Lumpable Hidden Markov Models—Model Reduction and Reduced Complexity Filtering

Langford B. White, *Member, IEEE*, Robert Mahony, and Gary D. Brushe, *Member, IEEE*

***Abstract*—This paper is concerned with filtering of hidden Markov processes (HMPs) which possess (or approximately possess) the property of lumpability. This property is a generalization of the property of lumpability of a Markov chain which has been previously addressed by others. In essence, the property of lumpability means that there is a partition of the (atomic) states of the Markov chain into aggregated sets which act in a similar manner as far as the state dynamics and observation statistics are concerned. We prove necessary and sufficient conditions on the HMP for exact lumpability to hold. For a particular class of hidden Markov models (HMMs), namely finite output alphabet models, conditions for lumpability of all HMPs representable by a specified HMM are given. The corresponding optimal filter algorithms for the aggregated states are then derived.**

The paper also describes an approach to efficient suboptimal filtering for HMPs which are approximately lumpable. By this we mean that the HMM generating the process may be approximated by a lumpable HMM. This approach involves directly finding a lumped HMM which approximates the original HMM well, in a matrix norm sense. An alternative approach for model reduction based on approximating a given HMM by an exactly lumpable HMM is also derived. This method is based on the alternating convex projections algorithm. Some simulation examples are presented which illustrate the performance of the suboptimal filtering algorithms.

***Index Terms*—Hidden Markov models, model reduction, optimal filtering, state reduction.**

I. Introduction

HIDDEN Markov models (HMMs) have been widely used to describe the nature of many random processes encountered in science and engineering (see [1] for general description and speech processing examples). In particular, HMMs have been applied in the area of estimation of communications signals such as convolutional coded signals [2] and analogue frequency modulated signals [3], to name a couple of examples. An HMM may be thought of as a system model where the observed data are statistically dependent on an unobserved sequence of "state" variables which themselves form a finite state Markov chain. We shall also use the term Hidden Markov Process (HMP) for a realization of the output (or measurement) sequence from an HMM. The difference is important because the same HMP (observed output sequence) may be associated with different HMMs. By making a suitable choice of HMM, the overall complexity of the associated filtering problem can be significantly reduced.

In many practical problems, particularly those involving a number of superimposed statistically independent signals, the computational complexity of the resulting optimal estimation algorithms can become prohibitive, with many being of exponential complexity in terms of some system parameter such as the number of superimposed signals present [4]. The central focus of this paper is to exploit a certain structure known as lumpability, which is often inherent in these models to derive good suboptimal algorithms which are computationally simpler. As stated in [5], exploitation of lumpability is concerned with "... a coarser analysis of possibilities," with the implicit property that such a "coarser" analysis will be generally less computationally demanding to perform. We first focus on filtering for HMPs which are exactly lumpable, then address the problem of approximation of a given HMM by a lumpable one, and the performance of estimation algorithms derived from these approximate models. We are interested in using approximate lumpability as a way of designing computationally efficient suboptimal filters for the states of the original HMM (referred to as atomic states) as well as the aggregated (lumped) states. The paper does not explicitly address smoothing or prediction; however, both smoothing and prediction are closely related to filtering, so in principle, reduced complexity smoothers and predictors for lumpable HMPs could be designed using the techniques of this paper.

The concept of lumpability has been addressed in [5]. One key use of this concept has been in the computation of the asymptotic distribution of subsets of the states of a large Markov chain [6], however only recently has the concept been mentioned in conjunction with HMMs [7]. In [7], lumpability of HMMs was not addressed explicitly, however the procedure mentioned in [7, Section 4] implicitly uses a lumpable approximation to an HMM. We comment further on this approach in Section IV. Implicit in this approach is the concept of a time scale separation in the Markov chain dynamics. The state reduction is achieved in a sense, by only examining the "slow" states. A similar concept is applied in continuous time in [8] and related papers referenced therein. Our formulation is similar in nature to [5], however we have concentrated on a linear subspace setting for the presentation of our results. Although

Manuscript received November 16, 1998; revised July 21, 1999 and February 2, 2000. Recommended by Associate Editor, G. G. Yin. This work was supported by the Australian Government through the Co-operative Research Centres programme and the Department of Defence.

L. B. White is with the Department of Electrical and Electronic Engineering, The University of Adelaide, SA 5005, Australia (e-mail: Lang.White@adelaide.edu.au).

R. Mahony is with the Department of Electrical and Computer Systems Engineering, Monash University, Clayton, Vic., 3168, Australia (e-mail: Robert.Mahony@eng.monash.edu.au).

G. D. Brushe is with Signals Analysis Discipline, Communications Division, Defence Science and Technology Organisation, Salisbury, SA 5108, Australia (e-mail: Gary.Brushe@dsto.defence.gov.au).

Publisher Item Identifier S 0018-9286(00)10668-3.

these results are relatively simple generalizations of the results of [5], the method of proof used, together with the procedure for testing lumpability which we outline, leads us naturally to the approximation algorithms presented subsequently. It also should be mentioned that we generally assume the Markov process is stationary, i.e., it is initialized with its asymptotic distribution. Thus we do not consider the concept of *weak lumpability* discussed in [5] and more recently by a number of authors including [9] for example, which are concerned with selection of appropriate initial conditions yielding a different form of state aggregation. A work which also exploits related subspace properties, but which applies such properties to identify an HMM, and to determine the fundamental complexity of an HMM is found in [10]. Lumpability is also mentioned in conjunction with both HMMs and approximation ideas in [11].

The layout of the paper is as follows: We define lumpability for a Markov chain and prove necessary and sufficient conditions for lumpability of a given chain. A finite algorithm is presented for testing exact lumpability of a Markov chain, and this algorithm yields all lumpings of a given chain. In [5], similar results are given; however, we provide a particularly simple test for lumpability of a given Markov chain, but our results, which are subspace based, offer two potential advantages over the results of [5]. Firstly, we can apply our results to yield lumpable approximations in the case where the Markov chain may be only approximately lumpable. Secondly, because linear (and some nonlinear) subspaces have well defined projection operators, the subspace formalism facilitates the specification of approximation algorithms based on these projections.

We extend the concept of lumpability of Markov chains to HMPs and derive the class of all lumped HMPs for a given HMP. A particular example of a HMP which is generated by a HMM possessing a finite parameterization is the discrete output HMP. In this case, we derive conditions on the *model* parameters (rather than data dependent parameters) which yield lumpability of the model, and thus the lumpability of all HMPs arising from that HMM. Finally we derive the optimal filters for the aggregated states.

One of the main applications of lumpability is model reduction. In Section IV of the paper, we describe an optimal lumped approximation to a given Markov chain, HMP or HMM. This gives rise to a two-pass filtering algorithm for approximately lumpable processes. The first pass computes filtered *a posteriori* probabilities for the aggregated states. The second pass uses maximum *a posteriori* probability estimates for the aggregated states to reduce the computational complexity of the standard optimal filter for the atomic states. The performance of the suboptimal filter is examined using simulation experiments.

In the final section, we examine an *explicit* model reduction technique, exploiting the subspace properties derived in earlier sections.

II. Lumpable Markov Chains

In this section, we define the concept of Markov chain m-lumpability where $2 \leq m < n$ where n is the number of states of the Markov chain. We describe a general procedure for testing m-lumpability, and deriving the associated aggregated states. We then specialize to the case $m = 2$. Although this is the most restrictive case, it is also the easiest to describe, and appropriately models practical applications such as two superimposed Markov chains for example. The case $m = 2$ also allows the important subspace properties we derive here to be more clearly visualized and understood. The necessity for much additional and cumbersome notation may then be avoided. We comment that the results of our main theorems generally apply for $m > 2$, but space limitations prevent us providing details herein. Readers are referred to [12] and [13] for examples of m-lumpable HMMs where $m > 2$.

A. Testing m-Lumpability

Definition: Let X_t be a n-state Markov chain taking values in the state set S. Then X_t is said to be m-lumpable if and only if there exist nonempty disjoint subsets $S_i, i = 1, \ldots, m$ such that $S = \bigcup_{i=1}^{m} S_i$ and for every $i = 1, \ldots, m$, $\Pr\{X_t \in S_i \,|\, X_{t-1} = q\}$ is independent of $q, \forall\, q \in S_j, \forall\, j \neq i$.

Definition: Let $n > 2$, then a class $\{I_k\}$ of $m < n$ disjoint nonempty subsets of $\{1, \ldots, n\}$ with union $\{1, \ldots, n\}$ is said to be an m-partition of $\{1, \ldots, n\}$.

Definition: Let $n > 2$, and $1 < m < n$. We denote by $\mathcal{Q}_n^m$ the set of all $L \in \{0,1\}^{m \times n}$ with elements

$$L_{i,j} = \begin{cases} 1 & j \in I_i \\ 0 & \text{else} \end{cases} \tag{1}$$

where $\{I_k\}$ is a m-partition of $\{1, \ldots, n\}$. The matrix L is called the lumping matrix defined by the m-partition $\{I_k\}$.

We now describe a finite algorithm for computing all lumpings of a specified Markov chain. Suppose $\{I_k\}$ is a m-lumping of an n-state process X_t, where $2 < m < n$, noting that an n-lumping is trivial. We are interested in testing whether an $(m-1)$-lumping exists. We suppose this is achieved (without loss of generality, by relabeling if necessary) by amalgamating I_{m-1} and I_m. It is easy to see from the definition of lumpability, that one may test whether this amalgamation may be made by selecting any $q \in I_{m-1}$ and $q' \in I_m$ and determining whether for each $i \in \{1, \ldots, m-2\}$ the condition

$$\Pr\{X_t \in S_i \,|\, X_{t-1} = q\} = \Pr\{X_t \in S_i \,|\, X_{t-1} = q'\} \tag{2}$$

holds. Thus it suffices to only test using arbitrary q and q' because the m-lumping guarantees that each of the above transition probabilities are constant across I_{m-1} and I_m, respectively. We then have that the $(m-1)$-partition

$$J_i = I_i, \quad i = 1, \ldots, m-2, \quad J_{m-1} = I_{m-1} \cup I_m \tag{3}$$

generates an $(m-1)$-lumping for X_t.

Comments:

1) The above result shows that it takes $m-2$ comparisons to check whether a given pair of subsets of the m-partition can be amalgamated. Thus, to determine all $(m-1)$-lumpings given an m-lumping requires at most $m(m-1)(m-2)/2$ comparisons.

2) Thus we may specify a procedure for obtaining all lumpings which exist.
   i) Let $m = n$ and define the trivial partition $I_k = \{k\}$ for $k = 1, \ldots, n$.
   ii) Let $\{I_k\}$ denote an m-partition generating a lumping. Apply the above test to each distinct pair of $\{I_k\}$. If there are no successes, then we stop the procedure, with the process being m-lumpable but not $(m-1)$-lumpable, and the set of all such m-partitions defines all m-lumpings. If there are successes, then amalgamate the appropriate partition subsets, and define the $(m-1)$-partition $\{J_k\}$ generated by the m-partition $\{I_k\}$ as above. There may be several such $(m-1)$-partitions for a given m-partition. Repeat this process for each m-partition.
   iii) This yields the minimal m yielding m-lumpability, together with all m-lumpings and associated p-lumpings for $n - 1 \geq p > m$.

B. The Case When $m = 2$

We now specialize to the case $m = 2$ and derive linear algebraic descriptions of lumpability. Generalizations to arbitrary m of most results presented in this paper for $m = 2$ can be found in [13].

Lemma 1: A n-state Markov chain X_t is 2-lumpable if and only if the state transition matrix is (up to re-labeling of the states) of the form

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

where $A_{11} \in \mathbb{R}^{k \times k}$, $A_{22} \in \mathbb{R}^{(n-k)\times(n-k)}$, and there are μ_1 and μ_2 such that

$$\begin{aligned} A_{11}\mathbf{1} &= \mu_1 \mathbf{1}, \quad A_{12}\mathbf{1} = (1-\mu_1)\mathbf{1} \\ A_{21}\mathbf{1} &= (1-\mu_2)\mathbf{1}, \quad A_{22}\mathbf{1} = \mu_2 \mathbf{1} \end{aligned} \tag{4}$$

where $\mathbf{1} = [1, \ldots, 1]^T$ is a vector of unity elements of the appropriate dimension.

Proof: Take $S_1 = \{q_1, \ldots, q_k\}$ and $S_2 = \{q_{k+1}, \ldots, q_n\}$ where $\{q_1, \ldots, q_n\}$ is the state set for X_t. Now for $i = k+1, \ldots, n$ we have

$$\Pr\{X_t \in S_1 \,|\, X_{t-1} = q_i\} = \sum_{j \in I_1} A_{i,j} = [A_{21}\mathbf{1}]_{i-k} \tag{5}$$

which is required to be independent of i, i.e., a scalar multiple of $\mathbf{1}$. This result holds with S_1 replaced by S_2, and A_{21} by A_{12}. The other equations follow from the fact that $A\mathbf{1} = \mathbf{1}$. Conversely if (4) holds, then the resulting class probabilities satisfy the lumping condition. □

Theorem 1: An n-state Markov chain X_t is 2-lumpable under lumping matrix L if and only if its state transition matrix A satisfies $A^T\mathcal{N}(L) \subseteq \mathcal{N}(L)$, i.e., the null space $\mathcal{N}(L)$, of L is A^T-invariant.

Proof: Suppose X_t is 2-lumpable under lumping matrix L which we can assume w.l.o.g. (by state relabeling) to be of the form $L = \operatorname{diag}(\mathbf{1}^T, \mathbf{1}^T)$, where there are k unit elements in the first row and $n-k$ on the second. The null space of L is given by the set of all n-vectors x of the form $x^T = [x_1^T \quad x_2^T]$ where $x_1 \in \mathbb{R}^k, x_2 \in \mathbb{R}^{n-k}$, with $\mathbf{1}^T x_1 = 0 = \mathbf{1}^T x_2$. Then by Lemma 1,

$$LA^T \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \mu_1 \mathbf{1}^T x_1 + (1-\mu_2)\mathbf{1}^T x_2 \\ (1-\mu_1)\mathbf{1}^T x_1 + \mu_2 \mathbf{1}^T x_2 \end{bmatrix} \tag{6}$$

which is clearly zero for all $x \in \mathcal{N}(L)$.

Conversely, suppose $LA^T x = 0$ for all $x \in \mathcal{N}(L)$, then in particular firstly taking $x^T = [x_1^T \quad 0_{n-k}^T]$ then taking $x^T = [0_k^T \quad x_2^T]$ we have from (5) that $\mathbf{1}^T A_{i,j}^T x_1 = 0$, $i, j = 1, 2$, $\forall x_1$ and x_2 orthogonal to (the appropriate dimension) $\mathbf{1}$. The dimension of the subspaces of such x_1 and x_2 are $k-1$ and $n-k-1$, respectively. Thus $A_{11}\mathbf{1} \perp x_1$, $\forall x_1 \in \mathbf{1}^\perp$. Thus $A_{11}\mathbf{1} \in \operatorname{Span}\{\mathbf{1}\}$, i.e., there is a constant μ_1 such that $A_{11}\mathbf{1} = \mu_1 \mathbf{1}$. The remainder of (4) can be similarly derived with the constant terms constrained by the unity row sums of A. Thus A corresponds to a 2-lumpable chain under L. □

Corollary: Consider $V^T A V$, where $V = [v_1, \ldots, v_n]$ are the right singular vectors of L, with $\mathcal{N}(L) = \operatorname{Span}\{v_3, \ldots, v_n\}$, then A is 2-lumpable under L if and only if A has the following form in the coordinate change generated by V:

$$V^T A V = \begin{bmatrix} E_{11} & E_{12} \\ 0_{n-2,2} & E_{22} \end{bmatrix} \tag{7}$$

with E_{11} nonnegative, i.e., there is a $(n-2) \times 2$ zero matrix in the bottom left block. Note also that $[v_1 \quad v_2] = L^T D^{1/2}$, i.e., v_1 and v_2 are the scaled (to unit norm) rows of L.

Proof: Follows from the fact that $v_j^T A^T v_i = 0$ for $i = 3, \ldots, n$ and $j = 1, 2$. Nonnegativity of E_{11} follows from the fact that $L^T = [v_1 \quad v_2] D^{-1/2}$ where $D = (LL^T)^{-1} = \operatorname{diag}(k, n-k)^{-1}$ is nonnegative. D represents a weighting dependent on the number of atomic states in each aggregated state. □

Comment: The number of candidate 2-lumpings is given by the number of distinct pairs of subsets of $\{1, \ldots, n\}$ at least one of which has at least 2 elements. This number is $2^{n-1} - 1$ for $n \geq 3$. Despite the fact that the number of candidate lumpings increases exponentially as the state dimension, a large number of candidate lumpings can be rejected by simple tests based directly on the definition as above. It is only when the number of atomic states which are being lumped into the two aggregated state sets are each comparable, is there advantages in using the results of the corollary above to test candidate lumpings.

The next result yields the lumped Markov chain model, and is analogous to the example [5, p. 125].

Lemma 2: Let $\overline{A}$ denote the transition matrix for the lumped model of X_t, then $\overline{A} = DLAL^T$ where A is the transition matrix for X_t, and D is a diagonal weighting matrix with element $D_{i,i}^{-1}$ being the number of atomic states in aggregated state set i.

Proof: The aggregated class transition probabilities are

$$\begin{aligned} \overline{A}_{i,j} &= \Pr\{X_t \in S_j \,|\, X_{t-1} \in S_i\} \\ &= \sum_{k \in S_j} \Pr\{X_t = q_k \,|\, X_{t-1} \in S_i\} \\ &= \sum_{k \in S_j} A_{p,k} \qquad \forall p \in I_i. \end{aligned} \tag{8}$$

Thus

$$\overline{A}_{i,j} = \sum_{k \in S_j} A_{p,k} = \frac{1}{\#S_i} \sum_{k \in S_j} \sum_{p \in S_i} A_{p,k} \left[DLAL^T \right]_{i,j} \quad (9)$$

where $\#S$ denotes the cardinality of the set S. This establishes the result. Note that $\overline{A} = D^{1/2} E_{11} D^{-1/2}$, where E_{11} is as defined in the corollary to Theorem 1. □

Comment: It is easy to show that the only single nonunity eigenvalue of $\overline{A}^T \in \mathbb{R}^{2\times 2}$ is identical to the eigenvalue of A^T corresponding to those eigenvectors of A^T not in $\mathcal{N}(L)$, i.e., $\mathcal{N}(A^T - \lambda \mathbb{I}) \cap \mathcal{N}(L)^{\perp} \neq \{0\} \Rightarrow \mathcal{N}(\overline{A}^T - \lambda \mathbb{I}) \neq \{0\}$. To see this, suppose $x \in \mathcal{N}(A^T - \lambda \mathbb{I})$, $x \neq 0$, where $\mathbb{I}$ denotes the appropriate size identity matrix. Decompose x into the direct sum $x = x_1 \oplus x_2$ where $x_1 \in \mathcal{N}(L)$ and $x_2 \in \mathcal{N}(L)^{\perp} = \mathcal{R}(L^T)$ with $x_2 \neq 0$. Here $\mathcal{R}$ denotes the range of a matrix. Then lumpability implies $A^T x = A^T x_1 + A^T L^T y_2$, for some $y_2 \in \mathbb{R}^2$, $y_2 \neq 0$. This holds because $A^T x_1 \in A^T \mathcal{N}(L) \Rightarrow A^T x_1 \in \mathcal{N}(L)$, and $x_2 \in \mathcal{R}(L^T)$. So $LA^T x = LA^T L^T y_2 = \overline{A}^T D^{-1} y_2$. But $LA^T x = \lambda L x = \lambda L x_2 = \lambda L L^T y_2 = \lambda D^{-1} y_2$. Now $x_2 \neq 0 \Rightarrow y_2 \neq 0 \Rightarrow D^{-1} y_2 \neq 0 \Rightarrow \mathcal{N}(\overline{A}^T - \lambda \mathbb{I}) \neq \{0\}$.

We will designate the lumped chain by χ_t.

III. Lumpable Hidden Markov Processes and Models

In this section we shall extend the concept of a lumpable Markov chain to a hidden Markov process (HMP). A general definition in terms of joint state transition and observation probabilities is given first. We shall prove a general theorem analogous to Theorem 1 for HMPs with continuous outputs. We then specialize the result to the case where the HMP has a finite number of discrete levels. For this particular subclass of HMMs, we find it convenient to distinguish between lumpability for a given observed HMP, and lumpability of the model which generated the observed HMP. This is because the concept of lumpability for the HMM can then be made explicit by virtue of the finite parameterization inherent in this specific case *vis a vis* the general continuous output case. An HMM defines the statistics of all HMPs it generates. We also comment that although our results are given for $m = 2$, the comments made in the previous section relating to m-lumpability of Markov chains, also apply here with the appropriate modifications [13].

Definition: An HMM is defined by a finite state level set $Q = \{q_1, \ldots, q_n\}$, a transition probability matrix $A \in \mathbb{R}^{N\times N}$ satisfying $A_{i,j} \geq 0$, for all $i, j \in \{1, \ldots, N\}$ and $A\mathbf{1} = \mathbf{1}$, and a set of probability measures $p_i(.), i = 1, \ldots, N$. We denote the HMM by the triple $(A, Q, p(.))$, where $p(.)$ denotes the vector comprising the $p_i(.)$. We will typically assume that $p(\cdot)$ is of specified form, perhaps depending on a finite parameterization θ.

Definition: A hidden Markov process (HMP) generated by the HMM $(A, Q, p(.))$ is a random process $Y_t, t \geq 0$, which satisfies: i) Y_t are conditionally independent given an underlying state process X_t taking values in the set Q with conditional measure $p(Y_t \,|\, X_t = q_i) = p_i(Y_t)$ and ii) the process X_t is a Markov chain with transition probabilities $\Pr\{X_{t+1} = q_j \,|\, X_t = q_i\} = A_{i,j}$, where $Q = \{q_1, \ldots, q_N\}$.

Comment: We have used the notation $\Pr\{.\}$ in a loose manner. To be precise, one should work with mixed continuous and discrete measures to handle continuous observations and discrete states but the context remains clear.

Definition: An HMP with state sequence $X_t \in S$ and observation sequence Y_t where $t = 0, 1, \ldots$, is said to be 2-lumpable if and only if there exist nonempty disjoint subsets S_1 and S_2 of S such that $S = S_1 \cup S_2$ and: i) $\Pr\{X_t \in S_1, Y_t \,|\, X_{t-1} = q\}$ is independent of q, $\forall q \in S_2$, $\forall t \geq 0$ and ii) $\Pr\{X_t \in S_2, Y_t \,|\, X_{t-1} = q\}$ is independent of q, $\forall q \in S_1, \forall t \geq 0$.

Comment: Notice that in contrast with the definition of lumpability of a (stationary) Markov chain, the above property must hold at all times. This is because the probabilities involved are dependent on the observed HMP realization, which is clearly time dependent. This is in contrast to the HMM generating the observed HMP, which specifies the (time-independent) statistics of the HMP. For the specific HMMs we address below, this requirement will be replaced by requirements made of the (finite number of) model parameters.

Definition: A hidden Markov model is said to be 2-lumpable if and only if every HMP generated by the model is 2-lumpable.

A. Continuous Output HMMs

A continuous output HMM will be parameterized by the state transition matrix A and observation conditional probability density $p(.)$. We define the observation matrices $B_t = \mathrm{diag}(p(Y_t \,|\, X_t = q_i))$.[1] An HMP is parameterized by $(A, \{B_t\})$.

Theorem 2: The HMP $(A, \{B_t\})$ is 2-lumpable if and only if for each $t \geq 0$ there is a partition

$$AB_t = \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix} \quad (10)$$

where there are $\nu_1, \ldots, \nu_4 \geq 0$ such that

$$\begin{aligned} E_{i,j}\mathbf{1} &= \nu_j \mathbf{1}, \qquad i = 1,\ j = 1, 2 \\ E_{i,j}\mathbf{1} &= \nu_{i+j}\mathbf{1}, \qquad i = 2,\ j = 1, 2. \end{aligned} \quad (11)$$

Here the E_{ij} and ν_i are in general time-dependent due to the dependence on the measurements sequence.

Proof: Follows as per Lemma 1 but using the fact that $\Pr\{X_t = q_i, Y_t \,|\, X_{t-1} = q_j\} = [B_t A^T]_{i,j}$. The fact that we do not have $\nu_2 = 1 - \nu_1$ and $\nu_3 = 1 - \nu_4$ in (11) in general, follows from the fact that AB_t is not row stochastic in general. However, regarding the ν_i terms as dependent on the observation $Y_t = y$, it is easy to see that $\int \nu_1(y) + \nu_2(y) dy = 1$, and similarly for $\nu_3 + \nu_4$. Nonnegativity of the ν_i follows from the fact that both A and B_t are nonnegative. □

Comment: The unnormalized filter recursion for computing $\alpha_t = \Pr(X_t, \mathbf{Y_t^-})$ where $\mathbf{Y_t^-} = \{Y_0, \ldots, Y_t\}$, is [14]

$$\alpha_{t+1} = B_{t+1} A^T \alpha_t. \quad (12)$$

It follows from Lemma 2 that the optimal (unnormalized) filter for the aggregated states $\mu_t = \Pr\{\chi_t, \mathbf{Y_t^-}\}$, is given by

$$\mu_{t+1} = LB_{t+1} A^T L^T D \mu_t = Q_{t+1}^T \mu_t. \quad (13)$$

[1] By this we mean $[B_t]_{i,j} = \delta_{i,j} p(y_t \,|\, X_t = q_i)$, where $\delta_{i,j} = 1$ if $i = j$ and zero otherwise.

Corollary: Let V denote the right singular vectors of L, as before. Then AB_t is 2-lumpable if and only if AB_t has the representation

$$V^T AB_t V = \begin{bmatrix} P_t & \eta_t \\ 0_{n-2,2} & \zeta_t \end{bmatrix}$$

where P_t is positive.

Proof: Follows directly from the application of the corollary to Theorem 1 to the product AB_t. The filter matrix Q_t as defined in (13) is given by $Q_t = D^{1/2} P_t D^{-1/2}$. □

B. Discrete Output HMMs

Consider a finite output HMM with observations Y_t taking one of M possible values $r_1, \ldots, r_M$ and conditional output probability matrix $C \in R^{M \times N}$ with elements

$$C_{i,j} = \Pr\{Y_t = r_i \,|\, X_t = q_j\} \tag{14}$$

so that $C^T \mathbf{1} = \mathbf{1}$. Let $B_t = \operatorname{diag}(C_{y_t,.})$,[2] which will be a diagonal matrix with one of the M rows of C on its diagonal specified by the observation, i.e., there are only M possible values which can be assumed by B_t. A discrete output HMM is parameterized by the matrix pair (A, C).

Theorem 3: The HMM (A, C) is 2-lumpable if and only if for each $i = 1, \ldots, M$, there is a partition

$$A \operatorname{diag}(c_i) = \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix}$$

where $C^T = [c_1, \ldots, c_M]$ there are $\nu_1, \ldots, \nu_4 \geq 0$ such that (11) holds. The matrices E_{jk} and scalars ν_j are dependent on i due to the explicit dependence on the measurements. In addition we have that $\sum_{i=1}^{M} \nu_1(i) + \nu_2(i) = 1$, and similarly for $\nu_3 + \nu_4$.

Proof: Follows directly from Theorem 2 since the required property for AB_t holds for all $t \geq 0$ and for all observation sequences $\{y_t\}$ if and only if $A \operatorname{diag}(c_i)$ has the form above $\forall i = 1, \ldots, M$. The final result follows by summing each side of (11) over the implicit i dependence. □

Corollary: $\mu_t = L\alpha_t, \forall t$ if and only if

$$V^T A \operatorname{diag}(c_i) V = \begin{bmatrix} P_i & R_i \\ 0_{n-2,2} & Z_i \end{bmatrix} \qquad \forall i \in \{1, \ldots, M\}$$

where the P_i are nonnegative.

Proof: Again this result follows directly from the corollary to Theorem 1, again with the same argument as Theorem 3. □

Comment: As in (13), the recursion for the filtered *a posteriori* probabilities for the aggregated states involves the matrices P_i one of which is determined by each measurement Y_t. In contrast to the continuous output case where an infinite number of possible values for P_i could exist, there are only M distinct values the measurements can take. Thus there are only M distinct matrices P_i which can result. These could therefore be precomputed and the measurements used to select the appropriate Q_t at each t [12].

[2]If $x \in \mathbb{R}^n$, then $\operatorname{diag}(x) \in \mathbb{R}^{n \times n}$ with $[\operatorname{diag}(x)]_{i,j} = \delta_{i,j} x_i$.

IV. Suboptimal Filtering for HMMs via Model Reduction

In this section, we shall be concerned with using the lumpable approximation to a HMM to derive efficient filters for the lumped or aggregated state sets as well as the original atomic states of the n-state HMM. Attention will be restricted to discrete output HMMs.

A. Filtering for Aggregated States

In this section, we examine optimal filtering for the aggregated states of a lumpable approximation to a given HMM. Since we are only interested in determining *a posteriori* probabilities for the aggregated states, we do not require the $n \times n$ *lumpable* approximation to the HMM, only the 2×2 *lumped* matrix. This leads naturally to an approximation procedure which yields directly the optimal 2×2 lumped matrix. An alternative procedure which yields the $n \times n$ lumpable transition matrix is given in Section VI.

Let $\mathcal{S}_n^+$ denote the set of all n-state transition probability matrices, i.e., $\mathcal{S}_n^+ = \{A \in \mathbb{R}^{n \times n} \colon A_{i,j} \geq 0, A\mathbf{1} = \mathbf{1}\}$. Denote by $\mathcal{S}_n$ the affine subspace of $\mathbb{R}^{n \times n}$ given by $\mathcal{S}_n = \{\Psi \in \mathbb{R}^{n \times n} \colon \Psi\mathbf{1} = \mathbf{1}\}$. This $n(n-1)$-dimensional subspace reflects one part of the probability constraint imposed by membership of $\mathcal{S}_n^+$; the other being the nonnegativity of elements of A. Denote by $\mathcal{T}_n$ the set of all $\Psi \in \mathbb{R}^{n \times n}$ satisfying $L\Psi^T x = 0, \forall x \in \mathcal{N}(L)$ then we show that $\mathcal{T}_n$ is an $n^2 - 2(n-2)$-dimensional linear subspace. Note that $\mathcal{T}_n$ depends on L. Let $\mathcal{P} = \{\Psi \in \mathbb{R}^{n \times n} \colon \Psi_{i,j} \geq 0 \; \forall i, j = 1, \ldots, n\}$. Evidently $\mathcal{S}_n^+ = \mathcal{S}_n \cap \mathcal{P}$.

Lemma 3: Let $L \in \mathcal{Q}_n, \Psi \in \mathcal{P}$, and let[3]

$$J_2(Q; \Psi) = \left\| L^T Q - \Psi L^T \right\|^2 \tag{15}$$

then

$$Q^* = \operatorname*{argmin}_Q J(Q; \Psi) = DL\Psi L^T \tag{16}$$

where $D = \operatorname{diag}(k, n-k)^{-1}$, and $Q^* \in \mathcal{P}$. Furthermore $\Psi \in \mathcal{S}_n^+ \Rightarrow Q^* \in \mathcal{S}_2^+$. Also $J_2(Q^*; \Psi) = 0$ if and only if Ψ is lumpable by L.

Proof: We write

$$\begin{aligned} J_2(Q; \Psi) &= \operatorname{Tr}\Big(\big(L^T Q - \Psi L^T\big) \big(L^T Q - \Psi L^T\big)^T \Big) \\ &= \operatorname{Tr}\big(Q^T L L^T Q - 2L\Psi L^T Q + \Psi L^T L \Psi^T \big). \end{aligned} \tag{17}$$

Differentiating wrt Q and using $LL^T = D^{-1}$ yields

$$0 = LL^T Q - L\Psi L^T \Rightarrow Q^* = DL\Psi L^T. \tag{18}$$

Clearly Q^* is nonnegative as it is the product of nonnegative matrices. To verify that $Q^* \in \mathcal{S}_2^+$, when $\Psi \in \mathcal{S}_n^+$ consider

$$Q^*\mathbf{1} = DL\Psi L^T \mathbf{1} = DL\Psi\mathbf{1} = DL\mathbf{1} = D[k \;\; n-k]^T = \mathbf{1}. \tag{19}$$

The fact that $J_2(Q^*; \Psi) = 0$ if and only if Ψ is lumpable by L follows from Lemma 1. □

Comments:

1) The above optimization problem arises if we were to assume that the lumped probabilities were correct at time t,

[3]Throughout the paper we use the Frobenius norm $\|\Psi\| = (\operatorname{Tr}(\Psi\Psi^T))^{1/2}$.

i.e., $\mu_t = L\alpha_t$ and we seek to compute $\mu_{t+1} = Q_{t+1}^T \mu_t$ which is as close to $L\alpha_{t+1}$ as possible. Hence minimization of the above norm is appropriate in terms of minimizing accumulated error in the approximate *a posteriori* probabilities for the lumped state sets. Correct probabilities for the aggregated states will thus be obtained if and only the relevant Ψ is lumpable.

2) Lemma 3 can be applied directly to $\Psi = A\,\mathrm{diag}(c_i), i = 1, \ldots, M$ in the case of discrete output HMMs to yield $Q_i, i = 1, \ldots, M$ for use in the aggregated state filter.
3) Suppose the underlying Markov chain X_t is itself lumpable with lumped matrix $\overline{A}$, then for the discrete output HMM the question arises whether to use $\overline{A}DL\,\mathrm{diag}(c_i)L^T$, or $DLA\,\mathrm{diag}(c_i)L^T$ as Q_i arises. By virtue of the optimality of the above approximation procedure, we conjecture that lumping $\Psi = A\,\mathrm{diag}(c_i)$ is the best idea.
4) In [7], the authors introduced a lumped approximation for a continuous output HMM by assuming a lumpable Markov chain, and approximating the output probability matrices B_t by the actual conditional observation probabilities given the aggregated states. The formula for the aggregated output matrix is then $\overline{B}_t^p = \mathrm{diag}(L\pi)^{-1}\,\mathrm{diag}(LB_t\pi)$, where π is the asymptotic distribution of the atomic chain X_t. This compares to the direct lumping on $\Psi = B_t$ using the formula from Lemma 3 (which inherently and implicitly assumes a uniform asymptotic distribution of the atomic states) $\overline{B}_t = \mathrm{diag}(L\mathbf{1})^{-1}\,\mathrm{diag}(LB_t\mathbf{1})$, i.e., in [7], L is replaced by $L\Pi$ where $\Pi = \mathrm{diag}(\pi)$. Given the optimality provided via Lemma 3, this approach would be expected to be inferior to a direct lumping of $\Psi = AB_t$. A heuristic argument is that using $Q_t = \overline{AB_t}$ does not permit any coupling between the Markov chain dynamics and the observation likelihoods when determining a lumpable approximation to AB_t. Use of $\overline{B}_t^p$ permits coupling via the stationary distribution of X_t, while lumping AB_t together uses full information of the dynamics of X_t. Thus it is suggested that lumping of B_t alone should not be used. The condition of a uniform asymptotic probability distribution for the underlying chain is not a restriction as we suggest lumping the product $A\,B_t$ (or the appropriate HMM parameters) directly.
5) In [12], the authors concentrate in particular on superimposed independent Markov processes in noise. In this case, the lumping proposed here reduces to a product of an averaged data likelihood and the dynamics of the component chain. Formulas for the lumping matrices are explicitly given. Also, a successive estimation scheme for multiuser estimation (i.e., estimation of all component Markov chains) is provided.

B. A Two-Pass Filter for the Atomic States

In this section, we propose a filter for the atomic states of a HMP. The filter is based on finding a good lumpable approximation and using the results of a coarse state estimation to yield a filter of reduced complexity for the atomic states. This filter uses two passes through the observed data. On the first pass we construct the optimal lumped approximation Q_t as described in Lemma 1 above. This yields approximate filtered probabilities μ_t for the aggregated states χ_t from which we can derive approximate maximum *a posteriori* probability (MAP) estimates $\hat{\chi}_t = \mathrm{argmax}\{\mu_t(i)\colon i = 1, 2\}$ for the aggregated state sets. Our algorithm for filtering for the atomic states is based on the assumption that $\hat{\chi}_t = \chi_t$ with probability 1 for all t. Thus pass 2 proceeds using the iteration

$$\hat{\alpha}_{t+1} = B_{t+1}\hat{A}_t^T\hat{\alpha}_t \tag{20}$$

where instead of using the transition matrix A for the atomic HMP, we use instead the quantity

$$\left[\hat{A}_t\right]_{i,j} = \begin{cases} A_{i,j}, & i \in \hat{\chi}_t, j \in \hat{\chi}_{t+1} \\ 0, & \text{else.} \end{cases} \tag{21}$$

One can thus view the MAP estimates from the first pass as *controls* for the second pass. The computational complexity for the first pass is $2m^2$, while the computational complexity for the second pass is $2n^2/m$. Fig. 1 shows the best savings achievable (i.e., proportion reduction over optimal atomic state filter) in computational requirements as a function of the normalized number of aggregated states m/n. This optimum occurs when each of the aggregated states has the same number of elements. This figure clearly displays that there is an optimal aggregation number (roughly $m = (n^2/2)^{1/3}$), and that potential benefits increase as n increases.

It should be noted that both filters can be run together since (21) requires only knowledge of the current and previous time estimates from pass 1. We will now examine the performance of this filter with an example.

V. Simulations

A. Example 1—Exact Lumpability

In this section, we give an example of an exactly lumpable HMM with 16 states. There are 4 discrete outputs, and we seek an 8-lumping, i.e., 8 aggregated state sets each having 2 atomic states as members. The parameters of this discrete output HMM are determined as follows. Let

$$J = 0.5\,\mathrm{diag}\left(\mathbf{1}^T, \mathbf{1}^T\right) \tag{22}$$

where each of the vectors $\mathbf{1}$ have 2 elements, and define

$$A = \begin{pmatrix} A_1 \\ A_1 \end{pmatrix}$$

where $A_1 = \mathrm{diag}(J, J, J, J)$. The observation probability matrix is specified by a matrix consisting of a single unity element in each row with zeros elsewhere. To model a stochastic (or noisy) output measurement, we perturb each element independently by a uniform random variable distributed on $(0, \epsilon)$ where $\epsilon > 0$ is termed a noise parameter, and rescale to ensure the matrix remains a (conditional) probability matrix. The form of C for noise parameter $\epsilon = 0.1$ is given by (23), shown at the bottom of the next page. It can be shown that (A, C) is exactly 2-lumpable under $L_8 = [\mathbb{I}_8 \;\; \mathbb{I}_8]$ where $\mathbb{I}_n$ denotes the identity

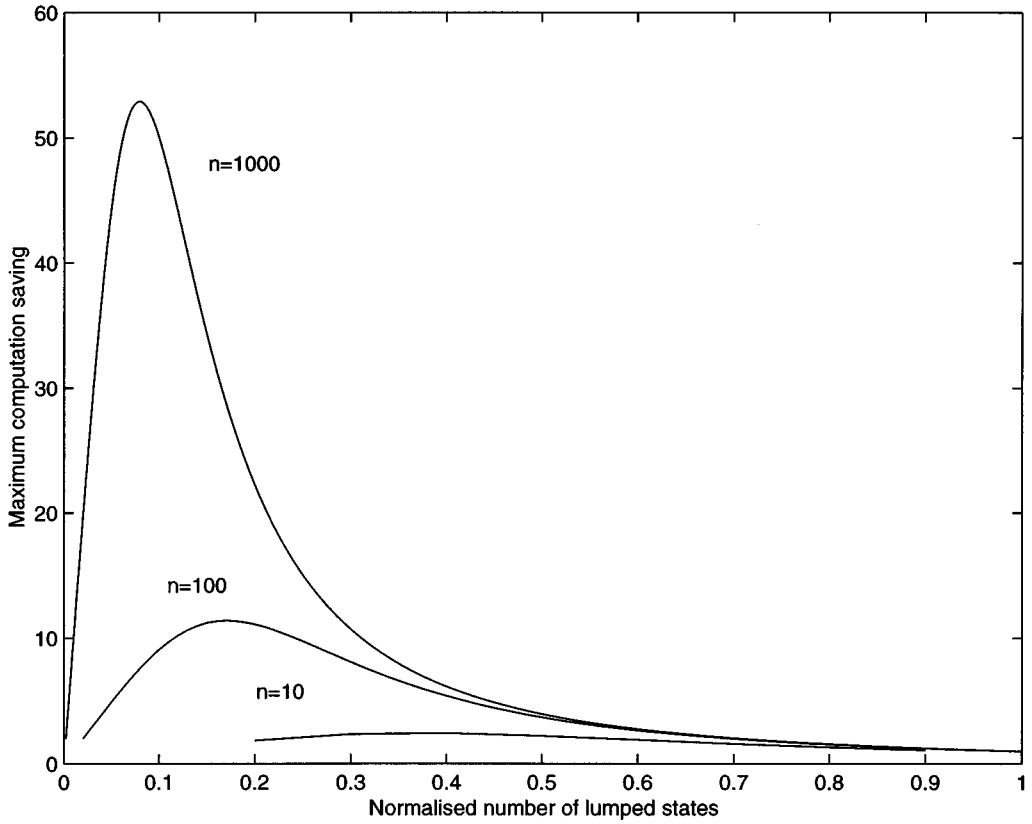

Fig. 1. Computational savings versus degree of aggregation.

matrix of size $n \times n$. The lumped model Q_1 corresponding to the first row of C in (23) is given by (24), shown at the bottom of the page, and similarly for the three remaining lumped models.

The optimal filter for MAP estimates of the atomic states was computed and applied to 1000 realizations of HMPs of length 200 samples produced from the above HMM. The two pass procedure was also applied to the same data, with state estimation error results tabulated in Table I. For pass 1 we say there is an error if the true atomic state does not lie in the estimated aggregated state set.

$$C^T = \begin{bmatrix} 0.8393 & 0.0416 & 0.0147 & 0.0315 & 0.0550 & 0.0666 & 0.0148 & 0.8397 \\ 0.0094 & 0.0311 & 0.9283 & 0.0730 & 0.0261 & 0.8095 & 0.0494 & 0.0688 \\ 0.0802 & 0.0681 & 0.0138 & 0.8614 & 0.8501 & 0.0668 & 0.0027 & 0.0021 \\ 0.0711 & 0.8593 & 0.0432 & 0.0341 & 0.0688 & 0.0571 & 0.9331 & 0.0694 \\ 0.0477 & 0.8622 & 0.0725 & 0.0581 & 0.0765 & 0.0355 & 0.8257 & 0.0724 \\ 0.0316 & 0.0331 & 0.0278 & 0.8787 & 0.9017 & 0.0082 & 0.0671 & 0.0710 \\ 0.0659 & 0.0457 & 0.8385 & 0.0059 & 0.0078 & 0.9084 & 0.0587 & 0.0393 \\ 0.8548 & 0.0590 & 0.0612 & 0.0574 & 0.0140 & 0.0479 & 0.0485 & 0.8173 \end{bmatrix} \tag{23}$$

$$Q_1 = \begin{bmatrix} 0.4601 & 0.028 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.0112 & 0 & 0 & 0.0153 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.4188 & 0 & 0 & 0.0343 \\ 0 & 0 & 0.0253 & 0 & 0 & 0.0311 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.0237 & 0 & 0 & 0.4124 \\ 0 & 0 & 0 & 0.0005 & 0 & 0 & 0.0335 & 0 \\ 0.0119 & 0.4147 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.0250 & 0 & 0 & 0.0049 & 0 \end{bmatrix} \tag{24}$$

TABLE I
STATE ESTIMATION ERROR RESULTS FOR TWO-PASS FILTER

Noise	Optimal		Pass 1		Pass 2	
Level	mean	std. dev.	mean	std. dev.	mean	std. dev.
0.01	0.0232	0.0181	0.0232	0.0181	0.0380	0.0240
0.03	0.0695	0.0322	0.0695	0.0322	0.1057	0.0405
0.05	0.1173	0.0428	0.1172	0.0427	0.1650	0.0526
0.07	0.1643	0.0507	0.1639	0.0506	0.2237	0.0599
0.09	0.2099	0.0561	0.2092	0.0559	0.2790	0.0642

The computation required is reduced by approximately 72% in this case with an increase in state estimation error ranging from about 65% at low noise to 33% at higher noise. It is interesting to note that the estimated error probabilities computed for the estimation of the aggregated states in pass 1, are approximately the same as the estimated probability of error in estimating the atomic states with the optimal filter. There appears to be a small decrease for higher noise, although the statistical variation from this experiment is too high to make any definitive comment. This effect may be more significant for aggregated states having a larger number of elements, rather than the case for this example, where each aggregated state contains only two atomic states.

B. Example 2—Approximate Lumpability

The second example addressed is one where the HMP is not exactly lumpable. We will consider an example of two superimposed Markov chains. We consider two statistically independent binary chains $X_t^{(1)}$ and $X_t^{(2)}$ with transition probability matrices given, respectively, by

$$A_1 = \begin{bmatrix} 0.7 & 0.3 \\ 0.05 & 0.95 \end{bmatrix}, \qquad A_2 = \begin{bmatrix} 0.2 & 0.8 \\ 0.4 & 0.6 \end{bmatrix}. \tag{25}$$

The transition probability matrix for the Cartesian product $X_t^{(1)} \times X_t^{(2)}$ is thus $A = A_1 \otimes A_2$ [4]. We will consider a continuous observation conditionally Gaussian HMP with output mapping

$$y_t = X_t^{(1)} + X_t^{(2)} + n_t \tag{26}$$

where n_t is an iid Gaussian zero mean process with variance σ^2. The Markov chains $X_t^{(1)}$ and $X_t^{(2)}$ assumes values in the level sets $q_1 = \{-1, 1\}$ and $q_2 = \{0.5, 1\}$, respectively.

The following filters were computed and applied to 1000 realizations of the process, each of 200 samples length: 1) the optimal filter; 2) the lumped approximation obtained by lumping AB_t at each time; 3) the lumped approximation obtained by lumping A and B_t separately; and 4) as in 3) but using the weighting method of [7]. Maximum *a posteriori* probability estimates of the Markov chain states were computed, and the percentage of incorrect decisions are tabulated in Table II below.

It appears from these simulations that the suboptimal method of [7] is superior, particularly at higher noise levels. This calls into question the argument that the optimal approximation obtained via Lemma 3 (i.e., lumping of AB_t) yields the best results in terms of error probabilities. In order to test this conjecture more thoroughly, further simulations are required. Reference [12] investigates these aspects in more detail. Also, in order to reconcile the algebraic approximation inherent in Lemma 3, with statistical properties of the resulting estimates, we argue that the measure change ideas inherent in [14] will be useful. This is an area of continuing research effort. It should be noted that in general, the lumping of matrices at each time is a computationally intensive procedure which would neutralize any computational savings made due to the two-pass procedure. In practice, one would use a discrete output model (obtained by some appropriate discretization of the continuous output model) and precompute all lumpings (see [12]).

TABLE II
PERCENTAGE ERROR PROBABILITIES FOR FILTERS

σ^2	Optimal	Suboptimal 1	Suboptimal 2	Suboptimal 3
0.01	0.131	0.353	0.413	0.413
0.03	2.035	4.967	5.030	4.545
0.05	4.069	9.213	9.167	8.517

VI. MODEL REDUCTION VIA APPROXIMATE 2-LUMPABILITY

In this section, we propose a model reduction technique for Markov chains and, by extension, hidden Markov processes. In the Markov chain case, the method is based on the approximation of a given transition probability matrix by a 2-lumpable probability matrix. In the HMP case, we approximate the product of a transition probability matrix A, and the data likelihood matrix B_t at each time, by a lumpable probability matrix. Even though the product AB_t is not in general row stochastic, the equations for the unnormalized *a posteriori* probabilities for the HMP states are invariant to row scaling, so we will still seek a 2-lumpable transition probability matrix as the approximation to AB_t. Approximation of HMMs can also be addressed in a similar way. We do not explicitly provide details for the HMP or HMM case, as they follow in much the same way as previous results.

Given $A \in \mathcal{S}_n^+$, we seek a 2-lumping matrix L and $\hat{A} \in \mathcal{S}_n^+$ such that $||A - \hat{A}||$ is minimized subject to $L\hat{A}^T x = 0, \forall x \in \mathcal{N}(T)$. We shall propose an alternating projection algorithm to yield feasible solutions to this problem. Optimality of the approximation appears to be an open issue.

Lemma 4: The affine projection onto $\mathcal{S}_n$ is given by

$$\mathbf{P}_{\mathcal{S}_n}(X) = X\left(\mathbb{I}_n - \frac{\mathbf{1}\mathbf{1}^T}{n}\right) + \frac{\mathbf{1}\mathbf{1}^T}{n}. \tag{27}$$

Proof: We seek the solution to the optimization problem

$$\min_{Y \in \mathbb{R}^{n \times n}} J_1(Y; X) = \|Y - X\|^2 + 2\lambda^T(\mathbf{1} - Y\mathbf{1}). \tag{28}$$

Differentiating with respect to Y yields the necessary condition $Y = X + \lambda\mathbf{1}^T$. The constraint is then forced by the choice $\lambda = (\mathbf{1} - Y\mathbf{1})/n$. Thus the form (27) is obtained. □

Theorem 4: The linear (orthogonal) projection onto $\mathcal{T}_n$ is given by

$$\mathbf{P}_{\mathcal{T}_n}(X) = X - (I - \Phi)X\Phi \tag{29}$$

where Φ is defined by

$$V^T \Phi V = \begin{bmatrix} \mathbb{I}_2 & 0_{2\times(n-2)} \\ 0_{(n-2)\times 2} & 0_{(n-2)\times(n-2)} \end{bmatrix} = \mathcal{I} \tag{30}$$

where $V = [v_1, \ldots, v_n]$ is the (unitary) matrix of right singular vectors of L.

Proof: Let $X \in \mathbb{R}^{n\times n}$ then we seek $Y \in \mathcal{T}_n$ such that $||X - Y||$ is minimized. Let $C = V^T X V$, and $F = V^T Y V$ then by the corollary to Theorem 1, the constraint $LY^T z = 0$, $\forall z \in \mathcal{N}(L)$ is equivalent to F having the structure

$$F = \begin{bmatrix} F_{11} & F_{12} \\ 0_{n-2,2} & F_{22} \end{bmatrix}$$

where $F_{11} \in \mathbb{R}^{2\times 2}$ and $F_{22} \in \mathbb{R}^{(n-2)\times(n-2)}$. Thus $\dim \mathcal{T}_n = n^2 - 2(n-2)$, there being $2(n-2)$ zero elements in the lower left corner of F. Similarly block the matrix C. Thus

$$\begin{aligned} ||X - Y||^2 &= ||C - F||^2 \\ &= ||C_{11} - F_{11}||^2 + ||C_{12} - F_{12}||^2 \\ &\quad + ||C_{22} - F_{22}||^2 + ||C_{21}||^2. \end{aligned} \tag{31}$$

This is clearly minimized by the choice

$$F = \begin{bmatrix} C_{11} & C_{12} \\ 0_{n-2,2} & C_{22} \end{bmatrix} \tag{32}$$

with the resulting error $||X - Y|| = ||C_{21}||$. To see that the projection has the form of (29) consider

$$\begin{aligned} V^T \mathbf{P}_{\mathcal{T}_n}(X) V &= V^T (X - (I - \Phi)X\Phi) V \\ &= C - (I - \mathcal{I})C\mathcal{I} = \begin{bmatrix} C_{11} & C_{12} \\ 0_{n-2,2} & C_{22} \end{bmatrix} \end{aligned} \tag{33}$$

which has the desired form (32). □

Lemma 5: The *convex projection* [16] onto the convex subset $\mathcal{P}$ of $\mathbb{R}^{n\times n}$ is given by

$$[\mathbf{P}_{\mathcal{P}}(X)]_{i,j} = \begin{cases} 0, & X_{i,j} < 0 \\ X_{i,j}, & X_{i,j} \geq 0. \end{cases} \tag{34}$$

Proof: Follows directly from the definition in [16]. □

Theorem 5: Let $A_0 \in \mathcal{S}_n^+$ be given, and define a sequence of iterates in $\mathbb{R}^{n\times n}$ by

$$A_{k+1} = \mathbf{P}_{\mathcal{P}_n} \mathbf{P}_{\mathcal{S}_n} \mathbf{P}_{\mathcal{T}_n} A_k, \qquad k \geq 0 \tag{35}$$

then $A_k \to A^* \in \mathcal{S}_n^+ \cap \mathcal{T}_n$.

Proof: Follows from the alternating convex projection [16, Theorem 2.3-4]. To see this we need only observe that the intersection $\mathcal{T}_n \cap \mathcal{S}_n \cap \mathcal{P}$ is nonempty since there are lumpable probability matrices (see earlier examples) and has finite dimension. □

Comments:

1) It can be shown that $\mathbf{P}_{\mathcal{S}_n^+} \neq \mathbf{P}_{\mathcal{S}_n} \mathbf{P}_{\mathcal{P}} \neq \mathbf{P}_{\mathcal{P}} \mathbf{P}_{\mathcal{S}_n}$ in general. The convex projection onto $\mathcal{S}_n^+$ restricted to $\mathcal{S}_n$ however can be determined analytically, and its use may be preferable for large problems where the speed of convergence of (35) in an issue.
2) Theorem 5 does not say anything about the optimality of the approximation. Optimality is equivalent to saying that A^* in Theorem 5 is the convex projection onto $\mathcal{S}_n^+ \cap \mathcal{T}_n$.

VII. Conclusion

This paper has generalized the concept of lumpability of a Markov chain to hidden Markov processes (HMPs). A vector space approach to the problem has been used to establish conditions for lumpability of a given HMP realization. For the case of a discrete output hidden Markov model (HMM) we have proven necessary and sufficient conditions for that HMM to be lumpable. A HMP arising from a lumpable HMM is always lumpable. The optimal filter for the aggregated (lumped) states has been derived.

We have also applied these results to obtain lumpable approximations in the case where a given HMP may only be approximately lumpable. We have derived an optimal approximation algorithm based on an algebraic criterion, yielding the lumped model or process, and the associated filter for the aggregated states. A two pass procedure has been derived which yields suboptimal filtered estimates for the atomic states of the original HMP. Significant computational savings are provided, although more simulation is required to make definitive statements about the performance of this suboptimal procedure. Finally, a general and explicit approximation result based on convex projections has been described.

Acknowledgment

The authors thanks B. D. O. Anderson for helpful comments. They also thank the associate editor and anonymous referees for helpful suggestions.

References

[1] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, 1989.
[2] G. D. Brushe and L. B. White, "Spatial filtering of superimposed convolutional coded signals," *IEEE Trans. Commun.*, vol. 45, pp. 1144–1153, Sept. 1997.
[3] L. B. White, "Maximum *a posteriori* probability line tracking for nonstationary processes," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Processing*, Toronto, Canada, 1991.
[4] ——, "Cartesian hidden Markov models with applications," *IEEE Trans. Signal Processing*, vol. 40, pp. 1601–1604, 1992.
[5] J. G. Kemeny and J. L. Snell, *Finite Markov Chains.* Princeton, NJ: Van Nostrand, 1960.
[6] H. A. Simon and A. Ando, "Aggregation of variables in dynamical systems," *Econometrica*, vol. 29, pp. 111–138, 1961.
[7] K. Dogancay and V. Krishnamurthy, "Quick aggregation of Markov chain functionals via stochastic complementation," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Processing*, Munich, Germany, 1997, pp. 63–66.
[8] G. Yin and Q. Zhang, "Singularly perturbed Markov chains (I): Asymptotic properties," in *Proc. IEEE Conf. Decision Control*, San Diego, CA, Dec. 1997, pp. 1103–1108.
[9] J. Ledoux, "On weak lumpability of denumerable Markov chains," *Statistics and Probability Letters*, vol. 25, pp. 329–339, 1995.
[10] H. Ito, S.-I. Amari, and K. Kobayashi, "Identifiability of hidden Markov information sources and their minimal degrees of freedom," *IEEE Trans. Inform. Theory*, vol. 38, pp. 324–333, 1992.

[11] M. Karan, "Frequency tracking and hidden Markov models," Ph.D. dissertation, Australian National Univ., 1995.
[12] L. B. White, G. D. Brushe, and R. Mahony, "Suboptimal filtering for large hidden Markov processes via aggregation," in *Third Biennial Eng. Math. Applicat. Conf.*, Adelaide, Australia, July 1998, pp. 515–518.
[13] ——, (1999, Nov.) m-lumpability for hidden Markov models. Department Electrical Electronic Eng. Rep., Univ. Adelaide. [Online]. Available: www.eleceng.adelaide.edu.au/personal/lwhite/reports/lumpyhmm.html
[14] R. J. Elliott, L. Aggoun, and J. B. Moore, *Hidden Markov Models: Estimation and Control*: Springer-Verlag Applications of Mathematics, 1995, vol. 29.
[15] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. Baltimore, MD: Johns Hopkins Univ. Press, 1996.
[16] D. C. Youla, "Mathematical theory of image restoration by the method of convex projections," in *Image Recovery: Theory and Application*, H. Stark, Ed. Orlando, FL: Academic, 1987.

Langford B. White (M'85) received the Bachelor of Science (Maths), Bachelor of Engineering (Hons), and Ph.D. degrees in electrical engineering from the University of Queensland, Brisbane, Australia in 1984, 1985, and 1989, respectively. From 1986 to 1999, he worked for the Defence Science and Technology Organization, Salisbury, South Australia in the areas of radar and communications electronic warfare. In 1994, he was awarded the Australian Telecommunications and Electronics Research Board medal for outstanding young investigator. Since 1999, he has been a Professor of Communication Networks in the Department of Electrical and Electronic Engineering, The University of Adelaide. His current research interests include communications signal processing, wireless data networks, and the internet.

Robert Mahony received the science degree majoring in applied mathematics and geology from the Australian National University (ANU) in 1989. After working for a year as a geophysicist processing marine seismic data he returned to study at ANU and obtained the Ph.D. degree in systems engineering in 1994. Between 1994 and 1997 he worked as a Research Fellow in the Cooperative Research Centre for Robust and Adaptive Systems based in the Research School of Information Sciences and Engineering, ANU, Australia. From 1997 to 1999 he had a position as a post-doctoral fellow in the CNRS laboratory for Heuristics Diagnostics and complex systems (Heudiasyc), Compiegne University of Technology, FRANCE. Having returned to Australia in late 1999, he is now a Logan Research Fellow in the Department of Engineering and Computer Science at Monash University, Melbourne, Australia. His main interests are in nonlinear control theory, with applications in mechanical systems and motion systems, mathematical systems theory, and optimization techniques with applications in linear algebra and geometry.

Gary D. Brushe (S'93–M'96) received the B.E. degree with first-class honor in electrical and electronic engineering from the James Cook University of North Queensland, Townsville, Australia, in 1989 and the Ph.D. degree in systems engineering from the Australian National University, Canberra, in 1996.

He completed an apprenticeship as an electrical fitter/mechanic in 1984. Since 1989, he has worked for the Defence Science and Technology Organization in Australia, where he is now a Senior Research Scientist in the Communications Division's Signals Analysis Discipline. His research interests include the analysis of digital communications signals, statistical and frequency domain signal processing, and reduced complexity processing.