. Similarly, for $j\geq k_0-\Delta$, an $O(\epsilon^2)$ approximation to the exact aggregate smoothed estimate $\zeta_{j|j+\Delta}$ is given by $\tilde{\Pi}_{j|j+\Delta}W_1$.

Finally, we note that the aggregate smoothed estimates can also be obtained by using Courtois' $N\times N$ aggregate matrix and the aggregate observation probability matrix of size $M\times N$. The subsequent smoothed estimate can be obtained by substituting

$$P^{\text{ag}}=I_N+\epsilon\left[V_1-V_1AW_2(V_2AW_2)^{-1}V_2\right]BW_1 \tag{22a}$$

$$C^{\text{ag}}=(C^{\text{ag}}_{mn})=(\Pr(Y_k=m|X_k\in S_n)) \tag{22b}$$

for P and C, respectively, into (1) and (2) with other appropriate changes in dimensions. Following [4], it can be shown that this aggregation technique results in the same $O(\epsilon^2)$ approximations to the exact aggregate smoothed estimates. In fact, one can use the aggregation technique of [10] to obtain aggregate smoothed estimates, which happens to provide a slightly better approximation. It is also worth noting that the aggregate smoothed estimates obtained through using these aggregation techniques of [5] or [10] require fewer computations than using the method suggested in Theorem 3.1 (see also Table I).[1]

However, we emphasize the facts that 1) we cannot obtain reduced-complexity *full-order* smoothed estimates using any of these aggregation techniques, and 2) we cannot easily extend these aggregation ideas to the case where the state to observation transition probability matrix C is a small perturbation of the "block-structured" form, as specified by Assumption 1.2. In this case, it was observed in [4] that some of the aggregation techniques may become *ad hoc*. It is also worth noting that our algorithm provides a *systematic* method of obtaining reduced-order computations to aggregate as well as full-order smoothed estimates. It was shown in [4] that in a special case where the sub-Markov chains given by $I_{n_i}+A_{ii}$ are independent and identically distributed (i.i.d.), our method can be used to obtain $O(\epsilon^3)$ approximations to the aggregate and full-order smoothed estimates, whereas none of the aggregation techniques of [5] or [10] can be adapted to achieve this.

We now indicate the savings in computations when the decoupling transformation is used with the assumption that the matrices in (13) can be precomputed and stored in memory due

[1]We have only indicated the number of divisions required in principle for normalization; in practice, more divisions are necessary to prevent numerical underflows in the calculations.

to the finite discrete nature of the observation process. The comparison of the savings in computations at each j for a given smoothing lag Δ are given in Table I. Note that while for aggregate smoothed estimates, an algorithm adapted from Courtois' method require fewer computations, for full-order smoothed estimates, our algorithm provides a significant reduction in computational requirements compared with exact calculations. The aggregation methods (including Courtois' method), of course, cannot be adapted to compute full-order smoothed estimates.

In the next section, we provide a comparative simulation study involving our algorithm and algorithms adapted from the aggregation techniques in [5] and [10].

IV. Simulations

In this section, we will compare the exact full-order and aggregate smoothed estimates with those obtained using our decoupling scheme, as well as commenting on the results obtainable from the aggregation schemes of [10] and [5]. The results were obtained using 20 000 data points, averaged over 20 sets, with a fixed smoothing lag of 50. This smoothing lag was chosen by noting that there was no significant improvement in smoothing performance when the lag exceeded this value. The aim is to illustrate the claim of $O(\epsilon^2)$ approximation that was made in Section III. The error criteria used is $(1/T)\sum_{k=0}^{T-1} |\Pi_{k|k+\Delta} - \Pi_k|$ (which is an estimate of $\mathbb{E}|\Pi_{k|k+\Delta} - \Pi_k|$ as $T \to \infty$), where Π_k denotes the state vector at time k; for the comparison of the various approximate schemes, Π_k is replaced by the approximate smoothed probability vectors.

TABLE II
Comparison of Exact (Full-Order) Smoothing WithDecoupling Scheme

	Average approximation error	
ϵ	Exact Smoothing	Decoupling Scheme
0.001	1.22759636430736	1.22759762653319
0.005	1.22291595821324	1.22293122070098
0.01	1.23981392134711	1.23981692582461
0.02	1.26983658972958	1.26983627705158
0.05	1.33095599692731	1.33094864102834
0.1	1.39437014954947	1.39440441531637

We will use the same example as in [4], shown in A, B, C, and $L(0)$ at the bottom of the page.

The difference between exact and approximate smoothing using our decoupling scheme is shown in Table II. A comparison of the approximate methods is tabulated in Table III. It is seen that as far as aggregate smoothed estimates are concerned, our decoupling method results in the same performance as a method adapted from Courtois' aggregation procedure. The $O(\epsilon^2)$ approximation can be clearly seen in Fig. 1. However, the aggregation matrix of Aldhaheri/Khalil is seen to consistently outperform our method by small amounts. Nevertheless, we reiterate that our algorithm is a *systematic* method for computing aggregate conditional densities with reduced-order computations. While other aggregation methods can be adapted to achieve comparable approximations to the aggregate smoothed estimates, this fact, along with the order of approximation achievable [$O(\epsilon^2)$ for an adapted version of Courtois' method] has not been established anywhere else. Furthermore, our method offers the provision of computing the

$$A = \begin{bmatrix} -0.35 & 0.25 & 0.1 & 0 & 0 & 0 & 0 & 0 \\ 0.15 & -0.65 & 0.5 & 0 & 0 & 0 & 0 & 0 \\ 0.55 & 0.15 & -0.7 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -0.3 & 0.3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.3 & -0.3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -0.4 & 0.25 & 0.15 \\ 0 & 0 & 0 & 0 & 0 & 0.3 & -0.42 & 0.12 \\ 0 & 0 & 0 & 0 & 0 & 0.15 & 0.35 & -0.5 \end{bmatrix}$$

$$B = \begin{bmatrix} 0.1 & 0.15 & -1.0 & 0.6 & 0.05 & 0 & 0.05 & 0.05 \\ 0 & 0.1 & -0.9 & 0.5 & 0.05 & 0.05 & 0.1 & 0.1 \\ 0.01 & 0.01 & -0.4 & 0.2 & 0.05 & 0.05 & 0.04 & 0.04 \\ 0.02 & 0.42 & 0.01 & 0.01 & -0.61 & 0.025 & 0.1 & 0.025 \\ 0.45 & 0.01 & 0.4 & -1.0 & 0.01 & 0.1 & 0.01 & 0.02 \\ 0.01 & 0.05 & 0.01 & 0.01 & 0.05 & 0.01 & -0.15 & 0.01 \\ 0.03 & 0.01 & 0.03 & 0.04 & 0.01 & 0.01 & 0.01 & -0.14 \\ 0.01 & 0.05 & 0.01 & 0.01 & 0.05 & -0.16 & 0.01 & 0.02 \end{bmatrix}$$

$$C = \begin{bmatrix} 0.25 & 0.25 & 0.25 & 0.4 & 0.4 & 0.57 & 0.57 & 0.57 \\ 0.32 & 0.32 & 0.32 & 0.5 & 0.5 & 0.16 & 0.16 & 0.16 \\ 0.43 & 0.43 & 0.43 & 0.1 & 0.1 & 0.27 & 0.27 & 0.27 \end{bmatrix}$$

$$L(0) = \begin{bmatrix} -0.25 & -0.25 & 0 & 0 & 0 \\ 0 & 0 & -0.5 & 0 & 0 \\ 0 & 0 & 0 & -0.4048 & -0.2121 \end{bmatrix}$$

TABLE III
AVERAGE APPROXIMATION ERROR OF VARIOUS AGGREGATE SMOOTHING SCHEMES RELATIVE TO EXACT AGGREGATE SMOOTHING

	Average approximation error		
ϵ	Decoupling Scheme	Adapted Courtois	Adapted Aldhaheri/Khalil
0.001	0.00000094826370	0.00000094826370	0.00000070769290
0.005	0.00002408049681	0.00002408049681	0.00002002193453
0.01	0.00009652988268	0.00009652988268	0.00007453286536
0.02	0.00034936827430	0.00034936827430	0.00027410254254
0.05	0.00165916284703	0.00165916284703	0.00129521052786
0.1	0.00467331907768	0.00467331907768	0.00351863089563

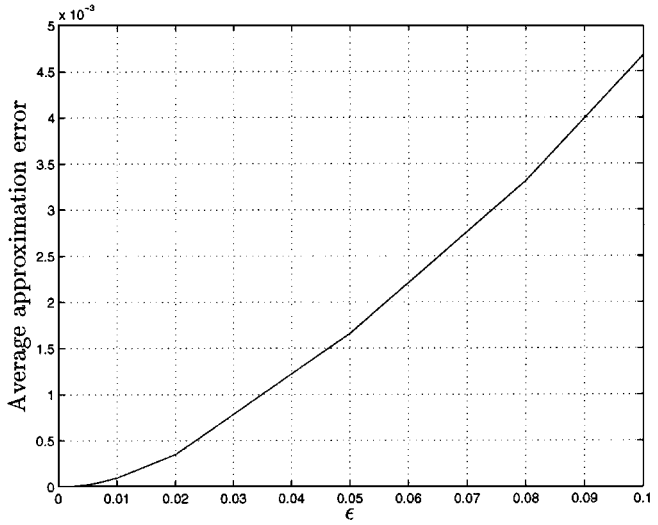

Fig. 1. Aggregate smoothing via decoupling relative to the exact aggregate smoothing.

full-order estimates, whereas the aggregation methods cannot be extended to achieve such full-order computations. It was also shown in [4] that for a special class of these HMMs, when the underlying NCDMC is i.i.d [i.e., $(I_{n_i} + A_{ii})$ has identical rows for each i], then one can obtain $O(\epsilon^3)$ to the aggregate filtered estimates with large computational savings using our method, whereas none of the aggregation methods discussed here can be adapted to achieve such savings. Such results also hold for smoothing. We do not include any simulation results here, but for similar filtering results, see [4].

V. CONCLUSIONS

In this paper, we propose an algorithm for obtaining approximate smoothed state estimates for a class of HMMs with (possibly large-scale) underlying NCDMCs. These approximations are of order $O(\epsilon^2)$, and they result in substantial computational savings.

REFERENCES

[1] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–285, Feb. 1989.
[2] V. Krishnamurthy, S. Dey, and J. P. LeBlanc, "Blind equalization of IIR channels using hidden Markov models and extended least squares," *IEEE Trans. Signal Processing*, vol. 43, pp. 2994–3006, Dec. 1995.
[3] S. Dey, V. Krishnamurthy, and T. Salmon-Leagagneur, "Estimation of Markov modulated time-series via EM algorithm," *IEEE Signal Processing Lett.*, vol. 1, pp. 153–155, Oct. 1994.
[4] S. Dey, "Reduced-complexity filtering for partially observed nearly completely decomposable Markov chains," *IEEE Trans. Signal Processing*, vol. 48, pp. 3334–3344, Dec. 2000.
[5] P. J. Courtois, *Decomposability, Queuing and Computer Systems Applications*. New York: Academic, 1977.
[6] G. Yin and Q. Zhang, *Continuous-Time Markov Chains and Applications: A Singular Perturbation Approach*. New York: Springer-Verlag, 1998.
[7] D. Tse, R. Gallager, and J. Tsitsiklis, "Statistical multiplexing of multiple time-scale Markov streams," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 1028–1038, June 1995.
[8] K. Kontovasilis and N. Mitrou, "Markov-modulated traffic with nearly complete decomposability characteristics and associated fluid queuing models," *Adv. Appl. Probab.*, vol. 27, pp. 1144–1185, 1995.
[9] C. D. Meyer, "Stochastic complementation, uncoupled Markov chains, and the theory of nearly reducible systems," *SIAM Rev.*, vol. 31, pp. 240–272, June 1989.
[10] R. W. Aldhaheri and H. K. Khalil, "Aggregation of the policy iteration method for nearly completely decomposable Markov chains," *IEEE Trans. Automat. Contr.*, vol. 36, pp. 178–187, Feb. 1991.
[11] V. G. Gatisgory and A. A. Prevozvanskii, "Aggregation of states in a Markov chain with weak interactions," *Kybernetica*, pp. 91–98, May–June 1975.
[12] F. Delebecque and J. P. Quadrat, "Optimal control of Markov chains admitting strong and weak interactions," *Automatica*, vol. 17, pp. 281–296, 1981.
[13] R. G. Phillips and P. V. Kokotovic, "A singular perturbation approach to modeling and control of Markov chains," *IEEE Trans. Automat. Contr.*, vol. AC-26, pp. 1087–1094, 1981.
[14] M. Coderch, A. S. Willsky, S. S. Sastry, and D. A. Casatanon, "Hierarchical aggregation of singularly perturbed finite state Markov processes," *Stochastics*, vol. 8, pp. 259–289, 1983.
[15] J. R. Rohlicek and A. S. Willsky, "The reduction of perturbed Markov generators: An algorithm exposing the role of transient states," *J. Assoc. Comput. Mach.*, vol. 35, pp. 675–696, 1988.
[16] F. Delebecque, J. P. Quadrat, and P. V. Kokotovic, "A unified view of aggregation and coherency in networks and Markov chains," *Int. J. Contr.*, vol. 40, pp. 939–952, 1984.
[17] M. Abbad, J. Filar, and T. R. Bielecki, "Algorithms for singularly perturbed limiting average Markov control problems," *IEEE Trans. Automat. Contr.*, vol. 37, pp. 1421–1425, Sept. 1992.
[18] M. Abbad and J. Filar, "Perturbation and stability theory for Markov control problems," *IEEE Trans. Automat. Contr.*, vol. 37, pp. 1415–1420, Sept. 1992.
[19] V. Krishnamurthy, "Adaptive estimation of hidden nearly completely decomposable Markov chains with applications in blind equalization," *Int. J. Adaptive Contr. Signal Process.*, vol. 8, pp. 237–260, Aug. 1994.
[20] L. White, R. Mahony, and G. Brushe, "Lumpable hidden Markov models—Model reduction and reduced complexity filtering," *IEEE Trans. Automat. Contr.*, vol. 45, pp. 2297–2306, Dec. 2000.
[21] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*. Englewood Cliffs, NJ: Prentice-Hall, 1979.
[22] L. Shue, B. D. O. Anderson, and S. Dey, "Exponential stability of filters and smoothers for hidden Markov models," *IEEE Trans. Signal Processing*, vol. 46, pp. 2180–2194, Aug. 1998.
[23] P. Kokotovic, H. K. Khalil, and J. O'Reilly, *Singular Perturbation Methods in Control: Analysis and Design*. New York: Academic, 1986.

Louis Shue was born in Vietnam in 1971. He received the B.Sc. and B.E. degrees, both with first class honors, from Monash University, Clayton, Australia, in 1994 and 1996, respectively, and the Ph.D. degree in systems engineering from the Department of Systems Engineering, the Australian National University, Canberra, Australia, in 1999.

He is currently a Research Fellow with the Centre for Signal Processing, Nanyang Technological University, Singapore. His current research interests include speech and image processing, small target tracking in infrared imagery, and smart human–PC interface.

Subhrakanti Dey (M'96) was born in Calcutta, India, in 1968. He the B.Tech. and M.Tech. degrees from the Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology, Kharagpur, in 1991 and 1993, respectively. He received the Ph.D. degree from the Department of Systems Engineering, Research School of Information Sciences and Engineering, Australian National University (ANU), Canberra, in 1996.

He is currently a Senior Lecturer with the Department of Electrical and Electronic Engineering, University of Melbourne, Parkville, Australia, where he has been since February 2000. From September 1995 to September 1997 and September 1998 to February 2000, he was a Postdoctoral Research Fellow with the Department of Systems Engineering, ANU. From September 1997 to September 1998, he was a Postdoctoral Research Associate with the Institute for Systems Research, University of Maryland, College Park. His current research interests include signal processing for eletrocommunications, wireless communications and networks, performance analysis of communication networks, stochastic and adaptive estimation and control, and statistical and adaptive signal processing. He is currently an Academic Staff Member with the ARC Special Research Centre for Ultra-Broadband Information Networks, Department of Electrical and Electronic Engineering, University of Melbourne.