REDUCED COMPLEXITY ESTIMATION FOR LARGE SCALE HIDDEN MARKOV MODELS

Subhrakanti Dey*, Iven Mareels*

* Dept. of Electrical & Electronic Engineering, University of Melbourne, Parkville, Victoria 3010 Australia
e-mail: sdey, iven@ee.mu.oz.au

Abstract

In this paper, we address the problem of reduced-complexity estimation of general large-scale hidden Markov models with underlying nearly completely decomposable discrete-time Markov chains and finite-state outputs. An algorithm is presented that computes $O(\epsilon)$ (where ϵ is the related weak coupling parameter) approximations to the aggregate and full-order filtered estimates with substantial computational savings. These savings are shown to be quite large when the chains have blocks with small individual dimensions. Some simulation studies are presented to demonstrate the performance of the algorithm.

1 Introduction

Reducing computational complexity in optimal estimation and control of large scale Markov chains has been a topic of substantial interest [1], [2], [3], [4]. A class of such Markov chains is also known as "nearly completely decomposable Markov chains". These Markov chains are usually large scale, and show strong interactions within groups and weak interactions between the groups. These chains are usually characterized by transition probability matrices which can be expressed as $P = I_n + A + \epsilon B$ where $I_n + A$ is a block diagonal matrix with the individual blocks being stochastic matrices, the number of blocks being N. These blocks are also termed as "superstates" (as in [5]). Typically $n \gg N$ where n is the total number of states in the chain. The parameter $\epsilon > 0$ is small and acts as the a weak coupling parameter that makes the chain "nearly" completely decomposable (the chain is completely decomposable if $\epsilon = 0$). Applications of such Markov chains to economic modelling, queueing networks and computer systems have been reported in early works such as [6] and [7]. Other applications can also be found in manufacturing systems operating with machines of varying speed [8], communication networks with variable bit rate video traffic [9] etc. In [1], [2] and also [10], [11] the authors focus on obtaining reduced-complexity computation of the stationary distribution of such Markov chains using various aggregation-decomposition methods. The singular perturbation approach to aggregation of finite-state Markov chains has been studied in [12] and [13] amongst others. Much of this work concentrated on obtaining reduced-complexity controllers for these Markov chains, very little attention was paid to the case of partially observed nearly completely decomposable Markov chains. Only [3] and more recently [4] address this problem. The research reported in [3], [4] partially address the case of reduced-complexity state estimation from imperfect observations, but does not provide a systematic way of obtaining reduced-order computations of the filtered estimates, that exploits the near complete decomposability of the Markov chain to arrive at substantial guaranteed computational savings with rigorous performance bounds. The trade-off between accuracy and computational complexity in state estimation of hidden nearly completely decomposable Markov chains is also not explored in [3], [4].

Computational complexity for filtering of hidden Markov models (HMM) with underlying nearly completely decomposable Markov chains was first addressed in our own work [5] and the following paper [14], where algorithms were presented for $O(\epsilon^2)$ approximations to the aggregate and full-order state estimates requiring $O(N^2)$ computations per discrete time instant as opposed to $O(n^2)$ computations for the optimal estimates. These results were proven using ergodicity assumptions on the hidden Markov models (HMM) and for sufficiently small ϵ. However, a restriction was imposed on the state-to-output observation probability matrices, in that they were assumed to have a "block" structure. This implied that the output transitions only depended on the superstates but did not distinguish amongst the states belonging to the same superstate. It was shown in [5] that how this assumption allowed us to obtain a nice decoupling transformation that is uniformly bounded and results in $O(\epsilon^2)$ approximations to the aggregate and full-order filtered estimates and smoothed estimates [14]. It is perhaps not surprising that this special structure of the state-to-output transition matrix allows one to obtain $O(\epsilon^2)$ approximation whereas generally speaking, one would expect to obtain $O(\epsilon)$ approximations. Mathematically, this structure helps us to solve for the decoupling transformation which satisfies a nonlinear matrix equation, using iterative approximation techniques (see [5]). These iterative techniques are easily implemented because the nonlinear term in the equation is weighted only by the weak coupling parameter ϵ, due to the special structure on the state-to-output transition matrix. It is also clear that under any small perturbation (which is of any polynomial order in ϵ) applied to the state-to-output transition matrix with the special block structure, the methods presented in [5] can be applied also, but not to any general state-to-output transition matrix that does not have a block structure or a nearly block

structure.

In this paper, we lift this restriction on the state-to-output transition probability matrix and allow it to be any general transition probability matrix with all positive elements. We concentrate on a discrete-time irreducible nearly completely decomposable Markov chain with finite-state outputs. We provide a systematic way for obtaining reduced-complexity filtered estimates for these HMMs with a prescribed degree of accuracy. The contribution of this work lies in (1) presenting an algorithm that computes guaranteed $O(\epsilon)$ approximations to the aggregate and full-order filtered estimates for a general hidden Markov model with finite-state outputs, (2) showing that these approximations can be calculated with substantial computational savings when the chain has superstates that have small individual dimensions, and (3) demonstrating through simulation studies the effectiveness of the algorithm. The main assumptions that are used involve ergodicity assumptions on the hidden Markov model and that ϵ is sufficiently small. We should also point out that these results have important implications for reduced-complexity estimation for HMMs with continuous-valued outputs as well. However, extensions of some of the rigorous analysis to the continuous-valued output case are still incomplete and part of our ongoing research.

2 Signal Model

A discrete-time nearly completely decomposable Markov chain X_k in a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ comprising of n states is characterized by a transition probability matrix $P^\epsilon \in \mathbb{R}^{n \times n}$

$$P^\epsilon = I_n + A + \epsilon B \qquad (1)$$

where I_n is the identity matrix of order $n \times n$, and A is block diagonal with

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

where $A_{ii} \in \mathbb{R}^{n_i \times n_i}$, $\forall i$, $\sum_i n_i = n$, $\epsilon > 0$ is a small perturbation parameter, and $B \in \mathbb{R}^{n \times n}$. It is clear that there are N blocks in the Markov chain within each of which the dynamics is fast and every so often, the chain leaves one block to visit another. Since ϵ is small, the rate at which these inter-block transitions occur is slow. For all i, $I_{n_i} + A_{ii}$ is row-stochastic, and so is P. Obviously, the row-sums of A and B are zero. Just as in [5], we make the following natural, key assumption:

Assumption 2.1 *The matrices P^ϵ and $I_{n_i} + A_{ii}$, $\forall i$ are irreducible.*

Remark 1 Notice that the above Assumption 2.1 guarantees the existence of a unique stationary distribution of P^ϵ and $I_{n_i} + A_{ii} \forall i$. There is one difference however, that should be noted: while the stationary distribution of P^ϵ depends on ϵ, those of $I_{n_i} + A_{ii}$, $i = 1, 2, \ldots, N$ do not.

The states of the Markov chain are observed through another stochastic process (observation or measurement process) Y_k. For the analytical development of the paper, we assume that Y_k belongs to a discrete set of finite cardinality. More specifically, $Y_k \in \{1, 2, \ldots, M\}$ and $\mathcal{P}(Y_k = i \mid X_k = j) = c_{ij}$, $i = 1, 2, \ldots, M$, $j = 1, 2, \ldots, n$. Such a signal model (irrespective of whether the underlying Markov chain is nearly completely decomposable or not) is also known as a hidden Markov model (HMM). Note also that $\sum_i c_{ij} = 1$, $\forall j$, that is, the observation probability matrix $C = (c_{ij})$ is column-stochastic. We assume that $c_{ij} > 0$, $\forall i, j$.

It was shown in [5] that reduced-order computations ($O(N^2)$ at each discrete time instant) for the filtered state estimate (for extension of these results to smoothing, see [14]) can be obtained with $O(\epsilon^2)$ approximation when the state-to-output transition matrix C has a special structure, namely, $c_{ij} = \bar{c}_{il}$, $\forall j \in S_l$, $\forall i$. Since typically, $n \gg N$, these computational savings are substantial. In this paper, we remove this restriction on C.

In what follows, we will show that even with no restrictions on C, one can obtain $O(\epsilon)$ approximations to the filtered estimates with substantial computational reductions when the individual block sizes n_i, $i = 1, 2, \ldots, N$ are not too large. As an example, one will obtain substantial reduction in computational complexity if $n = 200, N = 25, n_i = 8$, $\forall i$ instead of $n = 200, N = 4, n_i = 50$, $\forall i$. A table consisting of exact numbers of multiplications, additions and divisions per discrete time instant will be provided later to compare the computational complexity of exact state estimation and the approximate state estimation.

Like in [5], we term the N blocks as "superstates". Notice that the probability (or conditional probability) of the Markov chain belonging to a particular superstate is the sum of probabilities (or conditional probabilities) of the chain belonging to its constituent states. We denote the l-th superstate by S_l, $l = 1, 2, \ldots, N$. Without loss of generality, $S_1 = \{1, 2, \ldots, n_1\}, S_2 = \{n_1 + 1, n_1 + 2, \ldots, n_1 + n_2\}$ etc. We also term (like in [5]) the filtered estimate vector with l-th element being $P(X_k \in S_l | \mathcal{Y}_k)$ as the *aggregate* filtered estimate, $l = 1, 2, \ldots, N$. Here, $\mathcal{Y}_k$ is the complete filtration generated by the σ algebra $\sigma(Y_0, Y_1, \ldots, Y_k)$.

3 State estimation of hidden Markov models

It is well known that the conditional filtered state estimate for a hidden Markov model is defined in the following way:

$$\alpha_k(i) = \mathcal{P}(X_k = i \mid \mathcal{Y}_k) \qquad (2)$$

where $\mathcal{Y}_k$ is the complete filtration generated by the σ algebra $\sigma(Y_0, Y_1, \ldots, Y_k)$. Defining the row vector $\alpha_k \triangleq (\alpha_k(1)\ \alpha_k(2) \ldots \alpha_k(n))$, one can obtain the following recursion [15]

$$\alpha_{k+1} = \frac{1}{Z_{k+1}} \alpha_k P^\epsilon C(Y_{k+1}), \ \alpha_0 = \pi_0 C(Y_0) \qquad (3)$$

where $C(Y_{k+1}) = diag\{c_{i1} c_{i2} \ldots c_{in}\}$ if $Y_{k+1} = i$ and $Z_{k+1} = \alpha_k P^{\epsilon} C(Y_{k+1}) 1_n$ is the normalization factor (with 1_n being the n-length column vector of all 1-s) and π_0 is the row vector representing the initial distribution of X_0.

Note that the aggregate filtered state estimate is given by

$$\zeta_k(j) = \mathcal{P}(X_k \in S_j \mid \mathcal{Y}_k) = \sum_{l \in S_j} \mathcal{P}(X_k = l \mid \mathcal{Y}_k) \quad (4)$$

Obviously the row vector $\zeta_k \triangleq (\zeta_k(1)\zeta_k(2)\ldots\zeta_k(N))$ denotes the aggregate filtered state estimates and can be represented by

$$\zeta_k = \alpha_k W_1 \quad (5)$$

where $W_1 \in \mathbb{R}^{n \times N}$ is given by

$$W_1 = \begin{bmatrix} 1_{n_1} & 0 & 0 & 0 \\ 0 & 1_{n_2} & \cdot & \cdot \\ \cdot & 0 & \cdot & 0 \\ 0 & \cdot & \cdot & 1_{n_N} \end{bmatrix}$$

Following the same techniques as in [11] (also used in [5]), another matrix $W_2 \in \mathbb{R}^{n \times (n-N)}$ is chosen such that the transformation $\Gamma = [W_1 \ W_2]$ is nonsingular. Let $\eta_k = \alpha_k W_2$. Let also $\Gamma^{-1} = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}$ where obviously $V_1 \in \mathbb{R}^{N \times n}$ and $V_2 \in \mathbb{R}^{(n-N) \times n}$. We choose W_2 as in [5], the resulting V_1, V_2 can also be found in [5].

Now rewrite (3) as

$$[\zeta_{k+1} \ \eta_{k+1}] = \frac{1}{Z_{k+1}}[\zeta_k \ \eta_k] \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} P^{\epsilon} C(Y_{k+1})[W_1 \ W_2]$$

$$= \frac{1}{Z_{k+1}}[\zeta_k \ \eta_k] \begin{bmatrix} \tilde{A}_{11}^k & \tilde{A}_{12}^k \\ \tilde{A}_{21}^k & \tilde{A}_{22}^k \end{bmatrix} \quad (6)$$

where $\tilde{A}_{11}^k = \tilde{A}_1^k + \epsilon \tilde{B}_1^k$, $\tilde{A}_{12}^k = \tilde{A}_2^k + \epsilon \tilde{B}_2^k$, $\tilde{A}_{21}^k = \tilde{C}_1^k + \epsilon \tilde{D}_1^k$, $\tilde{A}_{22}^k = \tilde{C}_2^k + \epsilon \tilde{D}_2^k$ and they are given by the following equations:

$$\tilde{A}_1^k = V_1(I_n + A)C(Y_{k+1})W_1, \quad \tilde{B}_1^k = V_1 BC(Y_{k+1})W_1$$
$$\tilde{A}_2^k = V_1(I_n + A)C(Y_{k+1})W_2, \quad \tilde{B}_2^k = V_1 BC(Y_{k+1})W_2$$
$$\tilde{C}_1^k = V_2(I_n + A)C(Y_{k+1})W_1, \quad \tilde{D}_1^k = V_2 BC(Y_{k+1})W_1$$
$$\tilde{C}_2^k = V_2(I_n + A)C(Y_{k+1})W_2, \quad \tilde{D}_2^k = V_2 BC(Y_{k+1})W_2 \quad (7)$$

Clearly, (6) can be carried out in two steps ($\forall k$):
Step 1: Calculate the unnormalized quantities $\zeta_{k+1}^u, \eta_{k+1}^u$ according to the following recursion:

$$[\zeta_{k+1}^u \ \eta_{k+1}^u] = [\zeta_k \ \eta_k] \begin{bmatrix} \tilde{A}_{11}^k & \tilde{A}_{12}^k \\ \tilde{A}_{21}^k & \tilde{A}_{22}^k \end{bmatrix}$$

$$(8)$$

where $\zeta_0 = \alpha_0 W_1$, $\eta_0 = \alpha_0 W_2$.
Step 2: Normalize $\zeta_{k+1}^u, \eta_{k+1}^u$ by the normalization factor $Z_{k+1} = \sum_{j=1}^{N} \zeta_{k+1}^u(j)$. Note also that

$$\alpha_k = \zeta_k V_1 + \eta_k V_2 \quad (9)$$

As observed in [5], we remind the readers again that for the above choices of V_1, V_2, W_1, W_2, the matrices $\tilde{A}_1^k, \tilde{A}_2^k, \tilde{C}_1^k$ and $\tilde{C}_2^k$ are block diagonal matrices for all k, more specifically, $\tilde{A}_1^k$ is diagonal, $\tilde{A}_2^k$ is block diagonal with the i-th block being a row vector of size $n_i - 1$, $\tilde{C}_1^k$ is block diagonal with the i-th block being a column vector of size $n_i - 1$ and $\tilde{C}_2^k$ is block diagonal with the i-th block being a square matrix of size $(n_i - 1) \times (n_i - 1)$. Also, since the matrices $\tilde{A}_1^k, \tilde{A}_2^k, \tilde{C}_1^k$ and $\tilde{C}_2^k$ depend only on Y_{k+1} which is finitely-valued, one can essentially pre-compute the matrices $\tilde{A}_1^k, \tilde{A}_2^k, \tilde{C}_1^k$ and $\tilde{C}_2^k$ for each possible value of Y_{k+1} and store them in a lookup table. During the filtering operations, as and when we get a specific observation, we can obtain the corresponding matrices by table lookup.

4 $O(\epsilon)$ approximate reduced-order filters

In this section, we are concerned with obtaining reduced-complexity aggregate and full-order filters with a prescribed degree of accuracy when ϵ is "sufficiently" small. While the techniques used in this section are similar to that of [5], the results are substantially different due to the absence of a special structure on the state-to-output transition probability matrix C, as assumed in [5].

Following the approach in [5], we use a standard decoupling technique to obtain the transformed variables $[\bar{\zeta}_k \ \bar{\eta}_k]$ that are given by

$$[\bar{\zeta}_k \ \bar{\eta}_k] = [\zeta_k \ \eta_k] \begin{bmatrix} I_N & L_k \\ 0 & I_{n-N} \end{bmatrix} \quad (10)$$

This also implies that

$$[\zeta_k \ \eta_k] = [\bar{\zeta}_k \ \bar{\eta}_k] \begin{bmatrix} I_N & -L_k \\ 0 & I_{n-N} \end{bmatrix} \quad (11)$$

Note that one can relate the unnormalized versions of $\bar{\zeta}_k, \bar{\eta}_k$, denoted by $\bar{\zeta}_k^u, \bar{\eta}_k^u$ respectively, by the same decoupling transformation, (since the normalization factor is the same):

$$[\zeta_k^u \ \eta_k^u] = [\bar{\zeta}_k^u \ \bar{\eta}_k^u] \begin{bmatrix} I_N & -L_k \\ 0 & I_{n-N} \end{bmatrix} \quad (12)$$

Here $\{L_k \in \mathbb{R}^{N \times (n-N)}\}$ is assumed to be a sequence of uniformly bounded time-varying matrices to be solved for. Using this together with (6), one can obtain the following recursion in the transformed variables

$$[\bar{\zeta}_{k+1}\bar{\eta}_{k+1}] = \frac{1}{Z_{k+1}}[\bar{\zeta}_k\bar{\eta}_k] \begin{bmatrix} \tilde{A}_{11}^k - L_k \tilde{A}_{21}^k & 0 \\ \tilde{A}_{21}^k & \tilde{A}_{21}^k L_{k+1} + \tilde{A}_{22}^k \end{bmatrix}$$

$$(13)$$

where L_k was chosen such that the upper right hand element in RHS of (13) is zero, or

$$(\tilde{A}_{11}^k - L_k \tilde{A}_{21}^k)L_{k+1} = L_k \tilde{A}_{22}^k - \tilde{A}_{12}^k, \ L_0 = 0 \quad (14)$$

Note that one can recursively solve for L_k from the above equation provided $\tilde{A}_{11}^k - L_k \tilde{A}_{21}^k$ is invertible for every k. This involves multiplications of matrices which are not necessarily

sparse at each k and requires a large number of computations. A well known technique [16] is to exploit the fact that $\epsilon > 0$ is a small positive number and truncate a power series expansion of L_k in ϵ at some finite power. However, this approximation is only valid if L_k is uniformly bounded.

In [5] the uniform boundedness of $\{L_k\}$ was shown using the irreducibility condition stated in Assumption (2.1) and the special structure assumed on C (see [5] for details) and a number of inequalities which were jointly sufficient to guarantee that ϵ is sufficiently small. The proof in [5] crucially depended on the special structure of C which implied that the nonlinear term in (14) was of $O(\epsilon)$ ($C_1^k = 0$).

Due to the generalized nature of C in the current work, we do not have this convenience. We proceed along a different path to establish the boundedness of $\{L_k\}$ for a sufficiently small ϵ. We first make an assumption that guarantees the existence of a uniformly bounded $L_k(0)$, the solution to (14) when $\epsilon = 0$ ($\forall k$). Then, we argue that since L_k, as a solution of (14), is a continuous function of ϵ in a small neighbourhood of $\epsilon = 0$, L_k will also be uniformly bounded for a sufficiently small ϵ.

Notice that $L_k(0)$ satisfies the following recursion (see also (7)):

$$L_{k+1}(0) = (\tilde{A}_1^k - L_k(0)\tilde{C}_1^k)^{-1}(L_k(0)\tilde{C}_2^k - \tilde{A}_2^k), \ L_0(0) = 0 \tag{15}$$

The difference with [5] is that $\tilde{C}_1^k \neq 0$ in the current situation. Observing that $\tilde{A}_1^k$ is diagonal and nonsingular, one can rewrite (15) as the following:

$$
\begin{aligned}
L_{k+1}(0) &= (\tilde{A}_1^k)^{-1}L_k(0)\tilde{C}_2^k - (\tilde{A}_1^k)^{-1}\tilde{A}_2^k \\
&\quad + [(\tilde{A}_1^k)^{-1}L_k(0)\tilde{C}_1^k L_{k+1}(0)], \ L_0(0) = 0
\end{aligned} \tag{16}
$$

Below, we use the vec notation associated with a matrix ($\in \mathbb{R}^{m \times n}$) $A = [A_{*1}A_{*2}\ldots A_{*n}]$ which is defined as

$$vec(A) = \begin{bmatrix} A_{*1} \\ A_{*2} \\ \vdots \\ A_{*n} \end{bmatrix}$$

where A_{*i} denotes the i-th column of the matrix A. We also use the $\otimes$ notation to denote the Kronecker product of two matrices. Using the notations $l_k' = vec(L_k(0)')$, $m_0' = -vec\left(((\tilde{A}_1^0)^{-1}\tilde{A}_2^0)'\right)$, and $g_k'(L_{k-1}(0), L_k(0)) = vec\left((-(\tilde{A}_1^k)^{-1}\tilde{A}_2^k + (\tilde{A}_1^k)^{-1}L_{k-1}(0)\tilde{C}_1^k L_k(0))'\right)$, (17) can be rewritten as

$$l_{k+1} = l_k((\tilde{A}_1^k)^{-1} \otimes \tilde{C}_2^k) + g_{k+1}(L_k(0), L_{k+1}(0)), \ l_0 = 0 \tag{17}$$

Clearly, repeating this recursion, one can write $\forall k \geq 1$,

$$
\begin{aligned}
l_{k+1} &= m_0 \prod_{j=1}^{k} M_{j+1} + \Big[\sum_{p=0}^{k-1} g_{k+1-p}(L_{k-p}(0), L_{k+1-p}(0)) \\
&\quad \times \prod_{q=p-1}^{0} M_{k+1-q}\Big], \ l_1 = m_0
\end{aligned} \tag{18}
$$

where $M_{k+1} = (\tilde{A}_1^k)^{-1} \otimes \tilde{C}_2^k$ and $\prod_{q=-1}^{0} M_{k+1-q} = I_{N \times (n-N)}$.

Let us now introduce the following notations. We denote by $\|.\|_\infty$ the l_∞ vector norm, and by $\||.|\|_\infty$ the corresponding induced maximum absolute row sum matrix norm.

In what follows, we will be looking for sufficient conditions such that $L_k(0)$, $\forall k$ belongs to a compact set $\mathcal{D} \triangleq \{L : \|vec(L')\|_\infty < \bar{L}\}$. Notice also that using the special structures of $\tilde{A}_1^k, \tilde{A}_2^k, \tilde{C}_1^k$ and the fact that Y_{k+1}, $\forall k$ can only take finitely many values, one can easily obtain the following over bounds $\bar{a}_1, \bar{c}_1$ such that if $L_k(0) \in \mathcal{D}$, then $\forall k$, $\|g_{k+1}(L_k(0), L_{k+1}(0))\|_\infty < \bar{g} = (\bar{a}_1 + \bar{c}_1(\max_i n_i - 1)\bar{L}^2)$.

Denote $\rho(Y_{k+1}) = \||M_{k+1}|\|_\infty$. Denote π_Y as the stationary distribution of Y_k such that $\pi_Y(m) = \sum_i P(Y_k = m|X_k = i)\pi(i) = \sum_i c_{mi}\pi(i)$ where π_i is the i-th component of the stationary distribution of X_k such that $\pi P = \pi$. Denote the expectation taken with respect to the stationary distribution π_Y as E_{π_Y}.

We now make the following assumptions which are easy to verify and are sufficient to guarantee the uniform boundedness of $L_k(0)$, $\forall k$.

Assumption 4.1

$$E_{\pi_Y}(\ln \rho(Y_{k+1})) = -\bar{\lambda} < 0, \ \forall k \tag{19}$$

Remark 2 Note that a similar set of assumptions was made (Assumption 4.1 & 4.2) in [5] to obtain the uniform boundedness of L_k for a sufficiently small ϵ. These two assumptions played a crucial role in guaranteeing the uniform boundedness of L_k due to the special structure of the C matrix where the observation probabilities only depended on the superstates and were identical for all the states in the same superstate (for each observation). Since, in the current work, we do not assume such a special structure on C, we cannot prove a uniform boundedness of L_k directly, but make Assumption 4.1 that guarantees the uniform boundedness of $L_k(0)$. Later an additional Assumption 4.2 is made to prove that for sufficiently small $\epsilon > 0$, one can show the boundedness of L_k using continuity arguments.

Observe that Assumption 4.1 guarantees that $\exists$ a t^* such that $\forall k \geq t^* + 1$, $\||\prod_{j=1}^{k} M_{j+1}|\|_\infty < \exp(-(t^* + 1)\bar{\lambda})$ (see [5] for more details). Using this result, we now state the following Lemma (a similar proof can be found in [5]):

Lemma 4.1 Suppose Assumption 4.1 holds. Suppose also that there exists an $\bar{L} > 0$ such that

$$\|m_0\|_\infty \exp\left(-(t^* + 1)\bar{\lambda}\right) + \bar{g}\bar{S} + \bar{g}\frac{\exp\left(-(t^* + 1)\bar{\lambda}\right)}{1 - \exp(-\bar{\lambda})} < \bar{L} \tag{20}$$

where $\bar{S} = \sum_{p=0}^{t^*} \prod_{q=p-1}^{0} \||M_{t^*-q}|\|_\infty$ is a finite number and recall that $\bar{g} = \bar{a}_1 + \bar{c}_1(\max_i n_i - 1)\bar{L}^2$. Then there exists a uniformly bounded solution $L_k(0), \forall k$ to (17) such that $L_k(0) \in \mathcal{D}$.

Now that we have identified a reasonable condition under which the uniform boundedness of $L_k(0)$ holds, we argue that L_k, as a solution of (14), is a continuous function of ϵ in a small neighbourhood of $\epsilon = 0$, and therefore it will also be uniformly bounded for a sufficiently small ϵ. This can be shown easily if the following assumption holds:

Assumption 4.2 $a_L c_L < 1$ *and there exists a positive real number* L_δ *such that*

$$a_L c_L L_\delta + b_L < L_\delta \qquad (21)$$

where there exist positive real numbers a_L, b_L, c_L *such that* $\||(\tilde{A}_1^k - L_k(0)\tilde{C}_1^k)^{-1}\|| < a_L$, $\||L_k(0)\tilde{D}_2^k + L_k(0)\tilde{D}_1^k L_{k+1}(0) - \tilde{B}_2^k - \tilde{B}_1^k L_{k+1}(0)\|| < b_L$ *and* $\||\tilde{C}_2^k + \tilde{C}_1^k L_{k+1}(0)\|| < c_L$ *where* $\||.\||$ *denotes a suitable matrix norm.*

Note that Assumption 4.2 guarantees the uniform boundedness of $\frac{\partial L_k}{\partial \epsilon}|_{\epsilon=0}$ such that $\||\frac{\partial L_k}{\partial \epsilon}|_{\epsilon=0}\|| < L_\delta$. This allows us to use $L_k(0)$ as an $O(\epsilon)$ approximation to $L_k, \forall k$.

In addition, just as in [5] we make the following assumption which has been justified in [5] and is seen to be satisfied through simulation studies.

Assumption 4.3 *The evolution* $z_{k+1} = z_k(\tilde{A}_{21}^k L_{k+1} + \tilde{A}_{22}^k)$ *where* $z_k \in \mathbb{R}^{n-N}$ *has an exponentially stable solution.*

Note that Assumption 4.3 guarantees that $\bar{\eta}_k \to 0$ asymptotically (using (13)). The rate of this decay is determined by the eigenvalues of $\tilde{A}_{21}^k L_{k+1} + \tilde{A}_{22}^k$ and how close they are to the origin. This also implies that there is a finite integer k_0 such that for $k > k_0, |\bar{\eta}_k|$ is $O(\epsilon)$.

Using the above results, (12) and an unnormalized version of (13)), one can now write $O(\epsilon)$ approximations (for a sufficiently large k) to the unnormalized estimates ζ_k^u, η_k^u as

$$
\begin{aligned}
\tilde{\zeta}_{k+1}^u &= \tilde{\zeta}_k^u(\tilde{A}_1^k - L_k(0)C_1^k) \\
\tilde{\eta}_{k+1}^u &= -\tilde{\zeta}_{k+1}^u L_k(0)
\end{aligned}
\qquad (22)
$$

We also assume that ϵ is small enough such that the normalization procedure (division by $\tilde{Z}_{k+1}$ in (13)) does not affect the order of approximation of the unnormalized quantity, that is, $\frac{1}{Z_k} = \frac{1}{\tilde{Z}_k} + O(\epsilon)$ for some $k \geq k_1$, where $\tilde{Z}_k = \tilde{\zeta}_k^u 1_N$.

The above discussion may be summarized in the form of the following Theorem, establishing our main result on reduced-order approximations to the filtered estimates ζ_k, η_k:

Theorem 1 *Suppose Assumptions 2.1, 4.1, 4.2, and 4.3 hold. Also, suppose there exists an* $\bar{L} > 0$ *such that the inequality (20) holds. Then there exists a large enough but finite* k_2 *such that for* $k \geq k_2$, *an* $O(\epsilon)$ *approximation for* ζ_{k+1} *denoted by* $\tilde{\zeta}_{k+1}$ *(and the unnormalized version by* $\tilde{\zeta}_{k+1}^u$*) can be obtained recursively by the following two steps:*

$$
\begin{aligned}
\tilde{\zeta}_{k+1}^u &= \tilde{\zeta}_k[\tilde{A}_1^k - L_k(0)\tilde{C}_1^k], \quad \tilde{\zeta}_{k_2-1} = \zeta_{k_2-1} \\
\tilde{\zeta}_{k+1} &= \frac{1}{[\tilde{\zeta}_{k+1}^u 1_N]} \tilde{\zeta}_{k+1}^u
\end{aligned}
\qquad (23)
$$

Similarly, an $O(\epsilon)$ *approximation for* η_k *(for* $k \geq k_2$ *) is given by*

$$\tilde{\eta}_k = -\tilde{\zeta}_k L_k(0) \quad \tilde{\eta}_{k_2-1} = \eta_{k_2-1} \qquad (24)$$

Similarly, an $O(\epsilon)$ *approximation to* α_k *(the full-order normalized filtered estimate) is given by*

$$\tilde{\alpha}_k = \tilde{\zeta}_k V_1 + \tilde{\eta}_k V_2 \qquad (25)$$

We now provide a table comparing the numbers of additions, multiplications and divisions per discrete time instant in computing the exact aggregate filtered estimates (ζ_k) and the corresponding $O(\epsilon)$ approximation ($\tilde{\zeta}_k$, as given by Theorem 1).

No: of computations	Exact	$O(\epsilon)$
Addition	$n^2 - 1$	$2N^2 + \sum_{i=1}^{N} n_i^2 - n - 1$
Multiplication	n^2	$\sum_{i=1}^{N} n_i^2 + nN - N$
Division	N	$2N$

Table 1: Table of number of computations per time instant for exact and approximate $O(\epsilon)$ aggregate filtering

It is clear that there will be substantial savings in computations when the Markov chain has a reasonable number of superstates of small individual dimensions. For example, for $n = 200$ with $N = 40, n_i = 5, \forall i$, we have the number of additions, multiplications and divisions for the exact aggregate filtering as $39999, 40000$ and 40 respectively. The corresponding numbers for approximate aggregate filtering are $3999, 8960$ and 80 respectively. Similar computational savings can be demonstrated for full-order filtering as well.

5 Simulation studies

In this section, we present some simulation studies to demonstrate the performance of our reduced-complexity filter. The performance measures used are average approximation error in aggregate filtering and full-order filtering. The average aggregate filtering error is given by $E[||\zeta_k - \tilde{\zeta}_k||^2]$ which (from ergodicity assumptions) is estimated by the average approximation error $\lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} ||\zeta_k - \tilde{\zeta}_k||^2$. Similarly, the average full-order approximation error is given by $E[||\alpha_k - \tilde{\alpha}_k||^2]$ which in turn is estimated by $\lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} ||\alpha_k - \tilde{\alpha}_k||^2$. For our simulations, the choice for T is $100,000$. We also average our results over 10 simulations.

We perform our simulations with an 8-state ($n = 8, N = 3$) nearly completely decomposable Markov chain as in [5]. The discrete observation (or measurement) set contains 3 possible outputs where the state-to-observation probability matrix is given by

$$C = \begin{bmatrix} 0.2 & 0.15 & 0.5 & 0.3 & 0.2 & 0.8 & 0.75 & 0.25 \\ 0.26 & 0.7 & 0.25 & 0.45 & 0.47 & 0.16 & 0.15 & 0.68 \\ 0.54 & 0.15 & 0.25 & 0.25 & 0.33 & 0.04 & 0.1 & 0.07 \end{bmatrix}$$

Clearly, the choice for W_1, W_2, V_1 and V_2 here is the same as in [5] and is not repeated here.

Remark 3 Note that in this example, C does not have any special structure as in [5]. The only restriction on C is that all elements are positive. Therefore, the results of this paper are applicable to a very general class of state to observation probability matrices.

It was shown in [5] that the various aggregation methods that exist in the literature such as Courtois' method and Khalil's method ([11]) can be extended to obtain *ad hoc* approximate aggregate filters which often result in inexplicably large approximation errors specially in the case of a C matrix which cannot be readily aggregated unlike a block-structured C as in [5]. Therefore, here we do not show any comparisons with other aggregation methods for our approximate aggregate filter. Note also that the full-order approximate filtering scheme is unique to our approach and cannot be obtained by extending any of the aggregation methods.

Table 2 below shows the average approximation error for both the approximate aggregate filter and the full-order approximate filter as obtained by our approach for various values of ϵ ranging from 0.001 to 0.1. The $O(\epsilon)$ nature of the approximation

ϵ	Average approximation error	
	Aggregate filter	Full-order filter
0.001	3.751×10^{-6}	2.7483×10^{-6}
0.005	5.0458×10^{-6}	6.7955×10^{-6}
0.008	9.526×10^{-6}	1.562×10^{-5}
0.01	1.5132×10^{-5}	2.2048×10^{-5}
0.03	2.5031×10^{-5}	1.0583×10^{-4}
0.05	4.9878×10^{-5}	2.5571×10^{-4}
0.08	1.1945×10^{-4}	6.5083×10^{-4}
0.1	1.73252×10^{-4}	1.02574×10^{-3}

Table 2: Table of average approximation error in aggregate and full-order filtering

errors is illustrated in the table above.

Various numerical issues that can arise here in implementing our algorithm regarding stability of $L_k(0)$ or how large can ϵ be etc. will not be specifically addressed in this work as the corresponding discussions are very similar to those in [5]. One note should be added that if, for a particular k, $L_k(0)$ cannot be computed due to the matrix $(\tilde{A}_1^k - L_k(0)\tilde{C}_1^k)$ being singular, one should discard that observation and re-initialize the algorithm. We know due to the exponential forgetting property of HMM filters (see [17]) that such interruptions will not have a significant effect in the asymptotic performance of the algorithm.

References

[1] F. Delebecque and J. P. Quadrat, "Optimal control of Markov chains admitting strong and weak interactions," *Automatica*, vol. 17, pp. 281–296, 1981.

[2] R. G. Phillips and P. V. Kokotovic, "A singular perturbation approach to modelling and control of Markov chains," *IEEE Transactions on Automatic Control*, vol. 26, pp. 1087–1094, 1981.

[3] V. Krishnamurthy, "Adaptive estimation of hidden nearly completely decomposable Markov chains with applications in blind equalization," *International Journal of Adaptive Control and Signal Processing*, vol. 8, pp. 237–260, August 1994.

[4] L. White, R. Mahony, and G. Brushe, "Lumpable hidden Markov models - model reduction and reduced complexity filtering," *IEEE Transactions on Automatic Control*, vol. 45, pp. 2297–2306, December 2000.

[5] S. Dey, "Reduced-complexity filtering for partially observed nearly completely decomposable Markov chains," *IEEE Transactions on Signal Processing*, vol. 48, pp. 3334–3344, December 2000.

[6] H. A. Simon and A. Ando, "Aggregation of variables in dynamic systems," *Econometrica*, vol. 29, pp. 111–138, 1961.

[7] P. J. Courtois, *Decomposability Queueing and Computer System Applications*. New York: Academic Press, 1977.

[8] G. Yin and Q. Zhang, *Continuous-time Markov Chains and Applications: A Singular Perturbation Approach*. New York: Springer-Verlag, 1998.

[9] D. Tse, R. Gallager, and J. Tsitsiklis, "Statistical multiplexing of multiple time-scale Markov streams," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 1028–1038, 1995.

[10] C. D. Meyer, "Stochastic complementation, uncoupled Markov chains, and the theory of nearly reducible systems," *SIAM Review*, vol. 31, pp. 240–272, June 1989.

[11] R. W. Aldhaheri and H. K. Khalil, "Aggregation of the policy iteration method for nearly completely decomposable Markov chains," *IEEE Transaction on Automatic Control*, vol. 36, no. 2, pp. 178–187, 1991.

[12] 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.

[13] F. Delebecque, J. P. Quadrat, and P. V. Kokotovic, "A unified view of aggregation and coherency in networks and Markov chains," *International Journal of Control*, vol. 40, pp. 939–952, 1984.

[14] L. Shue and S. Dey, "Complexity reduction in fixed-lag smoothing of hidden markov models," *IEEE Transactions on Signal Processing*, vol. 50, pp. 1124–1132, May 2002.

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

[16] P. Kokotovic, H. K. Khalil, and J. O'Reilly, *Singular Perturbation Methods in Control: Analysis and Design*. New York: Academic Press, 1986.

[17] L. Shue, B. D. O. Anderson, and S. Dey, "Exponential stability of filters and smoothers for hidden Markov models," *IEEE Transactions on Signal Processing*, vol. 46, pp. 2180–2194, August 1998.