Limits of performance for the model reduction problem of Hidden Markov Models

Georgios Kotsalis Jeff S. Shamma

***Abstract*— We introduce system theoretic notions of a Hankel operator, and Hankel norm for hidden Markov models. We show how the related Hankel singular values provide lower bounds on the norm of the difference between a hidden Markov model of order n and any lower order approximant of order $\hat{n} < n$.**

I. Introduction

Hidden Markov Models (HMMs) are one of the most basic and widespread modeling tools for discrete-time stochastic processes, which take values on a finite alphabet. A comprehensive review paper is [3]. Applications of HMMs are met across the spectrum of engineering and science including bio-informatics, econometrics, speech recognition and telecommunications, see for instance [5], [12], [9], [11], [8].

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 [15] it was suggested that the concept of approximate lumpability can be used in the context of model reduction of HMMs. Further work on aggregation based model reduction of HMMs can be found in [13], [2]. In contrast to aggregation based methods, in [6] the authors develop a balanced truncation based model reduction algorithm for HMMs, that is characterized by an a priori computable upper bound to the approximation error and does not suffer from certain limitations of aggregation based model reduction as explained in [7].

The system theoretic view of HMMs established in [6] is continued in the current work where concepts of a Hankel operator and Hankel norm of HMMs are introduced. The related Hankel singular values provide lower bounds on the norm of the difference between a HMM of order n and any lower order approximant of order $\hat{n} < n$. While the upper and lower bound to the approximation error are expressed with regards to different norms they do provide the basis for rigorous model reduction of HMMs with a priori quantifiable certificates of fidelity for the low dimensional approximant.

Georgios Kotsalis is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, gkotsalis3@gatech.edu, Jeff S. Shamma is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, shamma@gatech.edu, and with King Abdullah University of Science and Technology (KAUST), jeff.shamma@kaust.edu.sa. Research was supported by AFOSR/MURI Projects FA9550-09-1-0538 and FA9550-10-1-0573.

The paper is organized as follows. The next section contains preliminary notions including the statistical description of a HMM and the class of linear systems evolving on homogeneous trees, which will provide the basis for the construction of the input-output Hankel operator of HMMs. A stability concept for the latter class of systems is developed in section III. Section IV contains the derivation of the lower bound with respect to the Hankel norm. We conclude with an example and a summary with considerations for future work.

II. Preliminaries

A. Notation

The set of integers is denoted by $\mathbb{Z}$, the set of positive integers by $\mathbb{Z}_+$, and the set of real numbers by $\mathbb{R}$. For $n \in \mathbb{Z}_+$, $\mathbb{R}^n$ denotes the Euclidean n-space. The transpose of a column vector $x \in \mathbb{R}^n$ is x^{T}. For $x \in \mathbb{R}^n$, $|x|^2 = x^{\mathrm{T}}x$ denotes the square of the Euclidean norm. For $A \in \mathbb{R}^{n\times n}$, $\|A\|_2 = \sup_{|x|=1} |Ax|$. The identity matrix in $\mathbb{R}^{n\times n}$ is written as I_n. Let $\mathbb{V}, \mathbb{U}$ be Banach spaces. The Banach space of all bounded linear operators from $\mathbb{V}$ into $\mathbb{U}$ is denoted by $\mathbf{B}(\mathbb{V}, \mathbb{U})$ and the associated induced norm is denoted by $\|\cdot\|$. When $\mathbb{V} = \mathbb{U}$ we write $\mathbf{B}(\mathbb{V})$ instead of $\mathbf{B}(\mathbb{V}, \mathbb{V})$. For a Hilbert space $\mathbb{V}$ the inner product is denoted by $\langle\cdot;\cdot\rangle$. The adjoint of the operator $\mathcal{L} \in \mathbf{B}(\mathbb{V})$ is denoted by $\mathcal{L}^*$. Let $\mathbb{S}^n = \{A \in \mathbb{R}^{n\times n} \mid A = A^{\mathrm{T}}\}$. For $P \in \mathbb{S}^n$, $P \succ 0$, $(P \succeq 0)$ indicates that it is a positive (semi-) definite matrix. Let $\mathbb{S}^n_+ = \{A \in \mathbb{S}^n \mid A \succeq 0\}$ and $\mathbb{S}^n_{++} = \{A \in \mathbb{S}^n \mid A \succ 0\}$. For $P \in \mathbb{S}^n_+$ the notation $|x|^2_P$ stands for $x^{\mathrm{T}}Px$, and $P^{\frac{1}{2}}$ is the positive semi-definite square root of P.

B. Alphabet, strings, language

Let $k \in \mathbb{Z}_+$ and consider the strictly ordered, finite set $\mathbb{A} = \{a_1, \ldots, a_k\}$. The set $\mathbb{A}$ will be referred to as the alphabet and its elements as letters. We denote by $\mathbb{A}^*$ the set of all finite sequences of elements of $\mathbb{A}$, including the empty sequence, denoted by $\varnothing$. The set $\mathbb{A}^*$ is called the language. The finite sequences of letters are called words or strings. Let $\wedge$ denote the concatenation operation, i.e. $\wedge : \mathbb{A}^* \times \mathbb{A}^* \to \mathbb{A}^*$. Words are read from right to left, i.e. in the word $w = w_r \wedge \ldots \wedge w_1$, where $w_1, \ldots, w_r \in \mathbb{A}$ the letter w_1 precedes w_2, etc.. The length of the word w is denoted by $|w|$. The set $\mathbb{A}^*$ equipped with the aforementioned concatenation operation, forms a semi-group. The empty word $\varnothing$ is the identity element. The set of words of length r is $\mathbb{A}_{(r)} = \{w \in \mathbb{A}^* \mid |w| = r\}$. By convention $\mathbb{A}_{(0)} = \{\varnothing\}$.

In this notation the language can be expressed as

$$\mathbb{A}^* = \bigcup_{r=0}^{\infty} \mathbb{A}_{(r)}.$$

One can think of the elements of $\mathbb{A}^*$ as nodes of an infinite acyclic graph, i.e. a homogeneous tree, rooted at $\varnothing$. Each node of the graph is labelled by a word in the language. Each level of the graph, consists of words of same length. Let $\mathbb{A}^*_{\rightarrow}$ correspond to a first, right to left, lexical ordering of $\mathbb{A}^*$ and $\mathbb{A}^*_{\leftarrow}$ correspond to a last, left to right, lexical ordering of $\mathbb{A}^* - \{\varnothing\}$. The ordered sets $\mathbb{A}^*_{\rightarrow}$, $\mathbb{A}^*_{\leftarrow}$, are used to distinguish between the evolution of the system starting at the present towards the future and starting at the distant past towards the present. To make this distinction even more pronounced we will append the symbol $\varnothing$ to all words in $\mathbb{A}^*_{\rightarrow}$ except $\varnothing$ as a suffix and to all words in $\mathbb{A}^*_{\leftarrow}$ as a prefix. For instance when $\mathbb{A} = \{0, 1\}$ one has $\mathbb{A}^*_{\rightarrow} = \{\varnothing, 0\varnothing, 1\varnothing, 00\varnothing, 10\varnothing, 01\varnothing, 11\varnothing, \ldots\}$ and $\mathbb{A}^*_{\leftarrow} = \{\varnothing 0, \varnothing 1, \varnothing 00, \varnothing 01, \varnothing 10, \varnothing 11, \ldots\}$.

C. Statistical description of a HMM

Hidden Markov Models can be defined in many equivalent ways. The basic definitions and notation introduced in the context of realization theory of HMMs will be used. One can find them for instance in slightly varying language in [10], [1], [14]. Let $\{Y(t)\}$ be a discrete-time, stationary stochastic process over some fixed probability space $\{\Omega, \mathcal{M}, \mathbb{P}\}$, with taking values on some alphabet $\mathbb{A} = \{a_1, \ldots, a_k\}$. The strict future of the process after time t is denoted by $Y_t^+ = \{\ldots, Y(t+2), Y(t+1)\}$ and $Y_t^- = \{Y(t), Y(t-1), \ldots\}$ denotes its past and present. Let $v = v_k \ldots v_1 \in \mathbb{A}^*$ the notation $\{Y_t^+ \equiv v\}$ stands for the event $\{\omega \in \Omega \mid Y(t+k) = v_k, \ldots, Y(t+1) = v_1\}$, by convention $\{Y_t^+ \equiv \varnothing\} = \Omega$.

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

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

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:

$$\begin{aligned} p[\emptyset] &= 1 \\ p[v] &\in [0, 1],\ \forall v \in \mathbb{A}^*, \\ p[v] &= \sum_{u \in \mathbb{A}^r} p[vu],\ \forall v \in \mathbb{A}^*, r \in \mathbb{Z}_+. \end{aligned}$$

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

$$\Pr[Y_t^+ \equiv v] = \Pr[\tilde{Y}_t^+ \equiv v]. \tag{1}$$

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 $\{\Omega, \mathcal{M}, \mathbb{P}\}$. In the context of this work when referring to a stationary stochastic process over the alphabet $\mathbb{A}$, one is thinking of an equivalence class of processes in the sense of (1). 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 $\mathbb{A}$ has a realization as a stationary **HMM** of size $n \in \mathbb{Z}_+$ if there exists a pair of discrete-time, stationary stochastic processes $\{X(t)\}$, $\{\tilde{Y}(t)\}$ taking values on the finite sets $\mathbb{X} = \{1, \ldots, n\}$ and $\mathbb{A}$ respectively, such that $\{Y(t)\}$ and $\{\tilde{Y}(t)\}$ are equivalent, the joint process $\{X(t), \tilde{Y}(t)\}$ is a Markov process and $\forall \sigma \in \mathbb{X}^*$, $\forall v \in \mathbb{A}^*$, the following "splitting property" holds

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

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 $M : \mathbb{A} \rightarrow \mathbb{R}_+^{n \times n}$ where

$$M[v]_{ij} = \Pr[X(t+1) = i, \tilde{Y}(t+1) = v | X(t) = j],$$

$i, j \in \mathbb{X}, v \in \mathbb{A}, t \in \mathbb{Z}$. The state transition matrix of the underlying Markov process $\{X(t)\}$ is given by

$$\Pi = \sum_{v \in \mathbb{A}} M[v].$$

Let $\pi \in \mathbb{R}_+^n$, such that $\Pi\ \pi = \pi,\ 1_n^{\mathrm{T}}\ \pi = 1$. 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

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

Lemma 2.1: Let $r \in \mathbb{Z}_+,\ v = v_r v_{r-1} \ldots v_1 \in \mathbb{A}_{(r)}$. The probability of that particular string can be computed recursively according to

$$p[v] = 1_n^{\mathrm{T}}\ M[v_r]\ M[v_{r-1}]\ \ldots\ M[v_1]\ \pi,$$

Proof: See for instance [1], [14]. ∎

The preceding lemma shows that if a given stationary process $\{Y(t)\}$ over the alphabet $\mathbb{A}$ has a realization as a stationary HMM of size n, then its probability function is completely encoded by the ordered triple

$$H = (M, \pi, 1_n^{\mathrm{T}}).$$

Accordingly in the following discussion referring to a HMM of size n will be in terms of the ordered triple $H = (M, \pi, 1_n^{\mathrm{T}})$ that contains its statistical parameters. The set of all HMMs of size n over the alphabet $\mathbb{A}$ is denoted by $\mathcal{H}_{n,\mathbb{A}}$.

D. Linear systems on homogeneous trees

We will consider linear systems with evolution on $\mathbb{A}^*$. This class of systems was introduced in [4], were various system theoretic properties where developed, including the notion of a Hankel operator. For the purposes of this work it suffices to focus on a particular subclass of the structured noncommutative multidimensional systems of [4], namely the noncommutative Fronasini-Marchesini systems. We will simply refer to them as a linear system with evolution on $\mathbb{A}^*$. In the following fix $n, m, p \in \mathbb{Z}_+$ and consider the signal spaces $\mathcal{X} = \{x \mid x : \mathbb{A}^* \to \mathbb{R}^n\}$, $\mathcal{U} = \{u \mid u : \mathbb{A}^* \to \mathbb{R}^m\}$, $\mathcal{Y} = \{y \mid y : \mathbb{A}^* \to \mathbb{R}^p\}$. A linear system Σ of order n evolving on $\mathbb{A}^*$ is defined by

$$\begin{aligned} x(i \wedge w) &= A(i)\ x(w) \ + \ B(i)\ u(w), \\ y(w) &= C\ x(w) \quad \text{for} \quad w \in \mathbb{A}^*, i \in \mathbb{A}. \end{aligned} \tag{2}$$

The state, input and output signals are $x \in \mathcal{X}, u \in \mathcal{U}, y \in \mathcal{Y}$ respectively. The parameters of the state space representation have appropriate dimensions, $A : \mathbb{A} \to \mathbb{R}^{n \times n}$, $B : \mathbb{A} \to \mathbb{R}^{n \times m}$, $C \in \mathbb{R}^{p \times n}$. If an initial condition $x(\varnothing) = x_\varnothing \in \mathbb{R}^n$ is specified then the output $y \in \mathcal{Y}$ is uniquely determined by the input $u \in \mathcal{U}$. The state space system Σ in (2) will sometimes be denoted solely by its parameters,

$$\Sigma = (A, B, C).$$

We will refer by $\Sigma_\rightarrow$ to the state space description (2) when $w \in \mathbb{A}^*_\rightarrow$. When $w \in \mathbb{A}^*_\leftarrow$ the associated state space description denoted by $\Sigma_\leftarrow$ is given by:

$$\begin{aligned} x(w) &= \sum_{i \in \mathbb{A}} [A(i)\ x(w \wedge i) \ + \ B(i)\ u(w \wedge i)], \\ y(w) &= C\ x(w) \quad \text{for} \quad w \in \mathbb{A}^*_\leftarrow, i \in \mathbb{A}. \end{aligned} \tag{3}$$

The set of all linear systems on $\mathbb{A}^*$ of order n is denoted by $\mathcal{S}_{n,\mathbb{A}}$. One can establish an immersion of $\mathcal{H}_\mathbb{A} = \bigcup_{n=2}^\infty \mathcal{H}_{n,\mathbb{A}}$ into $\mathcal{S}_\mathbb{A} = \bigcup_{n=2}^\infty \mathcal{S}_{n,\mathbb{A}}$ in the obvious way. Let

$$f : \mathcal{H}_\mathbb{A} \to \mathcal{S}_\mathbb{A},$$

where $f(H) = \Sigma = (M, \pi, 1_n^{\mathrm{T}})$. This rather innocuous step provides leverage since various system theoretic properties associated with linear systems on homogeneous trees can be transferred to HMMs. Furthermore consider $D : \mathcal{S}_\mathbb{A} \times \mathcal{S}_\mathbb{A} \to \mathbb{R}_+$ be a metric on the set of linear systems evolving on $\mathbb{A}^*$ then one obtains automatically a metric on HMMs by restricting the domain onto $f(\mathcal{H}_\mathbb{A}) \times f(\mathcal{H}_\mathbb{A})$. On the grounds of this observation the subsequent discussion in regards to linear systems evolving on $\mathbb{A}^*$ pertains to HMMs over the alphabet $\mathbb{A}$ as well.

III. Stability and signal spaces

For $A \in \mathbb{R}^{n \times n}$, the spectrum of A is defined as

$$\sigma(A) = \{\lambda \in \mathbb{C} \mid \mathcal{N}(\lambda I_n - A) \neq \varnothing\},$$

and the spectral radius of A is

$$r_\sigma[A] = \max\{|\lambda| \in \mathbb{R}_+ \mid \lambda \in \sigma(A)\}.$$

It follows from the definition that $r_\sigma[A] \leq \|A\|$ and if $A \in \mathbb{S}^n$, then $r_\sigma[A] = \|A\|_2$. Let $\mathbb{T}^n = \mathbf{B}[\mathbb{S}^n]$. For $\mathcal{F} \in \mathbb{T}^n$,

$$\|\mathcal{F}\| = \sup_{\|X\|=1} \|\mathcal{F}[X]\|.$$

The linear map $\mathcal{F} \in \mathbb{T}^n$ is called monotonic if for all $X \in \mathbb{S}^n$, $X \succeq 0 \Rightarrow \mathcal{F}[X] \succeq 0$.

Lemma 3.1: Consider $\mathcal{F} \in \mathbb{T}^n$, monotonic. For $i \in \mathbb{Z}_+$, $\mathcal{F}^i$ is monotonic and for any $X \in \mathbb{S}^n_+$, $\mathcal{F}^i(X) \succeq 0$.

Proof: The proof is immediate by induction. ■

Lemma 3.2: Consider $\mathcal{F} \in \mathbb{T}^n$, monotonic. The following statements are equivalent:

- (i) $r_\sigma[\mathcal{F}] < 1$,
- (ii) $\mathcal{F}^i[X] \to 0$, for all $X \in \mathbb{S}^n_+$,
- (iii) $\mathcal{F}^i(I_n) \to 0$.

Proof: (i) $\Rightarrow$ (ii). This follows directly by considering the Jordan decomposition of the of the finite dimensional linear map $\mathcal{F}$.
(ii) $\Rightarrow$ (iii). This is immediate given that $I_n \in \mathbb{S}^n_+$.
(iii) $\Rightarrow$ (i). Let $i \in \mathbb{Z}_+$, one has:

$$r_\sigma[\mathcal{F}]^i = r_\sigma[\mathcal{F}^i] \leq \|\mathcal{F}^i\|_2 = \sup_{\|X\|_2=1} r_\sigma[\mathcal{F}^i[X]].$$

For any $X \in \mathbb{S}^n$ with $\|X\|_2 = 1$ it holds that $I_n - X \succeq 0$. Given that $\mathcal{F}^i$ is linear and monotonic it follows that $\mathcal{F}^i[I_n] \succeq \mathcal{F}^i[X]$ and therefore $\|\mathcal{F}^i[I_n]\|_2 \geq \|\mathcal{F}^i[X]\|_2$. Since $\|I_n\|_2 = 1$ one obtains that $\|\mathcal{F}^i[I_n]\|_2 = \sup_{\|X\|_2=1} r_\sigma[\mathcal{F}^i[X]]$, and therefore $r_\sigma[\mathcal{F}]^i \leq \|\mathcal{F}^i[I_n]\|_2$, and the result follows. ■

Lemma 3.3: Let $F \in \mathbb{T}^n$, monotonic, $B \in \mathbb{S}^n_+$, and consider the equation

$$X = \mathcal{F}[X] + B.$$

- If $r_\sigma[F] < 1 \Rightarrow$ there exists a solution $X \succeq 0$.
- If there exists a solution $X \succeq 0$ and $B \succ 0$ then

$$r_\sigma[\mathcal{F}] < 1$$

and in fact $X \succ 0$.

Proof: For $k \in \mathbb{Z}_+$ let $X_k = \sum_{i=0}^k \mathcal{F}^i[B]$. Gelfand's formula states

$$\|\mathcal{F}^i\|^{\frac{1}{i}} \to r_\sigma[\mathcal{F}].$$

This allows one to conclude that there exists $\lambda \in (r_\sigma[F], 1)$ and $N \in \mathbb{Z}_+$ such that if $i > N$, $\|F^i\|_2 \leq \lambda^i$. Since

$$\|X_k\|_2 \leq \sum_{i=0}^k \|\mathcal{F}^i[B]\|_2 \leq \sum_{i=0}^k \|\mathcal{F}^i\|_2 \|B\|_2,$$

one can infer that the sequence of partial sums converges, say $X_k \to X$. Furthermore $X - \mathcal{F}[X] = B$ and $X \succeq B \succeq 0$.

For the second part any solution X satisfies the relations

$$X = \sum_{i=0}^{k-1} \mathcal{F}^i[B] + \mathcal{F}^k[X] \succeq \sum_{i=0}^{k-1} \mathcal{F}^i[B] \succeq B \succ 0.$$

The sequence of partial sums $X_k = \sum_{i=0}^{k} \mathcal{F}^i[B]$ therefore converges and furthermore $\mathcal{F}^i[B] \to 0$. By the previous lemma it follows that $r_\sigma[\mathcal{F}] < 1$. ∎

In the following we discuss a stability notion for a system evolving on an acyclic graph. We will consider the uncontrolled evolution of the state space recursion.

$$x(i \wedge w) = A(i)\ x(w), \quad w \in \mathbb{A}^*, i \in \mathbb{A}.$$

System Σ is said to be asymptotically stable if for $u(w) = 0,\ \forall w \in \mathbb{A}^*$, and $x(\varnothing) \in \mathbb{R}^n$,

$$\lim_{k\to\infty} \sum_{w\in\mathbb{A}_{(k)}} |x(w)|^2 = 0.$$

Consider $\mathcal{T} \in \mathbb{T}^n$, where $\mathcal{T}[X] = \sum_{i\in\mathbb{A}} A(i)^{\mathrm{T}}\ X\ A(i)$. For $k \in \mathbb{Z}_+$, $\mathcal{T}^k$ is linear and monotonic.

Lemma 3.4: Let R in $\mathbb{S}^n$ and consider the uncontrolled evolution of the state space recursion. One has

$$\sum_{w\in\mathbb{A}_{(k)}} x^{\mathrm{T}}(w)\ R\ x(w) = x[\varnothing]^{\mathrm{T}}\ \mathcal{T}^k[R]\ x[\varnothing].$$

Proof: The statement follows by induction. ∎

Lemma 3.5: A given system Σ is stable if and only if $r_\sigma[\mathcal{T}] < 1$.

Proof: The proof of the statement follows a series of equivalences. Consider the uncontrolled evolution of Σ.

$$\forall x(\varnothing) \sum_{w\in\mathbb{A}_{(k)}} |x(w)|^2 \to 0 \Leftrightarrow \mathcal{T}^k[I_n] \to 0 \Leftrightarrow r_\sigma[\mathcal{T}] < 1.$$

∎

For $p \in \mathbb{Z}_+$ we define

$$\begin{aligned} l_{2,p}^{\rightarrow} &= \{f : \mathbb{A}_{\rightarrow}^* \to \mathbb{R}^p \mid \sum_{w\in\mathbb{A}_+^*} |f(w)|^2 < \infty\}, \\ l_{2,p}^{\leftarrow} &= \{f : \mathbb{A}_{\leftarrow}^* \to \mathbb{R}^p \mid \sum_{w\in\mathbb{A}_-^*} |f(w)|^2 < \infty\} \end{aligned}$$

and set $l_{2,p} = l_{2,p}^{\rightarrow} \cup l_{2,p}^{\leftarrow}$. For $x, y \in l_{2,p}^{\rightarrow}$, their inner product is given by

$$\langle x, y\rangle = \sum_{w\in\mathbb{A}_{\rightarrow}^*} x^{\mathrm{T}}(w) y(w),$$

and similarly for square summable signals on $\mathbb{A}_{\leftarrow}^*$.

IV. A lower bound to the approximation error with respect to the Hankel norm

A. Controllability, observability

First we introduce an observability and a controllability concept for a stable system Σ. Consider the uncontrolled evolution of Σ. Let

$$\Psi_o : \mathbb{R}^n \to l_{2,p}^{\rightarrow}$$

denote the observability operator, where $\Psi_o(x) = y$ with $y(w), w \in \mathbb{A}_{\rightarrow}^*$ being the output of system Σ under the initial condition $x(\varnothing) = x$ and $u(w) = 0, \forall w \in \mathbb{A}_{\rightarrow}^*$. It is notationally convenient to extend the map A homomorphically from $\mathbb{A}$ to $\mathbb{A}^*$, i.e. for $k \in \mathbb{Z}_+$,

$$A(w_k \wedge \ldots \wedge w_1) = A(w_k) \ldots A(w_1),$$

and $A(\varnothing) = I_n$. Given this notation and taking the lexicographic ordering of $\mathbb{A}_{\rightarrow}^*$ in mind one can write the relationship $\Psi_o(x) = y$ in matrix form,

$$\begin{bmatrix} y(\varnothing) \\ y(1\varnothing) \\ \vdots \\ y(w) \\ \vdots \end{bmatrix} = \begin{bmatrix} C \\ CA(1) \\ \vdots \\ CA(w) \\ \vdots \end{bmatrix} x(\varnothing) = \mathcal{O}x(\varnothing).$$

The adjoint of the observability operator $\Psi_o^* : l_{2,p}^{\rightarrow} \to \mathbb{R}^n$ is defined through the relationship

$$\langle \Psi_o^*(z), x\rangle = \langle z, \Psi_o(x)\rangle,$$

and as such $\Psi_o^*(z) = \sum_{w\in\mathbb{A}^*} A(w)^{\mathrm{T}} C^{\mathrm{T}} z(w)$. We refer to $Y_o = \Psi_o^*\ \Psi_o$ as the observability grammian of the system Σ, and the system Σ will be called observable if $Y_o \succ 0$. The observability grammian satisfies the relationship $Y_o = \mathcal{T}^i[Y_o] + C^{\mathrm{T}}C$, and furthermore $Y_o = \sum_{i=0}^{\infty} \mathcal{T}^i[C^{\mathrm{T}}C]$. Let

$$\Psi_c : l_{2,m}^{\leftarrow} \to \mathbb{R}^n,$$

denote the controllability operator where $u \in l_{2,m}^{\leftarrow}$ is applied as an input to $\Sigma_{\leftarrow}$ and produces

$$x = \Psi_c(u) = \sum_{i\in\mathbb{A}} \sum_{w\in\mathbb{A}^*} A(w)B(i)u(w \wedge i).$$

Taking the lexicographic ordering of $\mathbb{A}_{\leftarrow}^*$ into account one can write

$$x = \mathcal{C} \begin{bmatrix} u(\varnothing 1) \\ \vdots \\ u(\varnothing r) \\ u(\varnothing 11) \\ \vdots \\ u(w \wedge i) \\ \vdots \end{bmatrix},$$

where

$$\mathcal{C} = \begin{bmatrix} B(1), \ldots, B(r), A(1)B(1), \ldots, A(w)B(i), & \ldots \end{bmatrix}.$$

The adjoint of the controllability operator $\Psi_c^* : \mathbb{R}^n \to l_{2,m}^{\leftarrow}$ is defined through the relationship

$$\langle \Psi_c^*(z), x\rangle = \langle z, \Psi_c(x)\rangle,$$

and as such $\Psi_c^*(z) = (\ldots, u(w \wedge i), \ldots)$ with

$$u(w \wedge i) = B(i)^{\mathrm{T}} A(w)^{\mathrm{T}} z, \quad w \in \mathbb{A}^*, i \in \mathbb{A}.$$

We refer to $X_c = \Psi_c\ \Psi_c^*$, as the controllability gramian of the system Σ, and the system Σ will be called controllable if $X_c \succ 0$. The controllability gramian satisfies the relationship $\mathcal{L}[X_c] + \sum_{i\in\mathbb{A}} B(i)B(i)^{\mathrm{T}} = X_c$ and furthermore $X_c = \sum_{i\in\mathbb{A}} \sum_{k=0}^{\infty} \mathcal{L}^k[B(i)B(i)^{\mathrm{T}}]$, where $\mathcal{L} = \mathcal{T}^*$.

B. Hankel Operator

Consider a stable system Σ that is both controllable and observable. The Hankel singular values of Σ are the square roots of the eigenvalues of X_cY_o, i.e.

$$i\text{th Hankel Singular Value } = \sigma_i = \sqrt{\lambda_i(X_cY_o)}$$

It will be assumed that the σ_i's are ordered such that $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n$.

Since the eigenvalues of X_cY_o are non-negative the Hankel singular values are well defined. They are also invariant under coordinate transformations of the particular realization of the system Σ. Let $G_\Sigma : l_{2,m} \to l_{2,p}$ be an operator that maps input sequences to output sequences $y = G_\Sigma u$ where y, u are constrained by the state space model $\Sigma_\leftarrow$, $\Sigma_\rightarrow$. Let $\Pi_{l_{2,p}^\rightarrow}$ denote the projection of a signal in $l_{2,p}$ onto $l_{2,p}^\rightarrow$. The Hankel operator associated with Σ, denoted by Γ_Σ is defined as

$$\Gamma_\Sigma : l_{2,m}^\leftarrow \to l_{2,p}^\rightarrow,$$

where $\Gamma_\Sigma(u) = \Pi_{l_{2,p}^\rightarrow} G_\Sigma u$. It follows from the definition that if Σ_1 and Σ_2 are two systems then $\Gamma_{\Sigma_1+\Sigma_2} = \Gamma_{\Sigma_1} + \Gamma_{\Sigma_2}$. The induced norm of the Hankel operator is defined as

$$\|\Gamma_\Sigma\| = \sup_{u \in l_{2,m}^\leftarrow, u \neq 0} \frac{\|\Gamma_\Sigma u\|_2}{\|u\|_2}.$$

This quantity will be referred to as the Hankel norm of Σ and denoted by $\|\Sigma\|_H$. The Hankel operator admits a decomposition in terms of the observability and controllability operator. Let $y = \Gamma_\Sigma u$, for $w \in \mathbb{A}_\rightarrow^*$ one has

$$y(w) = CA(w) \sum_{i \in \mathbb{A}} \sum_{v \in \mathbb{A}_\leftarrow^*} A(v)B(i)u(v \wedge i) = \Psi_o \ \Psi_c \ u.$$

Expressing the above relation in matrix form gives

$$\begin{bmatrix} y(\varnothing) \\ y(1\varnothing) \\ \vdots \\ y(w) \\ \vdots \end{bmatrix} = \mathcal{O} \ \mathcal{C} \begin{bmatrix} u(\varnothing 1) \\ \vdots \\ u(\varnothing r) \\ u(\varnothing 11) \\ \vdots \\ u(v \wedge i) \\ \vdots \end{bmatrix}$$

The above decomposition shows that despite the Hankel operator being a map between two infinite dimensional spaces, it is a finite dimensional one, since the dimension of its range cannot exceed n. In fact the dimension equals n when the realization is both controllable and observable.

Theorem 4.1: The Hankel singular values of Σ are the square roots of the non-zero eigenvalues of $\Gamma_\Sigma^* \ \Gamma_\Sigma$.

Proof: From the decomposition of Γ_Σ one has

$$\Gamma_\Sigma^* \ \Gamma_\Sigma = \Psi_c^* \ \Psi_o^* \ \Psi_o \ \Psi_c.$$

Let x_i be a right eigenvector of X_cY_o associated with σ_i, i.e. $X_cY_ox_i = \sigma_i^2 x_i$. Define the vector y_i as

$$y_i = \frac{1}{\sigma_i} Y_o x_i.$$

It follows that x_i, y_i satisfy $X_c y_i = \sigma_i x_i$, $Y_o x_i = \sigma_i y_i$. Furthermore $\langle y_j, x_i \rangle = 0$ if $i \neq j$. The x_i's can be normalized such that $\langle y_i, x_i \rangle = \frac{1}{\sigma_i}$. Now define the functions

$$\begin{aligned} \chi_j &= \Psi_o x_j = (\ldots CA(w)x_j \ldots)^{\mathrm{T}} \quad w \in \mathbb{A}_\rightarrow^* \\ \xi_j &= \Psi_c^* y_j = (\ldots B(i)^{\mathrm{T}} A(w)^{\mathrm{T}} y_j \ldots)^{\mathrm{T}} \quad w \in \mathbb{A}_\leftarrow^*, i \in \mathbb{A}. \end{aligned}$$

From the definition the collections $\{\chi_j\}$ and $\{\xi_j\}$ form orthonormal sets respectively. The functions χ_j, ξ_j are called the Schmidt pair associated with σ_j. By direct substitution one has

$$\Gamma_\Sigma \xi_i = \sigma_i \chi_i, \quad \Gamma_\Sigma^* \chi_i = \sigma_i \xi_i,$$

which implies that

$$\Gamma_\Sigma^* \Gamma_\Sigma \xi_i = \sigma_i^2 \xi_i, \quad \Gamma_\Sigma \Gamma_\Sigma^* \chi_i = \sigma_i^2 \chi_i.$$

This shows that every eigenvalue of X_cY_o is an eigenvalue of $\Gamma_\Sigma^*\Gamma_\Sigma$ and vice versa. ■

As a consequence of this result we obtain a singular value decomposition of Γ_Σ and compute its norm.

Lemma 4.1: The Hankel operator of system Σ can be written as

$$\Gamma_\Sigma u = \sum_{i=1}^{n} \sigma_i \langle \xi_i, u \rangle \chi_i.$$

Proof: Let $\{\xi_{i+n}, i \geq 1\}$ be a set of signals in $l_{2,m}^\leftarrow$ such that $\{\xi_i, i \geq 1\}$ forms an orthonormal basis for $l_{2,m}^\leftarrow$. The dimension of $\mathcal{R}(\Gamma_\Sigma^*\Gamma_\Sigma)$ is equal to n. From this, it follows that

$$\mathcal{R}(\Gamma_\Sigma^*\Gamma_\Sigma) = \text{Span}\{\xi_i, i = 1, \ldots, n\}.$$

Furthermore for any $u \in l_{2,m}^\leftarrow$

$$\langle u, \Gamma_\Sigma^*\Gamma_\Sigma \xi_{i+n} \rangle = \langle \Gamma_\Sigma^*\Gamma_\Sigma u, \xi_{i+n} \rangle = 0, \quad \forall i \geq 1.$$

The last equality holds because $\{\xi_{i+n}, i \geq 1\}$ is orthogonal to $\mathcal{R}(\Gamma_\Sigma^*\Gamma_\Sigma)$. As such $\Gamma_\Sigma \xi_{i+n} = 0$ for all $i \geq 1$. Then

$$\|\Gamma_\Sigma \xi_{i+n}\|_2^2 = \langle \Gamma_\Sigma \xi_{i+n}, \Gamma_\Sigma \xi_{i+n} \rangle = \langle \Gamma_\Sigma^*\Gamma_\Sigma \xi_{i+n}, \xi_{i+n} \rangle = 0,$$

$\forall i \geq 1$. Given any $u \in l_{2,m}^\leftarrow$, u can be written as

$$u = \sum_{i=1}^{\infty} \langle \xi_i, u \rangle \xi_i.$$

Then, it follows that

$$\Gamma_\Sigma u = \Gamma_\Sigma \sum_{i=1}^{\infty} \langle \xi_i, u \rangle \xi_i = \sum_{i=1}^{n} \langle \xi_i, u \rangle \Gamma_\Sigma \xi_i = \sum_{i=1}^{n} \langle \xi_i, u \rangle \sigma_i \chi_i$$

■

From the above calculation it follows that

$$\|\Gamma_\Sigma\| = \|\Sigma\|_H = \sigma_1.$$

Theorem 4.2: Let Σ be a stable system of order n that is controllable and observable. Let $\hat{\Sigma}_k$ be any controllable and observable system of order $k < n$. Then:

$$\|\Sigma - \hat{\Sigma}_k\|_H \geq \sigma_{k+1}.$$

Proof: The rank of $\Gamma_{\hat{\Sigma}_k}$ is k. Consider the subspace

$$M = \{f \mid f = \sum_{i=1}^{k+1} \alpha_i \xi_i, \quad \text{for} \quad \alpha_i \in \mathbb{C}\}.$$

The dimension of this subspace is $k+1$. This implies that there is an element $f \in M$ such that $\Gamma_{\hat{\Sigma}_k} f = 0$. Normalize f such that

$$f = \sum_{i=1}^{k+1} \alpha_i \xi_i, \qquad \sum_{i=1}^{k+1} |\alpha_i|^2 = 1.$$

It follows that

$$\begin{aligned} \|\Sigma - \hat{\Sigma}_k\|_H &\geq \|(\Gamma_\Sigma - \Gamma_{\hat{\Sigma}_k}) f\|_2 \geq \|\Gamma_\Sigma f\|_2 \\ &\geq \|\sum_{i=1}^{k+1} \alpha_i \chi_i\|_2 = (\sum_{i=1}^{k+1} |\alpha_i|^2 \sigma_i^2)^{\frac{1}{2}} \\ &\geq \sigma_{k+1}. \end{aligned}$$

■

V. Example

We will consider an example of a HMM where $\mathbb{X} = \{x_1, x_2, x_3\}$, $\mathbb{Y} = \{y_1, y_2\}$. The underlying Markov chain $\{X(t)\}$ has statistical parameters

$$\Pi = \begin{bmatrix} 0.50 & 0.25 & 0.25 \\ 0.25 & 0.50 & 0.25 \\ 0.25 & 0.25 & 0.50 \end{bmatrix}, \quad \pi = \frac{1}{3} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

The output process $\{Y(t)\}$ is a deterministic function of the state $\{X(t)\}$, i.e. $Y(t) = f(X(t))$, with $f(x_1) = f(x_2) = y_1$ and $f(x_3) = y_3$. Under this condition one has

$$M[y_1] = \begin{bmatrix} 0.50 & 0.25 & 0.25 \\ 0.25 & 0.50 & 0.25 \\ 0 & 0 & 0 \end{bmatrix},$$

$$M[y_2] = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0.25 & 0.25 & 0.50 \end{bmatrix}.$$

For this example ones has

$$Y_0 = \begin{bmatrix} 2.60 & 2.60 & 2.25 \\ 2.60 & 2.60 & 2.25 \\ 2.25 & 2.25 & 2.20 \end{bmatrix},$$

and

$$X_c = \begin{bmatrix} \frac{4}{5} & \frac{4}{5} & \frac{2}{9} \\ \frac{4}{5} & \frac{4}{5} & \frac{2}{9} \\ \frac{2}{9} & \frac{2}{9} & \frac{64}{90} \end{bmatrix}.$$

The Hankel singular values are

$$\sigma_1 \approx 3.43, \ \sigma_2 \approx 0.34, \ \sigma_3 = 0.$$

It should come to no surprise that the smallest Hankel singular value is 0. There exists a 2 state HMM that is statistically equivalent to the original model, in fact in can be obtained by aggregation based model reduction, one needs to form a cluster out of the states x_1, x_2. As such in this example our proposed bound is tight.

VI. Summary

We provide an immersion of HMMs into the class of linear systems on homogeneous trees which allows one to transfer various system theoretic properties from the latter class to the former, including input-output properties, such as the notion of a Hankel operator and Hankel norm. This allowed us to make use of the related Hankel singular values to provide lower bounds on the norm of the difference between a HMM of order n and any lower order approximant of order $\hat{n} < n$. Future work will focus on relating the Hankel norm presented in this paper to an l_2 induced gain norm presented in [6] in regards to the balanced truncation algorithm for HMMs, as well as the optimal Hankel norm model reduction problem for this class of systems.

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] 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.

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

[4] G. G. J. A. Ball and T. Malakorn, "Structured noncommutative multidimensional linear systems," *SIAM J. Control Optim.*, vol. 44, pp. 1474–1528, 2005.

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

[6] G. Kotsalis, A. Megretski, and M. Dahleh, "Balanced truncation for a class of stochastic jump linear systems and model reduction for hidden Markov models," *IEEE Trans. on Automatic Control*, vol. 53, pp. 2543–2558, 2008.

[7] G. Kotsalis and J. S. Shamma, "A counterexample to aggregation based model reduction of hidden markov models," in *Proceedings of the 50th IEEE Conference on Decision and Control and European Control Conference, CDC-ECC 2011, Orlando, FL, USA, December 12-15, 2011*, 2011, pp. 6558–6563.

[8] I. L. Macdonald and W. Zucchini, *Hidden Markov and Other Models for Discrete-Valued Time Series*. Chapman and Hall, 1997.

[9] E. M. O. Cappe and T. Ryden, *Inference in Hidden Markov Models*. Springer, 2005.

[10] 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.

[11] L. A. R. J. Elliot and J. B. Moore, *Hidden Markov Models: Estimation and Control*. Springer, 1995.

[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] M. Vidyasagar, "Reduced-order modeling of Markov and Hidden Markov processes via aggregation," in *Proc. IEEE Conf. Decision Control*, Atlanta, USA, Dec. 2010, pp. 1810 – 1815.

[14] ——, "The complete realization problem for hidden markov models: A survey and some new results," *Math. Control Signals Syst.*, vol. 23, no. 1, pp. 1–65, 2011.

[15] 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.