A Counterexample to Aggregation Based Model Reduction of Hidden Markov Models

Georgios Kotsalis Jeff S. Shamma

***Abstract*—This paper highlights a limitation of state space aggregation based model reduction of Hidden Markov Models. We construct an $\hat{n}$ dimensional Hidden Markov Model that is equivalent to a lower dimensional one of order $n < \hat{n}$ for which aggregation based reduction of the high dimensional model is guaranteed not to produce an error free low dimensional model of order n.**

I. Introduction

Hidden Markov Models (HMM's) are one of the most basic and widespread modeling tools for discrete-time stochastic processes that take values on a finite alphabet. A comprehensive review paper is [6]. Applications of HMM's are found across the spectrum of engineering and science in fields as diverse as speech processing, computational biology and financial econometrics (e.g. [12], [9] and [3]).

Very often the cardinality of the state space of the underlying Markov chain renders the use of a given HMM for statistical inference or decision making purposes as infeasible, motivating the investigation of possible algorithms that compress the state space without incurring much loss of information. In [18] it was suggested that the concept of approximate lumpability can be used in the context of model reduction of HMM's. Further work on aggregation based model reduction of HMM's can be found in [17], [4]. In contrast to aggregation based methods, in [8] the authors develop a balanced truncation based model reduction algorithm for HMM's, that is characterized by an a priori computable bound to the approximation error.

Apart from the a priori bound to the approximation error the balanced truncation type algorithm proposed in [8] offers several additional advantages over aggregation based reduction methods. In particular no structural assumptions are imposed to the high dimensional model and the appropriate projection operator is computed by solving Lyapunov like linear algebraic equations. In contrast, aggregation based reduction methods don't come with a priori bounds on the approximation error. They require in essence the solution of a combinatorial optimization problem whose complexity grows exponentially in the size of the underlying state space for the determination of the appropriate aggregation operator and relaxations to that problem are based only on qualitative arguments that involve the weak lumpability structural assumption.

Georgios Kotsalis and Jeff S. Shamma are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA. Email: georgios.kotsalis@ece.gatech.edu, and shamma@gatech.edu.

Research was supported in part by the AFOSR grants FA9550-05-1-0321, FA9550-09-1-0420, and FA9550-05-1-0321

In this paper a further shortcoming of aggregation based reduction methods is exposed. It is shown that there exist exactly reducible high dimensional models and aggregation based reduction is guaranteed not to produce an error free low dimensional model. In contrast to that, if exact reduction of a high dimensional model is possible, the balanced truncation type reduction method in [8] guarantees to produce the minimal equivalent model within the larger class of quasi-realizations (cf.).

A. Notation and mathematical preliminaries

The set of nonnegative integers is denoted by $\mathbf{N}$, the set of positive integers by $\mathbf{Z}_+$ and the set of real numbers by $\mathbf{R}$. The set of positive integers between 1 and n inclusive is denoted by $\mathbf{Z}_n$, i.e. $\mathbf{Z}_n = \{1, ..., n\}$. For $n \in \mathbf{Z}_+$ let $\mathbf{R}^n$ denote the Euclidean n-space. The transpose of a column vector $x \in \mathbf{R}^n$ is x'. For $x \in \mathbf{R}^n$ let $|x|^2 = x'x$ denote the square of the Euclidean norm. For $P \in \mathbf{R}^{n\times n}$ let $P > 0, (P \geq 0)$ indicate that it is a positive (semi-)definite matrix. The identity matrix in $\mathbf{R}^{n\times n}$ is written as I_n. The set of all permutation matrices of size n is denoted by $\mathcal{P}_n$. For $1_n \in \mathbf{R}^n$, $1'_n = (1, \ldots, 1)$. For $n, m \in \mathbf{Z}_+$, $a, b \in \mathbf{R}$, with $a < b$, let

$$[a, b]^{n\times m} = \{A \in \mathbf{R}^{n\times m} \mid A_{ij} \in [a, b],\ \forall i \in \mathbf{Z}_n,\ \forall j \in \mathbf{Z}_m\}.$$

Let $a \in \mathbf{R}^n$, $\mathrm{diag}[a]$ denotes the diagonal matrix

$$\begin{bmatrix} a_1 & & 0 \\ & \ddots & \\ 0 & & a_n \end{bmatrix}.$$

Let $A = \mathrm{rand}[S]$ stand for A being sampled uniformly out of the elements of S. For two sets A, B denote their difference as $A - B = \{x \in A \mid x \notin B\}$. For $n, m \in \mathbf{Z}_+$ and $A \in \mathbf{R}^{n\times m}$, the notation $A \gg 0$ indicates that $A_{ij} > 0, \forall i \in \mathbf{Z}_n, \forall j \in \mathbf{Z}_m$.

II. Preliminaries on Hidden Markov Models

A. HMM's and their statistical description

Hidden Markov Models can be defined in many equivalent ways. The basic definitions and notation introduced in the context of realization theory of HMM's will be used. One can find them for instance in slightly varying language in [11], [1], [15], [16] .

Let $\{Y(t)\}$ be a discrete-time, stationary stochastic process over some fixed probability space $\{\Omega, \mathcal{F}, \mathbf{P}\}$, with values on a finite set $\mathbf{Y} = \mathbf{Z}_m$, $m \geq 2$. The set $\mathbf{Y}$ is called the alphabet and its elements are referred to as letters. For a given $\mathbf{Y}$, define $\mathbf{Y}^*$ as the set of all finite sequences of elements of $\mathbf{Y}$, including the empty sequence, denoted by $\emptyset$.

The finite sequences of letters are called words or strings, if needed they are surrounded by quotation marks to avoid confusion. The set $\mathbf{Y}^*$, referred to as the language, is equipped with a non-commutative "multiplication" operation defined as concatenation of strings and the identity element is the empty sequence $\emptyset$, (i.e. $\mathbf{Y}^*$ is a monoid).

Let v be a word, its length is denoted by $|v|$, and by convention $|\emptyset| = 0$. The set of all strings of length $k \in \mathbf{N}$ is denoted by $\mathbf{Y}^k$. The concatenation of v and u is written as vu, and $|vu| = |v| + |u|$. Strings are read from right to left, in the sense that in the expression vu, u is followed by v. The strict future of the process after time t is denoted by $Y_t^+ = \{\dots, Y(t+2), Y(t+1)\}$ and $Y_t^- = \{Y(t), Y(t-1), \dots\}$ denotes its past and present. Let $v =$ "$v_k \dots v_1$" $\in \mathbf{Y}^*$ the notation $\{Y_t^+ \equiv v\}$ stands for the event $\{\omega \in \Omega \mid Y(t+k) = v_k, \dots, Y(t+1) = v_1\}$, by convention $\{Y_t^+ \equiv \emptyset\} = \Omega$.

Definition 2.1: The **probability function** of the process $\{Y(t)\}$ is a map $p : \mathbf{Y}^* \to \mathbf{R}_+$ where

$$p[v] = \Pr[Y_t^+ \equiv v], \ \forall v \in \mathbf{Y}^*, \forall t \in \mathbf{Z}.$$

Note that since the process is stationary, the value of $p[v]$ in the above definition does not depend on t. It can be readily verified, that the probability function satisfies the properties:

$$p[\emptyset] = 1 \tag{1}$$

$$p[v] \in [0,1], \ \forall v \in \mathbf{Y}^*, \tag{2}$$

$$p[v] = \sum_{u \in \mathbf{Y}^k} p[vu], \ \forall v \in \mathbf{Y}^*, k \in \mathbf{N}. \tag{3}$$

Definition 2.2: Let $\{Y(t)\}$, $\{\tilde{Y}(t)\}$ be discrete-time, stationary stochastic processes over the same alphabet $\mathbf{Y}$. The two stochastic processes are **equivalent** if $\forall t \in \mathbf{Z}, \ \forall v \in \mathbf{Y}^*$

$$\Pr[Y_t^+ \equiv v] = \Pr[\tilde{Y}_t^+ \equiv v]. \tag{4}$$

According to the definition above the two stochastic processes must only coincide in their probability laws in order to be equivalent. They don't have to be defined on the same underlying probability space $\{\Omega, \mathcal{F}, \mathbf{P}\}$. In the context of this work when referring to a stationary stochastic process over the alphabet $\mathbf{Y}$, one is thinking of an equivalence class of processes in the sense of (4). No explicit distinction between the members of the equivalence class is made, the concept of strong realization is not used, it is only the statistical description that matters.

Definition 2.3: A discrete-time, stationary process $\{Y(t)\}$ over the alphabet $\mathbf{Y}$ has a realization as a stationary HMM of size $n \in \mathbf{Z}_+, n \geq 2$ of the **joint Markov process type** if there exists a pair of discrete-time, stationary stochastic processes $\{X(t)\}$, $\{\tilde{Y}(t)\}$ taking values on the finite sets $\mathbf{X} = \mathbf{Z}_n$ and $\mathbf{Y}$ respectively, such that $\{Y(t)\}$ and $\{\tilde{Y}(t)\}$ are equivalent, the joint process $\{X(t), \tilde{Y}(t)\}$ is a Markov process and $\forall \sigma \in \mathbf{X}^*$, $\forall v \in \mathbf{Y}^*$, the following "splitting property" holds

$$\Pr[X_t^+ \equiv \sigma, \tilde{Y}_t^+ \equiv v | X_t^-, \tilde{Y}_t^-] = \Pr[X_t^+ \equiv \sigma, \tilde{Y}_t^+ \equiv v | X(t)].$$

The above definition insures that $\{X(t)\}$ is by itself a Markov chain of order n, meaning

$$\Pr[X_t^+ \equiv \sigma | X_t^-] = \Pr[X_t^+ \equiv \sigma | X(t)].$$

It also insures that $\{\tilde{Y}(t)\}$ is a probabilistic function of the Markov chain $\{X(t-1)\}$ in the sense that

$$\Pr[\tilde{Y}_t^+ \equiv v | X_t^-, \tilde{Y}_t^-] = \Pr[\tilde{Y}_t^+ \equiv v | X(t)].$$

Consider the map $M : \mathbf{Y}^* \to \mathbf{R}_+^{n \times n}$ where

$$M[v]_{ij} = \Pr[X(t + |v|) = i, \tilde{Y}_t^+ \equiv v | X(t) = j],$$

$i, j \in \mathbf{X}, v \in \mathbf{Y}^*, t \in \mathbf{N}$. Note that the state transition matrix of the underlying Markov process $\{X(t)\}$ is given by

$$\Pi = \sum_{v \in \mathbf{Y}} M[v].$$

Consider $\pi \in \mathbf{R}_+^n$, such that $\Pi\pi = \pi$, $1_n'\pi = 1$. The vector π corresponds to an invariant distribution of $\{X(t)\}$, which is unique if the Markov process has a single ergodic class. Since the processes $\{Y(t)\}$ and $\{\tilde{Y}(t)\}$ are equivalent, one has

$$p[v] = \Pr[Y_t^+ \equiv v] = \Pr[\tilde{Y}_t^+ \equiv v], \forall t \in \mathbf{Z}, \ \forall v \in \mathbf{Y}^*.$$

Assumption In the following it assumed that the entries of any stochastic transition matrix are strictly positive. This is a sufficient condition for uniqueness of the stationary distribution of the corresponding Markov process.

Lemma 2.1: Consider $k \in \mathbf{N}$, $v =$ "$v_k v_{k-1} \dots v_1$" $\in \mathbf{Y}^k$. The probability of that particular string can be computed recursively according to

$$p[v] = 1_n' M[v] \pi,$$

where

$$M[v] = M[v_k] \dots M[v_1], \qquad M[\emptyset] = I_n.$$

Proof: See for instance [1], [16].

The preceding lemma shows that if a given stationary process $\{Y(t)\}$ over the alphabet $\mathbf{Y}$ has a realization as a stationary HMM of size n of the joint Markov process type, then its probability function is completely encoded by the ordered triple $\mathbf{H} = (1_n, \{M[v], v \in \mathbf{Y}\}, \pi)$. Accordingly in the following discussion referring to a HMM of size n of the joint Markov process type will be in terms of the ordered triple $\mathbf{H} = (1_n, \{M[v], v \in \mathbf{Y}\}, \pi)$. The space of all HMM's of size n of the joint Markov process type over the alphabet $\mathbf{Y}$ is denoted by $\mathcal{H}_{n,\mathbf{Y}}$. An element of that space will be abbreviated as a JMP HMM. The preceding definition of a HMM leads to the most economical description in terms of the size of the underlying state space. The next two definitions of a HMM are common in the literature and are used in the context of aggregation.

Definition 2.4: A discrete-time stationary process $\{Y_t\}$ has a realization as stationary HMM of size $n \in \mathbf{Z}_+, n \geq 2$ of the **deterministic function of a Markov chain type**, if there exists a discrete-time, stationary Markov process $\{X(t)\}$, taking values on the finite set $\mathbf{X}$ and a function $f : \mathbf{X} \to \mathbf{Y}$ such that

$Y_t = f(X_t)$. The statistical description of a HMM $\mathbf{H}$ of the deterministic function of a Markov chain type is given by the tuple $\mathbf{H} = (1_n, O, \Pi, \pi)$ where $O \in \{0,1\}^{n\times m}, \Pi \in \mathbf{R}^{n\times n}$ and $\pi \in \mathbf{R}^{n\times 1}$. In particular

$$\begin{aligned} \Pi_{ij} &= \Pr[X_{t+1} = i | X_t = j] \\ O_{ij} &= \begin{cases} 1 & \text{if } j = f(i), \\ 0 & \text{otherwise.} \end{cases} \\ \pi_i &= \Pr[X_0 = i] \end{aligned}$$

The space of all HMM's of size n of the deterministic function of a Markov chain type over the alphabet $\mathbf{Y}$ is denoted by $\mathcal{H}^D_{n,\mathbf{Y}}$. An element $\mathbf{H} = (1_n, O, \Pi, \pi) \in \mathcal{H}^D_{n,\mathbf{Y}}$ will be abbreviated as a DFMC HMM.

Definition 2.5: A discrete-time stationary process $\{Y_t\}$ has a realization as stationary HMM of size $n \in \mathbf{Z}_+, n \geq 2$ of the **random function of a Markov chain type**, if there exists a discrete-time, stationary stochastic process $\{X(t)\}$, taking values on the finite set $\mathbf{X}$ and a discrete-time, stationary stochastic process $\{\tilde{Y}(t)\}$, taking values on the finite set $\mathbf{Y}$, as well as matrices $O \in \mathbf{R}^{n\times m}, \Pi \in \mathbf{R}^{n\times n}$ and $\pi \in \mathbf{R}^{n\times 1}$, such that

$$\begin{aligned} \Pi_{ij} &= \Pr[X_{t+1} = i | X_t = j] \\ O_{ij} &= \Pr[\tilde{Y}_t = j | X_t = i] \\ \pi_i &= \Pr[X_0 = i] \end{aligned}$$

and the processes $Y_t, \tilde{Y}_t$ have the same probability law. The i'th column of O will be denoted as o_i, so that

$$O = [o_1, \ldots, o_m].$$

The space of all HMM's of size n of the random function of a Markov chain type over the alphabet $\mathbf{Y}$ is denoted by $\mathcal{H}^R_{n,\mathbf{Y}}$. An element $\mathbf{H} = (1_n, O, \Pi, \pi) \in \mathcal{H}^R_{n,\mathbf{Y}}$ will be abbreviated as a RFMC HMM. Let $\mathbf{H} \in \mathcal{H}^D_{n,\mathbf{Y}}$, it is immediate that $\mathbf{H} \in \mathcal{H}^R_{n,\mathbf{Y}}$, and therefor $\mathcal{H}^D_{n,\mathbf{Y}} \subset \mathcal{H}^R_{n,\mathbf{Y}}$. Given $\mathbf{H} = (1_n, O, \Pi, \pi) \in \mathcal{H}^R_{n,\mathbf{Y}}$ there exists $\tilde{\mathbf{H}} = (1_n, \{\tilde{M}[v], v \in \mathbf{Y}\}, \tilde{\pi}) \in \mathcal{H}_{n,\mathbf{Y}}$ such that

$$p_{\mathbf{H}}[v] = p_{\tilde{\mathbf{H}}}[v], \quad \forall v \in \mathbf{Y}^*.$$

To see this one can set

$$\begin{aligned} \tilde{\pi} &= \pi, \\ \tilde{M}[y] &= \mathrm{diag}[o_y]\ \Pi, \quad y \in \mathbf{Y}. \end{aligned}$$

On the above grounds one has $\mathcal{H}^R_{n,\mathbf{Y}} \subset \mathcal{H}_{n,\mathbf{Y}}$.

B. Generalized Automata and Quasi-realizations of finite stochastic systems.

The concept of a generalized automaton was formally introduced in [14]. Prior to that, with a slightly different terminology the same objects appeared in connection wih quasi-realizations of discrete-time, finite valued stochastic processes of finite rank, see [16] for more information and the references therein. Generalized automata (GA) are equivalent to recognizable Formal Power Series in several noncommuting indeterminates with real coefficients, that have been frequently used in the study of formal languages in theoretical computer science, see for instance [2], [5], [13].

Definition 2.6: A **generalized automaton** of size n over the alphabet $\mathbf{Y}$ is defined as an ordered triple $\mathbf{G} = (c, \{A[v], v \in \mathbf{Y}\}, b)$, where $c \in \mathbf{R}^n$, $A : \mathbf{Y} \to \mathbf{R}^{n\times n}$, $b \in \mathbf{R}^n$.

Some of the references in the literature including [14] incorporate a finite state space $\mathbf{X}$ of cardinality n in the definition, thus making $\mathbf{G}$ an ordered quadruple. This is not pursued in this work since no explicit use of the state space $\mathbf{X}$ is being made. Let $v =$ "$v_k \ldots v_1$", $\in \mathbf{Y}^*$, where $k \in \mathbf{N}$, the domain of A is extended from $\mathbf{Y}$ to $\mathbf{Y}^*$ by means of the homomorphism

$$A[v] = A[v_k] \ldots A[v_1], \ A[\emptyset] = I_n.$$

Definition 2.7: The **word function** of $\mathbf{G}$ is a map $q_{\mathbf{G}} : \mathbf{Y}^* \to \mathbf{R}$, where

$$q_{\mathbf{G}}[v] = c' A[v] b, \ \forall v \in \mathbf{Y}^*.$$

The set of all GA of size n over the alphabet $\mathbf{Y} = \{1, \ldots, m\}$ is denoted by $\mathcal{G}_{n,\mathbf{Y}}$.

Definition 2.8: Let $n_i \geq 2$, $i \in \{1, 2\}$. Two generalized automata $\mathbf{G}_1 \in \mathcal{G}_{n_1,\mathbf{Y}}$, $\mathbf{G}_2 \in \mathcal{G}_{n_2,\mathbf{Y}}$ are **equivalent** if

$$q_{\mathbf{G}_1}[v] = q_{\mathbf{G}_2}[v], \quad \forall v \in \mathbf{Y}^*.$$

Definition 2.9: The GA $\mathbf{G} \in \mathcal{G}_{n,\mathbf{Y}}$ is **minimal** if for all $\tilde{n}$, $2 \leq \tilde{n} < n$ there is no $\tilde{\mathbf{G}} \in \mathcal{G}_{\tilde{n},\mathbf{Y}}$ that is equivalent to $\mathbf{G}$.

Lemma 2.2: Two generalized automata $\mathbf{G}_1 \in \mathcal{G}_{n,\mathbf{Y}}$, $\mathbf{G}_2 \in \mathcal{G}_{n,\mathbf{Y}}$ are equivalent if and only if there exists a non singular matrix $T \in \mathbf{R}^{n\times n}$ such that

$$\begin{aligned} A_1[y] &= T^{-1} A_2[y] T, \quad \forall y \in \mathbf{Y}, \\ b_1 &= T^{-1} b_2, \\ c_1 &= c_2 T. \end{aligned}$$

Proof: See for instance [16].

A HMM can be interpreted as a generalized automaton. Let $\mathbf{H} = (1_n, \{M[v], v \in \mathbf{Y}\}, \pi) \in \mathcal{H}_{n,\mathbf{Y}}$ then clearly $\mathbf{G} \in \mathcal{G}_{n,\mathbf{Y}}$ with $c = 1_n$, $A[y] = M[y]$, $y \in \mathbf{Y}$, $b = \pi$ satisfies $q_{\mathbf{G}}[v] = p_{\mathbf{H}}[v]$, $\forall v \in \mathbf{Y}^*$, thus $\mathcal{H}_{n,\mathbf{Y}} \subset \mathcal{G}_{n,\mathbf{Y}}$. This observation motivated the concept of a quasi-realization.

Definition 2.10: A **quasi-realization** of order n of a discrete-time, finite rank, finite valued, stationary stochastic process $\{Y(t)\}$ over the alphabet $\mathbf{Y}$ is a generalized automaton $\mathbf{G} \in \mathcal{G}_{n,\mathbf{Y}}$, whose word function satisfies

$$q[v] = \Pr[Y_t^+ \equiv v] = c' A[v] b, \ \forall v \in \mathbf{Y}^*,$$

and additionally

$$\begin{aligned} c' &= c' (\sum_{v \in \mathbf{Y}} A[v]), \\ b &= (\sum_{v \in \mathbf{Y}} A[v]) b. \end{aligned}$$

The quasi-realization is minimal if the size of the automaton equals the rank of the given process. Minimal quasi-realizations are also termed as regular. The underlying parameters of the automaton, being arbitrary real numbers, do not necessarily have a probabilistic interpretation. The connection between discrete-time, finite valued, stationary stochastic processes of finite rank and GA, has been long recognized in the literature, essentially in the work of [7].

III. Aggregation of Hidden Markov Models

The concept of state space aggregation as means of reducing the dimensionality of HMM's has been studied in [18], [17], [4]. Let $\mathbf{Y} = \mathbf{Z}_m$ with $m \geq 2$. Consider two HMM's $\mathbf{H} = (1_n, O, \Pi, \pi) \in \mathcal{H}^R_{n,\mathbf{Y}}$ and $\hat{\mathbf{H}} = (1_n, O, \Pi, \pi) \in \mathcal{H}^R_{\hat{n},\mathbf{Y}}$ with state spaces $\mathbf{S} = \{s_1, \ldots, s_n\}$ and $\hat{\mathbf{S}} = \{\hat{s}_1, \ldots, \hat{s}_{\hat{n}}\}$ respectively, where $\hat{n} \geq 3$, $n < \hat{n}$. For the two models to be related by aggregation one needs to introduce a surjective map

$$\phi : \hat{\mathbf{S}} \to \mathbf{S},$$

that naturally partitions the domain into equivalence classes. For a given $s \in \mathbf{S}$, define $\Phi_s = \{\hat{s} \in \hat{\mathbf{S}} \mid \phi(\hat{s}) = s\}$, one has

$$\Phi_s \cap \Phi_{\bar{s}} = \emptyset \;\text{ if }\; s \neq \bar{s} \;\text{ and }\; \bigcup_{s \in \mathbf{S}} \Phi_s = \hat{\mathbf{S}}.$$

The surjective map ϕ is called a partition function. The set of all partition functions involving the sets $\hat{\mathbf{S}}$ and $\mathbf{S}$ is denoted by $\mathcal{S}_{\hat{n},n}$,

$$\mathcal{S}_{\hat{n},n} = \{\phi : \hat{\mathbf{S}} \to \mathbf{S} \mid \phi \;\text{ is surjective}\}.$$

Definition 3.1: For a given $\phi \in \mathcal{S}_{\hat{n},n}$, the corresponding **aggregation operator** L_ϕ is a map $L_\phi : \mathbf{R}^{\hat{n}} \to \mathbf{R}^n$ with

$$L_{\phi_{ij}} = \begin{cases} 1 & \text{if } \phi(\hat{s}_j) = s_i, \\ 0 & \text{otherwise.} \end{cases}$$

The set of all possible aggregation operators between the sets $\hat{\mathbf{S}}$ and $\mathbf{S}$ is denoted by $\mathcal{L}_{\hat{n},n}$,

$$\mathcal{L}_{\hat{n},n} = \{L_\phi \mid \phi \in \mathcal{S}_{\hat{n},n}\}.$$

The corresponding set of dilation operators is denoted by $\mathcal{D}_{n,\hat{n}}$,

$$\mathcal{D}_{n,\hat{n}} = \{D_\phi : \mathbf{R}^n \to \mathbf{R}^{\hat{n}} \mid L_\phi D_\phi = I_n,\; L_\phi \in \mathcal{L}_{\hat{n},n}\}.$$

Frequently the dilation operators are restricted to be of the form

$$D_{\phi_{ij}} = \begin{cases} \mu_i & \text{if } \phi(\hat{s}_i) = s_j, \\ 0 & \text{otherwise,} \end{cases}$$

where

$$\sum_{i \in \Phi_{s_j}} \mu_i = 1, \quad \mu_i \geq 0, \quad \forall i \in \{1, \ldots, \hat{n}\}.$$

In that case μ_i admits the interpretation of conditional probability of state $\hat{s}_i$ within the cluster $\phi(\hat{s}_i) = s_j$. These conditional probabilities relate to the nonstationary Markovian evolution of the aggregated state process, see for instance [10]. This structural restriction on dilation operators is not imposed in this work.

Definition 3.2: Consider two HMM's, $\hat{\mathbf{H}} = (1_n, O, \Pi, \pi) \in \mathcal{H}^R_{\hat{n},\mathbf{Y}}$ with state space $\hat{\mathbf{S}} = \{\hat{s}_1, \ldots, \hat{s}_{\hat{n}}\}$ and $\mathbf{H} = (1_n, O, \Pi, \pi) \in \mathcal{H}^R_{n,\mathbf{Y}}$ with state space $\mathbf{S} = \{s_1, \ldots, s_n\}$. Model $\mathbf{H}$ is the outcome of an **aggregation based model reduction** applied to $\hat{\mathbf{H}}$ if there exists $L_\phi \in \mathcal{L}_{\hat{n},n}$ and $D_\phi \in \mathcal{D}_{n,\hat{n}}$ such that

$$\pi = L_\phi \hat{\pi} \tag{5}$$

$$\Pi = L_\phi \hat{\Pi} D_\phi \tag{6}$$

$$\mathrm{diag}[o_y] = L_\phi \,\mathrm{diag}[\hat{o}_y]\, D_\phi, \quad \forall y \in \mathbf{Y}. \tag{7}$$

Accordingly define for a fixed $\phi \in \mathcal{S}_{\hat{n},n}$ and $D_\phi \in \mathcal{D}_{n,\hat{n}}$

$$\mathcal{A}_{\phi,D_\phi} : \mathcal{H}^R_{\hat{n},\mathbf{Y}} \to \mathcal{H}^R_{n,\mathbf{Y}}$$

where $\mathcal{A}_{\phi,D_\phi}[\hat{\mathbf{H}}] = \mathbf{H}$, and the parameters of $\hat{\mathbf{H}}$ and $\mathbf{H}$ are related by (5) - (7).

IV. Main result

A. A counterexample to aggregation based model reduction of Hidden Markov Models.

In this section we construct an $\hat{n}$ dimensional Hidden Markov Model that is equivalent to a lower dimensional one of order $n < \hat{n}$ for which aggregation based reduction of the high dimensional model is guaranteed not to produce the low dimensional model nor one that is equivalent to it and has order n.

Some more notation needs to be introduced. Given $\mathbf{H} \in \mathcal{H}^R_{n,\mathbf{Y}}$, let $\mathcal{E}^R_{\mathbf{H}}$ denote the set of RFMC HMM's that are equivalent to $\mathbf{H}$ and have same size, i.e.

$$\mathcal{E}^R_{\mathbf{H}} = \{\tilde{\mathbf{H}} \in \mathcal{H}^R_{n,\mathbf{Y}} \mid p_{\mathbf{H}} \equiv p_{\tilde{\mathbf{H}}}\}.$$

Similarly define

$$\mathcal{E}_{\mathbf{H}} = \{\tilde{\mathbf{H}} \in \mathcal{H}_{n,\mathbf{Y}} \mid p_{\mathbf{H}} \equiv p_{\tilde{\mathbf{H}}}\},$$

$$\mathcal{E}^G_{\mathbf{H}} = \{\tilde{\mathbf{G}} \in \mathcal{G}_{n,\mathbf{Y}} \mid p_{\mathbf{H}} \equiv q_{\tilde{\mathbf{G}}}\}.$$

Lemma 4.1: Consider $\mathbf{H} = (1_n, O, \Pi, \pi) \in \mathcal{H}^R_{n,\mathbf{Y}}$. Suppose that an arbitrary $\mathbf{G} \in \mathcal{E}^G_{\mathbf{H}}$ is minimal. Furthermore suppose that Π is a full rank matrix and that for some $y^* \in \mathbf{Y}$, if $i \neq j \to O_{iy^*} \neq O_{jy^*}$, $\forall i, j \in \mathbf{Z}_n$. Every $\tilde{\mathbf{H}} = (1_n, \tilde{O}, \tilde{\Pi}, \tilde{\pi}) \in \mathcal{E}^R_{\mathbf{H}}$ is obtained by permuting the states of $\mathbf{H}$.

Proof: Lemma 2.2 is employed. In particular there exists a non singular matrix $T \in \mathbf{R}^{n \times n}$, such that

$$\mathrm{diag}[\tilde{o}_y]\tilde{\Pi} = T^{-1}\mathrm{diag}[o_y]\Pi T, \quad \forall y \in \mathbf{Y} \tag{8}$$

$$1'_n = 1'_n T \tag{9}$$

Using the fact that

$$\sum_{y \in \mathbf{Y}} \mathrm{diag}[\tilde{o}_y] = \sum_{y \in \mathbf{Y}} \mathrm{diag}[o_y] = I_n$$

one obtains from (8)

$$\tilde{\Pi} = T^{-1}\Pi T$$

which shows that $\tilde{\Pi}$ is a full rank matrix as well. Moreover combining the above equation and (8) gives

$$T\mathrm{diag}[\tilde{o}^*_y] = \mathrm{diag}[o^*_y]T. \tag{10}$$

Since T is full rank $\forall j \in \{1,\dots,n\}\ \exists k : T_{kj} \neq 0$. Using (10) this implies that $\tilde{O}_{y^*j} = O_{ky^*}$. Suppose that $\exists m \neq k : T_{mj} \neq 0$. This would imply $\tilde{O}_{y^*j} = O_{my^*}$ and therefor $O_{my^*} = O_{ky^*}$ which contradicts the assumption that $\forall m \neq k\ O_{my^*} \neq O_{ky^*}$. So one has that $\forall j \in \{1,\dots,n\}\ \exists! k : T_{kj} \neq 0$ and using (9) one gets that $T_{kj} = 1$ which shows that $T \in \mathcal{P}_n$.

Let $\mathcal{H}^{R,P}_{n,\mathbf{Y}} \subset \mathcal{H}^{R}_{n,\mathbf{Y}}$ denote the set of models that fulfill the assumptions of lemma 4.1. If $\mathbf{H} \in \mathcal{H}^{R,P}_{n,\mathbf{Y}}$ then $\mathcal{E}^R_{\mathbf{H}}$ is a finite set.

Lemma 4.2: Consider $\mathbf{H} = (1_n, \{M[v], v \in \mathbf{Y}\}, \pi) \in \mathcal{H}_{n,\mathbf{Y}}$. Let $\hat{n} = n \times |\mathbf{Y}| = n \times m$. There exists $\hat{\mathbf{H}} = (1_{\hat{n}}, \hat{O}, \hat{\Pi}, \hat{\pi}) \in \mathcal{H}^R_{\hat{n},\mathbf{Y}}$ that is equivalent to $\mathbf{H}$.

Proof: The lemma is intuitively true based on the following reasoning. For a given JMP HMM the joint process $\{X_t, Y_t\}$ is a Markov process, and one can clearly write $Y(t) = f(X(t), Y(t))$ for in fact a deterministic function f. Formally for $\mathbf{H} = (1_n, \{M[v], v \in \mathbf{Y}\}, \pi) \in \mathcal{H}_{n,\mathbf{Y}}$ let $\hat{\mathbf{H}} = (1_{\hat{n}}, \hat{O}, \hat{\Pi}, \hat{\pi}) \in \mathcal{H}^R_{\hat{n},\mathbf{Y}}$ where

$$\hat{\Pi} = \begin{bmatrix} M[1] & \dots & M[1] \\ \vdots & & \vdots \\ M[m] & \dots & M[m] \end{bmatrix} \tag{11}$$

$$\hat{O}_{iy} = \begin{cases} 1 & \text{if } (y-1)n+1 \leq i \leq ny,\ y \in \mathbf{Y}, \\ 0 & \text{otherwise,} \end{cases} \tag{12}$$

$$\hat{\pi} = \hat{\Pi}\hat{\pi}. \tag{13}$$

Let $\hat{\pi}$ be partitioned as

$$\hat{\pi}' = [\hat{\pi}'_1, \dots, \hat{\pi}'_m].$$

It follows that

$$M[y](\sum_{k \in \mathbf{Y}} \hat{\pi}_k) = \hat{\pi}_y,\ y \in \mathbf{Y},$$

and therefor

$$\sum_{k \in \mathbf{Y}} \hat{\pi}_k = \pi.$$

Let

$$\hat{M}[y] = \mathrm{diag}[\hat{o}_y]\hat{\Pi}, \quad y \in \mathbf{Y},$$

and note that

$$1'_{\hat{n}} = [1'_n, \dots, 1'_n].$$

For an arbitrary $v =$ "$v_k \dots v_1$", $\in \mathbf{Y}^*$, where $k \in \mathbf{N}$ one has

$$\begin{aligned} p_{\hat{\mathbf{H}}}[v] &= 1'_{\hat{n}}\hat{M}[v_k]\dots\hat{M}[v_1]\hat{\pi} \\ &= [\dots, 1'_n M[v_k]\dots M[v_1], \dots] \begin{bmatrix} \hat{\pi}_1 \\ \vdots \\ \hat{\pi}_m \end{bmatrix} \\ &= 1'_n M[v_k]\dots M[v_1](\sum_{k \in \mathbf{Y}} \hat{\pi}_k) \\ &= 1'_n M[v_k]\dots M[v_1]\pi = p_{\mathbf{H}}[v], \end{aligned}$$

proving that $p_{\hat{\mathbf{H}}} \equiv p_{\mathbf{H}}$.

The above construction will be summarized by introducing the map

$$\mathcal{C}_{n,\hat{n}} : \mathcal{H}^R_{n,\mathbf{Y}} \to \mathcal{H}_{\hat{n},\mathbf{Y}}$$

where $\mathcal{C}_{n,\hat{n}}[\mathbf{H}] = \hat{\mathbf{H}}$, and the parameters of $\hat{\mathbf{H}}$ and $\mathbf{H}$ are related by (11) - (13).

Theorem 4.1: There exists HMM's $\mathbf{H} = (1_n, O, \Pi, \pi) \in \mathcal{H}^R_{n,\mathbf{Y}}$ and $\hat{\mathbf{H}} = (1_n, O, \Pi, \pi) \in \mathcal{H}^R_{\hat{n},\mathbf{Y}}$ with state spaces $\mathbf{S} = \{s_1, \dots, s_n\}$ and $\hat{\mathbf{S}} = \{\hat{s}_1, \dots, \hat{s}_{\hat{n}}\}$ respectively, that are equivalent, i.e. $p_{\mathbf{H}} \equiv p_{\hat{\mathbf{H}}}$, however $\forall \phi \in \mathcal{S}_{\hat{n},n}$ and $D_\phi \in \mathcal{D}_{n,\hat{n}}$ one has

$$\mathcal{A}_{\phi,D_\phi}[\hat{\mathbf{H}}] \notin \mathcal{E}^R_{\mathbf{H}}.$$

Proof: An example will be provided that justifies the theorem. Let $n = 2, \hat{n} = 4, \mathbf{Y} = \{1, 2\}$. Consider the model $\mathbf{H} = (1_2, O, \Pi, \pi) \in \mathcal{H}^{R,P}_{2,\mathbf{Y}}$, with

$$\begin{aligned} \pi &= \begin{bmatrix} 0.25 \\ 0.75 \end{bmatrix} \\ \Pi &= \begin{bmatrix} 0.70 & 0.10 \\ 0.30 & 0.90 \end{bmatrix} \\ O &= \begin{bmatrix} 0.20 & 0.80 \\ 0.10 & 0.90 \end{bmatrix} \end{aligned}$$

Note that $\mathbf{H}$ fulfills the assumptions of lemma 4.1 and therefor within $\mathcal{H}^R_{2,\mathbf{Y}}$ the only other model equivalent to $\mathbf{H}$ is obtained by permuting the states and will be denoted by $\tilde{\mathbf{H}}$, so $\mathcal{E}^R_{\mathbf{H}} = \{\tilde{\mathbf{H}}, \mathbf{H}\}$. Consider now $\bar{\mathbf{H}} = (1_2, \{\bar{M}[v], v \in \mathbf{Y}\}, \bar{\pi}) \in \mathcal{H}_{2,\mathbf{Y}}$ that is equivalent to $\mathbf{H}$, i.e. $\bar{\mathbf{H}} \in \mathcal{E}_{\mathbf{H}}$. With

$$T = \begin{bmatrix} 1.20 & -0.20 \\ -0.20 & 1.20 \end{bmatrix}$$

the parameters of $\bar{\mathbf{H}}$ are

$$\begin{aligned} \bar{\pi} &= T\pi = \begin{bmatrix} 0.15 \\ 0.85 \end{bmatrix} \\ \bar{M}[1] &= T\mathrm{diag}[o_1]\Pi T^{-1} = \begin{bmatrix} 0.14 & 0.02 \\ 0.03 & 0.09 \end{bmatrix} \\ \bar{M}[2] &= T\mathrm{diag}[o_2]\Pi T^{-1} = \begin{bmatrix} 0.56 & 0.08 \\ 0.27 & 0.81 \end{bmatrix}. \end{aligned}$$

Now consider the model $\hat{\mathbf{H}} = (1_4, \hat{O}, \hat{\Pi}, \hat{\pi}) \in \mathcal{H}^R_{4,\mathbf{Y}}$ that is equivalent to $\bar{\mathbf{H}}$ and therefor to $\mathbf{H}$, where

$$\hat{\mathbf{H}} = \mathcal{C}_{2,4}[\bar{\mathbf{H}}].$$

The unique steady state distribution of $\hat{\Pi}$ is

$$\hat{\pi} = \begin{bmatrix} 0.045 \\ 0.080 \\ 0.105 \\ 0.770 \end{bmatrix},$$

A quick calculation shows that $\forall L_\phi \in \mathcal{L}_{4,2}$ one has

$$L_\phi\hat{\pi} \notin \{\begin{bmatrix} 0.25 \\ 0.75 \end{bmatrix}, \begin{bmatrix} 0.75 \\ 0.25 \end{bmatrix}\}.$$

and therefor $\forall \phi \in \mathcal{S}_{4,2}$ and $D_\phi \in \mathcal{D}_{2,4}$ one has

$$\mathcal{A}_{\phi,D_\phi}[\hat{\mathbf{H}}] \notin \mathcal{E}^R_{\mathbf{H}},$$

completing the proof. In the above proof we established the existence of a pair of equivalent HMM's $\hat{\mathbf{H}}$ and $\mathbf{H}$ of different order for which aggregation is guaranteed not to recover

an error free low dimensional model. Next we establish the generic nature of such instances by providing a randomized algorithmic process that produces further such examples with probability 1. For a given $n \in \mathbf{Z}_+$, $n \geq 2$ and $m \in \mathbf{Z}_+$, $2 \leq m \leq n$, let $\hat{n} = n \times m$.

Step 1: Construct $\mathbf{H} \in \mathcal{H}^{R,P}_{n,\mathbf{Y}}$.

In other words create an HMM that fulfills the assumptions of lemma 4.1. Set

$$\bar{\Pi} = \text{rand}[\ [0,1]^{n\times n}\ ].$$

With $p_s = 1_n'\bar{\Pi}$, let

$$\Pi = \bar{\Pi}\ \text{diag}[\frac{1}{p_{s_1}}, \ldots, \frac{1}{p_{s_n}}].$$

The matrix Π is a column stochastic matrix, thus it is valid state transition matrix, and further $\Pr[\Pi \text{ is full rank}] = 1$. Set

$$\bar{O} = \text{rand}[\ [0,1]^{n\times m}\ ].$$

With $q_s = \bar{O}1_n$, let

$$O = \text{diag}[\frac{1}{q_{s_1}}, \ldots, \frac{1}{q_{s_m}}]\bar{O}.$$

The matrix O is row stochastic, thus it is a valid emission matrix, and further with probability 1 one of its columns consists of distinct entries. Compute the unique steady state probability $\pi \in \mathbf{R}^{n\times 1}$ by solving $\pi = \Pi\pi$. Set $\mathbf{H} = (1_n, O, \Pi, \pi)$, one has that $\mathbf{H} \in \mathcal{H}^{R,P}_{n,\mathbf{Y}}$ with probability 1. Let

$$\mathcal{M}_{\mathbf{H}} = \{v \in \mathbf{R}^{n\times 1} \mid \exists P \in \mathcal{P}_n : v = P\pi\}.$$

Note that if $\tilde{\mathbf{H}} = (1_n, \tilde{O}, \tilde{\Pi}, \tilde{\pi}) \in \mathcal{E}^R_{\mathbf{H}}$ then $\tilde{\pi} \in \mathcal{M}_{\mathbf{H}}$. Let

$$\epsilon_1 = \min_{\pi_1,\pi_2\in\mathcal{M}_{\mathbf{H}},\pi_1\neq\pi_2} ||\pi_1 - \pi_2||_1.$$

This is the minimum distance in terms of the 1 norm between two distinct elements in $\mathcal{M}_{\mathbf{H}}$. By construction $\Pr[\epsilon_1 > 0] = 1$.

Step 2: Construct $\hat{\mathbf{H}} \in \mathcal{H}^R_{\hat{n},\mathbf{Y}} : \exists \phi_0 \in \mathcal{S}_{\hat{n},n}$ with $L_{\phi_0}\hat{\pi} = \pi$. This can be achieved by setting

$$\hat{\mathbf{H}} = \mathcal{C}_{n,\hat{n}}[\mathbf{H}].$$

The required aggregation operator $L_{\phi_0} \in \mathcal{L}_{n,\hat{n}}$ is given by

$$L_{\phi_0} = [I_n, \ldots, I_n].$$

Let

$$\mathcal{L}^0_{n,\hat{n}} = \{L \in \mathcal{L}_{n,\hat{n}} \mid L = PL_{\phi_0}, \quad P \in \mathcal{P}_n\}.$$

With probability 1, it holds that $\forall L \in \mathcal{L}_{n,\hat{n}}$ if $L\hat{\pi} \in \mathcal{M}_{\mathbf{H}}$ then $L \in \mathcal{L}^0_{n,\hat{n}}$. Define

$$\mathcal{L}^c_{n,\hat{n}} = \mathcal{L}_{n,\hat{n}} - \mathcal{L}^0_{n,\hat{n}}.$$

Let

$$\epsilon_2 = \min_{\pi_1\in\mathcal{M}_{\mathbf{H}},L\in\mathcal{L}^c_{n,\hat{n}}} ||L\hat{\pi} - \pi_1||_1.$$

By construction $\Pr[\epsilon_2 > 0] = 1$.

Step 3: Construct $\hat{\mathbf{H}}_T$ in the vicinity of $\hat{\mathbf{H}}$ such that $\forall L \in \mathcal{L}_{n,\hat{n}}$ it holds that $L\hat{\pi}_T \notin \mathcal{M}_{\mathbf{H}}$ with probability 1.

Let

$$\Delta = \text{rand}[\ [-\frac{1}{2}, \frac{1}{2}]^{n\times n}\ ].$$

For $k \in \mathbf{Z}_+$ define

$$\bar{T}_k = I_n + 2^{-k}\Delta,$$

and with $r_s(k) = 1_n'\bar{T}_k$ let

$$T_k = \bar{T}_k \text{diag}[\frac{1}{r_{s_1}(k)}, \ldots, \frac{1}{r_{s_n}(k)}].$$

Note that the entries in each column of T_k sum up to 1. Let

$$B_k[y] = T_k \text{diag}[o_y]\Pi T_k^{-1}, \quad y \in \mathbf{Y}.$$

Find the smallest $k \in \mathbf{Z}_+$ such that

$$B_k[y] >> 0, \quad \forall y \in \mathbf{Y},$$

and call it k_0. Note that k_0 is guaranteed to exist with probability 1 since as k increases T_k, T_k^{-1} converge towards I_n and $\Pr[\Pi >> 0] = \Pr[O >> 0] = 1$. Let

$$M_k[y] = B_k[y], \quad y \in \mathbf{Y},\ k \geq k_0$$

and $\pi_k = T_k\pi,\ k \geq k_0$. Let $\mathbf{H}_k = (1_n, \{M_k[v], v \in \mathbf{Y}\}, \pi_k)$ and

$$\hat{\mathbf{H}}_k = \mathcal{C}_{n,\hat{n}}[\mathbf{H}_k].$$

Consider the following conditions:

$$\Pr[L_{\phi_0}(\hat{\pi}_k - \hat{\pi})) \neq 0] = 1, \tag{14}$$

$$max_{L\in\mathcal{L}^0_{n,\hat{n}}} ||L(\hat{\pi}_k - \hat{\pi})||_1 < \epsilon_1, \tag{15}$$

$$max_{L\in\mathcal{L}^c_{n,\hat{n}}} ||L(\hat{\pi}_k - \hat{\pi})||_1 < \epsilon_2. \tag{16}$$

Condition (14) holds by virtue of the random perturbation induced by T_k. Let Find the smallest $k \geq k_0$ such that (15) holds and call it k_1. Similarly find the smallest $k \geq k_0$ such that (16) holds and call it k_2. Note that k_1, k_2 are guaranteed to exist by continuity arguments. As k increases $\hat{\pi}_k$ converges to $\hat{\pi}$. Set $k^* = \max\{k_1, k_2\}$. Let

$$M_T[y] = M_{k^*}[y], \quad y \in \mathbf{Y},$$

$\pi_T = T_{k^*}\pi$ and $\mathbf{H}_T = (1_n, \{M_T[v], v \in \mathbf{Y}\}, \pi_T)$. The model $\hat{\mathbf{H}}_T = \mathcal{C}_{n,\hat{n}}[\mathbf{H}_T]$ is equivalent to $\mathbf{H}$. By virtue of (14) and (15) it holds with probability 1 that

$$\forall L \in \mathcal{L}^0_{n,\hat{n}}, \quad L\hat{\pi}_T \notin \mathcal{M}_{\mathbf{H}}.$$

By virtue of (16) it holds with probability 1 that

$$\forall L \in \mathcal{L}^c_{n,\hat{n}}, \quad L\hat{\pi}_T \notin \mathcal{M}_{\mathbf{H}}.$$

Thus with probability 1 $\forall \phi \in \mathcal{S}_{\hat{n},n}$ and $D_\phi \in \mathcal{D}_{n,\hat{n}}$ one has

$$\mathcal{A}_{\phi,D_\phi}[\hat{\mathbf{H}}] \notin \mathcal{E}^R_{\mathbf{H}}.$$

.

B. Applying the balanced truncation type reduction algorithm of [8] to the example of theorem 4.1

The balanced truncation type of model reduction method for HMM's developed in [8] will be now applied to the example of theorem 4.1. The algorithm starts out with $\hat{\mathbf{H}} = (1_{\hat{n}}, \{\hat{M}[y],\ y \in \mathbf{Y}\}, \hat{\pi}) \in \mathcal{H}_{\hat{n},\mathbf{Y}}$ and produces $\mathbf{G} = (c, \{A[v], v \in \mathbf{Y}\}, b) \in \mathcal{G}_{n,\mathbf{Y}}$, where $n < \hat{n}$. The individual steps as well as the guarantee of fidelity are summarized below.
Step 1: Compute the *gramian* like matrices.
Let $W_o \in \mathbf{R}^{\hat{n}\times\hat{n}}, W_o \geq 0$ and $W_c \in \mathbf{R}^{\hat{n}\times\hat{n}}, W_c \geq 0$ be the unique solutions to the respective Lyapunov like linear algebraic equations

$$\begin{aligned} W_o &= \sum_{y\in\mathbf{Y}} \hat{M}'[y] W_o \hat{M}[y] + 1_n 1_n', \\ W_c &= \sum_{y\in\mathbf{Y}} \hat{M}[y] W_c \hat{M}'[y] + \pi\pi'. \end{aligned}$$

Let

$$\begin{aligned} W_o &= L_o' L_o, \\ W_c &= L_c L_c' \end{aligned}$$

be the corresponding Cholesky decompositions.

Step 2 : Compute the singular numbers that control the bound to the approximation error.
First perform an eigenvalue decomposition to

$$W_{co} = L_c' W_o L_c.$$

Let

$$\rho_1 \geq \rho_2 \geq \ldots \geq \rho_n > \rho_{n+1} \geq \ldots \rho_{\hat{n}} \geq 0$$

be the ordered, possibly repeated, eigenvalues of W_{co}. Let $\{\psi_1, \ldots, \psi_n\}$ be the corresponding first n normalized eigenvectors of W_{co}, i.e.

$$W_{co}\psi_i = \rho_i \psi_i, \quad |\psi_i|^2 = 1,\ \psi_i'\psi_k = 0,\ i \neq k$$

The relevant singular numbers are given by

$$\sigma_i = \sqrt{\rho_i}, \quad i \in \mathbf{Z}_{\hat{n}}.$$

Let $\{\chi_1, \ldots, \chi_n\}$ be the corresponding first n normalized row eigenvectors of $L_o' W_c L_o$, defined by

$$\chi_i = \frac{1}{\sigma_i}\psi_i' L_c' L_o', \quad i \in \mathbf{Z}_n.$$

Step 3 : Extract the reduced order model.

In order to extract the low order model, one needs to compute a projection operator $U \in \mathbf{R}^{n\times\hat{n}}$ and dilation operator $V \in \mathbf{R}^{\hat{n}\times n}$, where $UV = I_n$. These operators are given by

$$\begin{aligned} V &= L_c[\psi_1 \frac{1}{\sqrt{\sigma_1}}, \ldots, \psi_n \frac{1}{\sqrt{\sigma_n}}], \\ U &= \begin{bmatrix} \frac{1}{\sqrt{\sigma_1}}\chi_1 \\ \vdots \\ \frac{1}{\sqrt{\sigma_n}}\chi_n \end{bmatrix} L_o \end{aligned}$$

The parameters of the reduced order GA $\mathbf{G} = (c, \{A[v], v \in \mathbf{Y}\}, b) \in \mathcal{G}_{n,\mathbf{Y}}$ are given by

$$\begin{aligned} A[y] &= U\hat{M}[y]V, \quad y \in \mathbf{Y}, \\ b &= U\hat{\pi}, \\ c &= 1_{\hat{n}}' V. \end{aligned}$$

In [8] it was proven that the following a priori computable bound to the approximation error holds.

$$\sqrt{\sum_{v\in\mathbf{Y}^*} (q_{\mathbf{G}}[v] - p_{\hat{\mathbf{H}}}[v])^2} \leq 2(\sigma_{n+1} + \ldots + \sigma_{\hat{n}})$$

The above algorithm will applied to the example of theorem 4.1. The algorithm is guaranteed to produce a low order GA of size 2 equivalent to $\mathbf{H}$, i.e. $\mathbf{G} \in \mathcal{E}_{\mathbf{H}}^G$. Given $\hat{\mathbf{H}} = (1_4, \hat{O}, \hat{\Pi}, \hat{\pi}) \in \mathcal{H}_{4,\mathbf{Y}}^R$ one needs first transform it to an equivalent model $\tilde{\mathbf{H}} \in \mathcal{E}_{\hat{\mathbf{H}}}$. To this end let

$$\tilde{M}[y] = \mathrm{diag}[\hat{o}_y]\hat{\Pi}, \quad y \in \mathbf{Y},$$

and $\tilde{\pi} = \hat{\pi}$. The singular numbers that control the error bound between $\tilde{\mathbf{H}}$ and $\mathbf{G}$ are

$$\sigma_1 = 4.6299, \quad \sigma_2 = 0.0117, \quad \sigma_3 = \sigma_4 = 0.$$

The fact that the last 2 singular numbers are zero indicates that exact reduction is possible. The reduced order GA of order 2 $\mathbf{G} = (c, \{A[v], v \in \mathbf{Y}\}, b)$ has the parameters

$$\begin{aligned} A[1] &= \begin{bmatrix} 0.1219 & 0.0343 \\ 0.0343 & 0.1081 \end{bmatrix}, \\ A[2] &= \begin{bmatrix} 0.8773 & -0.0160 \\ -0.0160 & 0.4927 \end{bmatrix}, \\ b &= \begin{bmatrix} -0.9990 \\ -0.0458 \end{bmatrix}, \\ c &= \begin{bmatrix} -0.9990 & -0.0458 \end{bmatrix} \end{aligned}$$

and it holds that

$$\sqrt{\sum_{v\in\mathbf{Y}^*} (q_{\mathbf{G}}[v] - p_{\hat{\mathbf{H}}}[v])^2} \leq 2(\sigma_3 + \sigma_4) = 0.$$

Thus by construction $q_{\mathbf{G}} \equiv p_{\mathbf{H}} \equiv p_{\hat{\mathbf{H}}}$.

V. Conclusion

In this paper we constructed a counterexample to aggregation based model reduction of HMM's. We established the existence of a pair of equivalent HMM's $\hat{\mathbf{H}}$ and $\mathbf{H}$ of different order for which aggregation is guaranteed not to recover an error free low dimensional model. We also established the generic nature of such instances by providing a randomized algorithmic process that produces further such examples with probability 1. Further more we showed how the balanced truncation type algorithm of [8] can be employed to these examples and produce an exact low dimensional model within the class of quasi-realizations.

REFERENCES

[1] B. D. O. Anderson, "The realization problem for hidden Markov models," *Math. Control Signals Syst.*, vol. 12, no. 1, pp. 80–122, Apr. 1999.

[2] J. Berstel and C. Reutenauer, *Rational series and their languages*. Springer-Verlag, 1988.

[3] R. Bhar and S. Hamori, *Hidden Markov Models : applications to financial economics*. Kluwer Academic Publishers, 2004.

[4] K. Deng, G. Mehta, and S. P. Meyn, "Aggregation based model reduction of a hidden Markov model," in *Proc. IEEE Conf. Decision Control*, Atlanta, USA, Dec. 2010.

[5] S. Eilenberg, *Automata, languages, and machines*. Academic Press, 1974.

[6] Y. Ephraim and N. Merhav, "Hidden Markov processes," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1518–1569, Jun. 2002.

[7] E. J. Gilbert, "The identifiability problem for functions of Markov chains," *Ann. Math. Statist.*, vol. 30, pp. 688–697, 1959.

[8] G.Kotsalis, A. Megretski, and M. A. Dahleh, "Balanced truncation for a class of stochastic jump linear systems and model reduction of hidden Markov models," *IEEE Trans. Autom. Control*, vol. 43, no. 11, pp. 2543–2557, Dec. 2008.

[9] T. Koski, *Hidden Markov Models for Bioinformatics*. Kluwer Academic Publishers, 2001.

[10] G. Kotsalis and M. Dahleh, "Model reduction of irreducible Markov chains." in *Proc. IEEE Conf. Decision Control*, Hawai, USA, Dec. 2003, pp. 5727 – 5728.

[11] G. Picci, "On the internal structure of finite state stochastic processes," in *Recent Developments in Variable Structure Systems*, ser. Lecture Notes in Economics and Mathematical Systems. Springer, 1978, vol. 162, pp. 288–304.

[12] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257 – 286, Feb. 1989.

[13] E. D. Sontag, "On certain questions of rationality and decidability," *Journal of Computer and System Science*, vol. 11, pp. 375–381, 1975.

[14] P. Turakainen, "Generalized automata and stochastic languages," *Proc. Amer. Math. Soc.*, vol. 21, no. 2, pp. 303 – 309, May 1969.

[15] M. Vidyasagar, "The realization problem for hidden Markov models: The complete realization problem," in *Proc. IEEE Conf. Decision Control*, Seville, Spain, Dec. 2005, pp. 6632 – 6637.

[16] ——, "The complete realization problem for hidden Markov models. a survey and some new results," working paper, 2009.

[17] ——, "Reduced order modeling of markov and hidden Markov processes via aggregation," in *Proc. IEEE Conf. Decision Control*, Atlanta, USA, Dec. 2010.

[18] L. B. White, R. Mahony, and G. D. Brushe, "Lumpable hidden Markov models - model reduction and reduced complexity filtering," *IEEE Trans. Autom. Control*, vol. 45, no. 12, pp. 2297–2306, Dec. 2000.