2011 American Control Conference
on O'Farrell Street, San Francisco, CA, USA
June 29 - July 01, 2011

# Modeling of Symbolic Systems: Part II - Hilbert Space Construction for Model Identification and Order Reduction★

Yicheng Wen yxw167@psu.edu
Asok Ray axr2@psu.edu
Ishanu Chattopadhyay ixc128@psu.edu
Shashi Phoha sxp26@psu.edu

The Pennsylvania State University
University Park, PA 16802, USA

*Abstract*— **This paper, which is the second of two parts, is built upon the vector space of symbolic systems represented by probabilistic finite State automata (PFSA) reported in the first part. This second part addresses the Hilbert space construction for model identification, where order reduction is achieved via orthogonal projection. To this end, a family of inner products is constructed and the norm induced by an inner product is interpreted as a measure of information contained in the PFSA, which also quantifies the error due to model order reduction. A numerical example elucidates the process of model order reduction by orthogonal projection from the space of PFSA onto a subspace that belongs to the class of shifts of finite type.**

## I. Introduction

System identification and model order reduction are active areas of research in many fields of science and engineering. For example, in the classical continuous domain, dynamical models of physical processes and human-engineered systems frequently involve a very large number of differential or difference equations. Due to the constraint of computational resources, model reduction is essential for such large-order systems [1]. In this respect, symbolization-based techniques [2] have been developed for probabilistic representation of dynamical systems to compensate for certain inadequacies of classical time-domain and frequency-domain system identification. Probabilistic finite state automata (PFSA) have emerged as a useful mathematical model for identification of uncertain dynamical systems.

The work reported in this paper is built upon the concepts of symbolic dynamics [3] and language theory [4][5], instead of classical continuous-domain modeling [6]. The basic approach is symbolic dynamic filtering (SDF) [7][8] that partitions the (possibly pre-processed) time series or image data observed from the underlying physical process to generate strings of symbols. Then, semantic models are constructed in the symbolic domain as PFSA.

The model reduction techniques for symbolic systems are not well studied as those in the classical continuous domain except for a few cases for Markov chains. Kotsalis and Dahleh [9] reported a reduction scheme for partially observed irreducible Markov chains, based on the $L_1$-metric of the asymptotic observed outputs. This method only applies to nearly completely decomposable Markov chains, which is a very restricted model. Deng et al. [10] proposed an information-theoretic framework to aggregate large-scale Markov chains., where the Kullback-Leibler (K-L) divergence [11] was employed as a metric to measure the distance between two stationary Markov chains; however, the K-L divergence is not a true metric in the mathematical sense. Chattopadhyay and Ray [12] introduced the concept of projective composition for projection of PFSA to an arbitrary structure. Although the projective composition has a nice property of preserving the long-term distribution over the states of the projected model, it is not an orthogonal projection and the projection error is difficult to interpret.

The main contribution of this paper lies in the construction of a family of inner products on the vector space of PFSA [13]. An inner product on the vector space would allow formulation of model order reduction problems by orthogonal projection in a Hilbert space setting, where quantification of the projection error is numerically efficient. This error due to model reduction could also be interpreted in terms of information loss, analogous to the entropy rate [14]. In addition, the proposed analytical approach has the following potential benefits.

1) Development of a mathematical tool to enhance the understanding of stochastic regular languages for solving the associated problems in symbolic dynamics.
2) Establishment of a link between the theories of formal languages and functional analysis.

The technical contents of this second part are built upon the vector space of PFSA, which is reported in the first part [13]. Therefore, it is necessary to refer to the algebraic framework of the vector space formulated in the first part to build the topological framework of an inner product space in this paper.

## II. Preliminaries

Given a (finite) alphabet $\Sigma$ of symbols, the set $\mathscr{B}_\Sigma \triangleq 2^{\Sigma^\star}\Sigma^\omega$ is defined to be the $\sigma$-algebra generated by the set $\{L : L = x\Sigma^\omega \text{ where } x \in \Sigma^\star\}$. Let the space of all positive probability measures on $\mathscr{B}_\Sigma$, which contains a finite number of Nerode equivalence classes [12], be denoted by $\mathscr{P}_f^+$. More explicitly, $\mathscr{P}_f^+ = \{p : \mathscr{B}_\Sigma \to (0,1] \text{ such that } |\mathbf{N}_p| < \infty\}$,

★ This work has been supported in part by the U.S. Office of Naval Research under Grant No. N00014-09-1-0688, and the U.S. Army Research Laboratory and the U.S. Army Research Office under Grant No. W911NF-07-1-0376

where $\mathbf{N}_p$ denotes a partition induced by Nerode equivalence [13].

Let $\widetilde{\mathscr{A}} \triangleq \{G = (Q, \Sigma, \delta, q_0, \tilde{\pi}) : \tilde{\pi}(q, \sigma) > 0$ for all $q \in Q$ and all $\sigma \in \Sigma\}$. Algorithm 1 in the first part [13] has been constructed in the context of the probabilistic Nerode equivalence $\mathcal{N}_p$ such that a map $\widetilde{\mathbb{H}} : \widetilde{\mathscr{A}} \rightarrow \mathscr{P}_f^+$ is surjective.

**Definition II.1 (PFSA equivalence)** *Two PFSA $\widetilde{G}$ and $G$ are said to be equivalent if the associated probabilities are equal, i.e., $\widetilde{\mathbb{H}}(\widetilde{G}) = \widetilde{\mathbb{H}}(G)$. The equivalence class of $G$ is denoted as $\Xi(G) \triangleq \{\widetilde{G} \in \widetilde{\mathscr{A}} : \widetilde{\mathbb{H}}(\widetilde{G}) = \widetilde{\mathbb{H}}(G)\}$.*

By defining a quotient space $\mathscr{A} \triangleq \widetilde{\mathscr{A}}/\Xi$, the associated quotient map is obtained as:

$$\mathbb{H} : \mathscr{A} \longrightarrow \mathscr{P}_f^+ \tag{1}$$

Since a quotient map is injective, $\mathbb{H}$ becomes a bijection.

**Remark II.1** *For a PFSA $G$, each state $q \in Q$ is a Nerode equivalence class $S \in \mathbf{N}_{\mathbb{H}(G)}$. By Algorithm 1 in the first part [13], it follows that*

$$\tilde{\pi}(q, \sigma) = p(\sigma|S) \triangleq \frac{p(x\sigma)}{p(x)}, \forall x \in S \tag{2}$$

So far the vector space $(\mathscr{P}_f^+, \oplus, \odot)$ is established. By use of the bijection $\mathbb{H}$ and its inverse $\mathbb{F}$ (see Algorithm 1 in the first part [13]), new vector addition and scalar multiplication operations are introduced on the quotient space $\mathscr{A}$.

**Definition II.2 (Vector space $\mathscr{A}$)** *Let $G_1, G_2 \in \mathscr{A}$ and $k \in \mathbb{R}$. Then, following the definitions of vector addition $\oplus$ and scalar multiplication $\odot$ in the space of PFSA in the first part [13],*

- *The addition operation $+ : \mathscr{A} \times \mathscr{A} \rightarrow \mathscr{A}$ is defined as*

$$G_1 + G_2 = \mathbb{F}(\mathbb{H}(G_1) \oplus \mathbb{H}(G_2))$$

- *The scalar multiplication operation $\cdot : \mathbb{R} \times \mathscr{A} \rightarrow \mathscr{A}$ is defined as*

$$k \cdot G_1 = \mathbb{F}(k \odot (\mathbb{H}(G_1))$$

**Remark II.2** *Since $\mathbb{F} \triangleq \mathbb{H}^{-1}$, it follows from the Definition II.2 that $\mathbb{H}(G_1 + G_2) = \mathbb{H}(G_1) \oplus \mathbb{H}(G_2)$ and $\mathbb{H}(k \cdot G_1) = k \odot \mathbb{H}(G_1)$. Therefore, the bijection $\mathbb{H}$ is linear and hence $\mathbb{H}$ is an isomorphism between the vector spaces $(\mathscr{P}_f^+, \oplus, \odot)$ and $(\mathscr{A}, +, \cdot)$.*

**Definition II.3 (Symbolic White Noise)** *The zero element $\underline{e}$ of the vector space $\mathscr{P}_f^+$, defined as, $\underline{e}(x) = \frac{1}{|\Sigma|^{|x|}}, \forall x \in \Sigma^\star$, is called symbolic white noise.*

Note that $\underline{e}(\sigma|x) = \frac{1}{|\Sigma|}, \forall \sigma \in \Sigma, x \in \Sigma^\star$. The symbolic white noise can be modeled by a one-state PFSA, with the uniform distribution of symbol generation. For this process, every string of the same length has equal probability of occurrence and the knowledge of the history does not provide any information for predicting the future.

## III. Construction of a family of Inner Products

In order to build a framework for generating a family of inner products, a measure space $(\Sigma^\star, 2^{\Sigma^\star}, \mu)$ is constructed, where the finite measure $\mu : 2^{\Sigma^\star} \rightarrow [0, 1]$ has the following properties.

- $\mu(\emptyset) = 0$ and $\mu(\Sigma^\star) = 1$;
- $\mu\big(\bigcup_{k=1}^{\infty}\{x_k\}\big) = \sum_{k=1}^{\infty} \mu\big(\{x_k\}\big) \quad \forall x_k \in \Sigma^\star$

The second condition in the above statement implies that a non-negative measure is assigned to each of the singleton strings (including the null string $\epsilon$), which are considered to be mutually disjoint measurable sets. Thus, for any collection of strings, $L \in 2^{\Sigma^\star}$, $\mu(L) = \sum_{x \in L} \mu(\{x\})$.

The conditional probability $p(\cdot|\cdot) : \Sigma \times \Sigma^\star \rightarrow [0, 1]$ is now treated as a vector function for a given string $x \in \Sigma^\star$, and is denoted by $p(\cdot|x) : \Sigma^\star \rightarrow \mathbb{R}^{|\Sigma|}$ such that

$$p(\cdot|x) \triangleq [p(\sigma_1|x), p(\sigma_2|x), \ldots, p(\sigma_{|\Sigma|}|x)] \tag{3}$$

It follows from Eq. (3) that $\sum_{j=1}^{|\Sigma|} p(\sigma_j|x) = 1 \ \forall x \in \Sigma$.

**Definition III.1** *Given $p_1, p_2 \in \mathscr{P}_f^+$, an equivalence relation is defined as: $p_1 \sim p_2$ if $p_1(\cdot|x) = p_2(\cdot|x), \mu-a.e.$, which implyies that*

$$\mu\left(\{x \in \Sigma^\star : p_1(\cdot|x) \neq p_2(\cdot|x)\}\right) = 0 \tag{4}$$

*A quotient space is defined as: $\mathscr{Q} = \mathscr{P}_f^+/_\sim$ based on the above equivalence relation $\sim$.*

**Remark III.1** *If the following condition is imposed on the measure $\mu$ in Definition III.1:*

$$\mu(x) > 0, \forall x \in \Sigma^\star \tag{5}$$

*then the $\mu$-a.e. condition becomes everywhere. That is, the relation $p_1 \sim p_2$ becomes equivalent to to $p_1 = p_2$. In other words, $\mathscr{Q} = \mathscr{P}_f^+$ provided that Eq. (5) holds.*

**Proposition III.1** *$(\mathscr{Q}, \oplus, \odot)$ is a subspace of the vector space $(\mathscr{P}_f^+, \oplus, \odot)$.*

*Proof:* It follows from the definitions of vector addition $\oplus$ and scalar multiplication $\odot$ in the space of PFSA in the first part [13] that

$$(p_1 \oplus p_2)(\tau|x) = \frac{p_1(\tau|x)p_2(\tau|x)}{\sum_{\alpha \in \Sigma} p_1(\alpha|x)p_2(\alpha|x)} \tag{6}$$

$$(k \odot p_1)(\tau|x) = \frac{p_1^k(\tau|x)}{\sum_{\alpha \in \Sigma} p_1^k(\alpha|x)} \tag{7}$$

for all $\tau \in \Sigma$, $x \in \Sigma^\star$ and $k \in \mathbb{R}$.

Both of the above equations are consistent under the equivalence relation $\sim$, which means, $p_1 \sim \tilde{p}_1$ and $p_2 \sim \tilde{p}_2$ implies $(p_1 \oplus p_2) \sim (\tilde{p}_1 \oplus \tilde{p}_2)$ and $(k \odot p_1) \sim (k \odot \tilde{p}_1)$. ■

**Definition III.2** *On the vector space $\mathscr{Q}$ over the real field $\mathbb{R}$, a function $\langle \cdot, \cdot \rangle : \mathscr{Q} \times \mathscr{Q} \rightarrow \mathbb{R}$ is defined as*

$$\langle p_1, p_2 \rangle = \frac{1}{2} \sum_{\sigma_i, \sigma_j \in \Sigma} \sum_{x \in \Sigma^\star} \log \frac{p_1(x\sigma_i)}{p_1(x\sigma_j)} \log \frac{p_2(x\sigma_i)}{p_2(x\sigma_j)} \mu(\{x\}) \tag{8}$$

**Remark III.2** *In Eq. (8), the summation over $x \in \Sigma^\star$ could be recognized as an integration over $\Sigma^\star$ with the measure $\mu$. In this case, the integration degenerates to a summation because $\Sigma^\star$ is countable.*

**Theorem III.1 (Pre-Hilbert Space)** *In Definition III.2, the function $\langle \cdot, \cdot \rangle : \mathscr{Q} \times \mathscr{Q} \to \mathbb{R}$ is an inner product. That is, $(\mathscr{Q}, \oplus, \odot, \langle \cdot, \cdot \rangle)$ forms a pre-Hilbert space over the real field $\mathbb{R}$.*

*Proof:* While the symmetry property, i.e., $\langle p_1, p_2 \rangle = \langle p_2, p_1 \rangle$, is obvious, positive-definiteness is established as follows.

$$\langle p, p \rangle = \frac{1}{2} \sum_{\sigma_i, \sigma_j \in \Sigma} \sum_{x \in \Sigma^\star} \left( \log \frac{p(x\sigma_i)}{p(x\sigma_j)} \right)^2 \mu(\{x\}) \geq 0 \quad (9)$$

If $\langle p, p \rangle = 0$, non-negativity of each term in the summation mandates that, for $\mu$-almost every $x \in \Sigma^\star$, $\log \frac{p(x\sigma_i)}{p(x\sigma_j)} = 0$. Therefore, $p(\sigma_i | x) = \frac{1}{|\Sigma|}, \forall \sigma_i \in \Sigma$ and it follows from Definition II.3 that $p \sim \underline{e}$.

The linearity property is established as follows.

$$\begin{aligned}
&\langle a \odot p_1, p_2 \rangle \\
&= \frac{1}{2} \sum_{\sigma_i, \sigma_j \in \Sigma} \sum_{x \in \Sigma^\star} \log \frac{(a \odot p_1)(x\sigma_i)}{(a \odot p_1)(x\sigma_j)} \log \frac{p_2(x\sigma_i)}{p_2(x\sigma_j)} \mu(\{x\}) \\
&= \frac{1}{2} \sum_{\sigma_i, \sigma_j \in \Sigma} \sum_{x \in \Sigma^\star} \log \frac{p_1^a(x\sigma_i)}{p_1^a(x\sigma_j)} \log \frac{p_2(x\sigma_i)}{p_2(x\sigma_j)} \mu(\{x\}) \\
&= \frac{a}{2} \sum_{\sigma_i, \sigma_j \in \Sigma} \sum_{x \in \Sigma^\star} \log \frac{p_1(x\sigma_i)}{p_1(x\sigma_j)} \log \frac{p_2(x\sigma_i)}{p_2(x\sigma_j)} \mu(\{x\}) \\
&= a \langle p_1, p_2 \rangle \qquad (10)
\end{aligned}$$

and

$$\begin{aligned}
&\langle p_1 \oplus p_2, p_3 \rangle \\
&= \frac{1}{2} \sum_{\sigma_i, \sigma_j \in \Sigma} \sum_{x \in \Sigma^\star} \log \frac{(p_1 \oplus p_2)(x\sigma_i)}{(p_1 \oplus p_2)(x\sigma_j)} \log \frac{p_3(x\sigma_i)}{p_3(x\sigma_j)} \mu(\{x\}) \\
&= \frac{1}{2} \sum_{\sigma_i, \sigma_j \in \Sigma} \sum_{x \in \Sigma^\star} \log \frac{p_1(x\sigma_i) p_2(x\sigma_i)}{p_1(x\sigma_j) p_2(x\sigma_j)} \log \frac{p_3(x\sigma_i)}{p_3(x\sigma_j)} \mu(\{x\}) \\
&= \frac{1}{2} \sum_{\sigma_i, \sigma_j \in \Sigma} \sum_{x \in \Sigma^\star} \left( \log \frac{p_1(x\sigma_i)}{p_1(x\sigma_j)} + \log \frac{p_2(x\sigma_i)}{p_2(x\sigma_j)} \right) \\
&\qquad \cdot \log \frac{p_3(x\sigma_i)}{p_3(x\sigma_j)} \mu(\{x\}) \\
&= \langle p_1, p_3 \rangle + \langle p_2, p_3 \rangle \qquad (11)
\end{aligned}$$

■

Now the map $\mathbb{H} : \mathscr{A} \longrightarrow \mathscr{P}_f^+$ in Eq. (1) is used to define an inner product on the space $\mathscr{A}$ in terms of $\langle \cdot, \cdot \rangle$.

**Definition III.3** *The inner product $\langle \cdot, \cdot \rangle_A : \mathscr{A} \times \mathscr{A} \to \mathbb{R}$ is defined as*

$$\langle G_1, G_2 \rangle_A = \langle \mathbb{H}(G_1), \mathbb{H}(G_2) \rangle \quad (12)$$

By Remark III.1, this makes the map $\mathbb{H}$ an isometric isomorphism between the two pre-Hilbert spaces provided that Eq. (5) holds. Otherwise, a quotient space for $\widetilde{\mathscr{A}}$ needs to be defined to construct $\mathscr{A}$ similar to what is done in the standard $L_p$ space.. To simplify the notations, we abuse the notation $\mathscr{A}$ to represent a well-defined quotient space of $\widetilde{\mathscr{A}}$.

The following result is generated based on Algorithm 1 in the first part [13].

**Proposition III.2** *Let $G_i = (Q_i, \Sigma, \delta_i, q_0^i, \tilde{\pi}_i) \in \mathscr{A}$, $i = 1, 2$. The inner product can be computed as:*

$$\langle G_1, G_2 \rangle_A = \frac{1}{2} \sum_{\sigma_i, \sigma_j \in \Sigma} \sum_{q \in Q} \log \frac{\tilde{\pi}_1(q, \sigma_i)}{\tilde{\pi}_1(q, \sigma_j)} \log \frac{\tilde{\pi}_2(q, \sigma_i)}{\tilde{\pi}_2(q, \sigma_j)} \mu(q) \quad (13)$$

*where $Q = \{q_1 \bigcap q_2 : q_1 \in Q_1, q_2 \in Q_2\}$.*

If $G_1$ and $G_2$ have the same structure, then $Q = Q_1 = Q_2$ and Eq. (13) can be used directly to compute $\langle G_1, G_2 \rangle_A$. If $G_1$ and $G_2$ do not have the same structure, a synchronized composition [12] $G_1 \otimes G_2$ and $G_2 \otimes G_1$ needs to be constructed to have a common structure, namely,

$$\langle G_1, G_2 \rangle_A = \langle G_1 \otimes G_2, G_2 \otimes G_1 \rangle_A \quad (14)$$

The measure $\mu$ in Eq. (8) is user-selectable such that any choice of the finite measure $\mu$ yields a valid inner product.

As an example, if the measure $\mu$ is selected as

$$\mu(\{x\}) \triangleq \frac{|x|}{2^{|x|+1} |\Sigma|^{|x|}}, \quad (15)$$

then $\mu$ depends only on the length of the string. It makes sense to assign a smaller measure on a longer string since the probability of its occurrence is small. On the other hand, the rationale of assigning zero measure on the null string $\epsilon$, i.e., $\mu(\epsilon) = 0$ is as follows. Since the null string $\epsilon$ is in the initial state $q_0$, imposing $\mu(\epsilon) = 0$ puts less weight on the initial state. Alternatively, if the initial state $q_0$ is important, then a non-zero measure must be assigned on the null string. The following measure, for example, could be considered in this regard:

$$\mu(\{x\}) \triangleq \frac{1}{2^{|x|} |\Sigma|^{|x|}}$$

## IV. Computation of the Measure $\mu$

The following definitions are introduced to compute the measure $\mu$ defined in Eq. (15) for each state of a given PFSA.

**Definition IV.1** *Let the map $m_n : 2^{\Sigma^\star} \to [0, 1]$ be defined as*

$$m_n(L) \triangleq \frac{|\{x \in L : |x| = n\}|}{|\Sigma|^n} \quad \forall L \subseteq \Sigma^\star \quad (16)$$

**Remark IV.1** *$m_n(L)$ is the ratio of the number of strings of length $n$ in the set $L$ to the total number of strings of length $n$ in $\Sigma^\star$ and represents the size of the set $L$ in terms of strings of length $n$.*

**Definition IV.2 (Uniformizer of PFSA)** *Given a PFSA $G = (Q, \Sigma, \delta, q_0, \tilde{\pi})$, the PFSA $G'$ is called the* *uniformizer of $G$ if $G' = (Q, \Sigma, \delta, q_0, \tilde{\pi}')$, where $\tilde{\pi}'(q,\sigma) = \frac{1}{|\Sigma|}, \forall q \in Q, \forall \sigma \in \Sigma$.*

The uniformizer of a PFSA $G$ is denoted by $\mathbb{U}(G)$. The uniformizer simply modifies the original probability morph function to a uniform distribution over the symbols to each state. Note that $\mathbb{U}(G)$ retains the graph connectivity of $G$.

**Proposition IV.1** *Given a PFSA $G = (Q, \Sigma, \delta, q_0, \tilde{\pi}_G)$, then*

$$\boldsymbol{m}_n = \boldsymbol{m}_0 \left(\Pi^{\mathbb{U}(G)}\right)^n \tag{17}$$

*where $\boldsymbol{m}_n \triangleq [m_n(q_1), m_n(q_2), \ldots, m_n(q_{|Q|})]$, $\Pi^{\mathbb{U}(G)}$ is the state transition matrix for the uniformizer $\mathbb{U}(G)$, and*

$$\boldsymbol{m}_0(q) = \begin{cases} 1 & \textit{if } q = q_0 \\ 0 & \textit{if } q \neq q_0 \end{cases}$$

*Proof:* For any $q_i \in Q$ and $n \in \mathbb{N}$, it follows that

$$\begin{aligned} |\Sigma|^{n+1} m_{n+1}(q_i) &= |\{x \in q_i : |x| = n+1\}| \\ &= |\Sigma|^n \sum_{\delta(q_j,\sigma)=q_i} m_n(q_j) \end{aligned}$$

$$\begin{aligned} m_{n+1}(q_i) &= \frac{1}{|\Sigma|} \sum_{\delta(q_j,\sigma)=q_i} m_n(q_j) \\ &= \sum_{\delta(q_j,\sigma)=q_i} \tilde{\pi}^{\mathbb{U}(G)}(q_j,\sigma) m_n(q_j) \end{aligned} \tag{18}$$

Following the definition of a state transition probability matrix in the first part [13], a matrix representation of Eq. (18) is obtained as

$$\boldsymbol{m}_{n+1} = \boldsymbol{m}_n \Pi^{\mathbb{U}(G)} \tag{19}$$

by following Eq. (17) and this completes the proof. ■

Let us define $f_a(q) \triangleq \sum_{i=0}^{\infty} m_i(q) \cdot a^i$, where $0 < a < 1$. Then, given a PFSA $G = (Q, \Sigma, \delta, q_0, \tilde{\pi}) \in \mathscr{A}$, let us denote $\boldsymbol{f}_a \triangleq [f_a(q_1), f_a(q_2), \ldots, f_a(q_{|Q|})]$. Then,

$$\boldsymbol{f}_a = \sum_{i=0}^{\infty} \boldsymbol{m}_i \cdot a^i \tag{20}$$

Now, it follows from Eq. (17) that

$$\boldsymbol{f}_a = \boldsymbol{m}_0 \sum_{i=0}^{\infty} \left(a \Pi^{\mathbb{U}(G)}\right)^i = \boldsymbol{m}_0 \left(I - a \cdot \Pi^{\mathbb{U}(G)}\right)^{-1} \tag{21}$$

The last step is valid since $\|a \cdot \Pi^{\mathbb{U}(G)}\|_\infty < 1$.

**Proposition IV.2** *Given a PFSA $G = (Q, \Sigma, \delta, q_0, \tilde{\pi}) \in \mathscr{A}$, let us denote the measure $\boldsymbol{\mu} \triangleq [\mu(q_1), \mu(q_2), \ldots, \mu(q_{|Q|})]$. Then,*

$$\boldsymbol{\mu} = \frac{\boldsymbol{m}_0}{4} \left(I - \frac{1}{2}\Pi^{\mathbb{U}(G)}\right)^{-1} \Pi^{\mathbb{U}(G)} \left(I - \frac{1}{2}\Pi^{\mathbb{U}(G)}\right)^{-1} \tag{22}$$

*Proof:* For any $q \in Q$, we have

$$\begin{aligned} \mu(q) &= \sum_{x \in q} \mu(\{x\}) \\ &= \sum_{i=0}^{\infty} \left(m_i(q)|\Sigma|^i\right) \frac{i}{2^{i+1}|\Sigma|^i} \\ &= \sum_{i=0}^{\infty} m_i(q) \frac{i}{2^{i+1}} \end{aligned} \tag{23}$$

It follows from Eq. (21) that

$$\frac{d\boldsymbol{f}_a}{da} = \frac{1}{a^2} \sum_{i=0}^{\infty} \boldsymbol{m}_i \cdot i \cdot a^{i+1} \tag{24}$$

Comparing Eqs. (23) and (24), we obtain

$$\boldsymbol{\mu} = \left(a^2 \frac{d\boldsymbol{f}_a}{da}\right)_{a=\frac{1}{2}} \tag{25}$$

Since the convergence regions of $\frac{d\boldsymbol{f}_a}{da}$ and $\boldsymbol{f}_a$ are the same, convergence of $\mu(q)$ is guranteed. The fact that $\frac{dA^{-1}}{dt} = -A^{-1}\frac{dA}{dt}A^{-1}$ for the invertible matrix $A$ depending on a parameter $t$ implies

$$\begin{aligned} \boldsymbol{\mu} &= \left(a^2 \boldsymbol{m}_0 \left(I - a \cdot \Pi^{\mathbb{U}(G)}\right)^{-1} \Pi^{\mathbb{U}(G)} \right. \\ &\quad \left. \cdot \left(I - a \cdot \Pi^{\mathbb{U}(G)}\right)^{-1}\right)_{a=\frac{1}{2}} \\ &= \frac{\boldsymbol{m}_0}{4} \left(I - \frac{1}{2}\Pi^{\mathbb{U}(G)}\right)^{-1} \Pi^{\mathbb{U}(G)} \left(I - \frac{1}{2}\Pi^{\mathbb{U}(G)}\right)^{-1} \end{aligned} \tag{26}$$

■

## V. Interpretation of the Inner Product

In the literature of information theory [14], the entropy rate of a PFSA $G$ is defined as

$$h(G) \triangleq -\sum_{q \in Q} \wp(q) \left[\sum_{\sigma \in \Sigma} \tilde{\pi}(q,\sigma) \log \tilde{\pi}(q,\sigma)\right] \tag{27}$$

while in the present formulation, the induced norm of $G$ is

$$\|G\|_A \triangleq \sqrt{\sum_{q \in Q} \left[\frac{\mu(q)}{2} \sum_{\sigma_i,\sigma_j \in \Sigma} \left(\log\left(\frac{\tilde{\pi}(q,\sigma_i)}{\tilde{\pi}(q,\sigma_j}\right)\right)^2\right]} \tag{28}$$

Equations (27) and (28) have structural similarity in the sense that both are represented as a weighted sum over the states. However, the following two major differences are noteworthy.

1) For each state $q \in Q$, a weight $\mu(q)$ is used in Eq. (28) instead of the stationary probability $\wp(q)$ in Eq. (27).
2) The root mean square (rms) difference of logarithm of the probabilities of a pair of symbols conditioned on each state is used instead of the expectation of logarithm of a symbol's conditional probability. This rms value is a norm that is a consequence of the inner product.

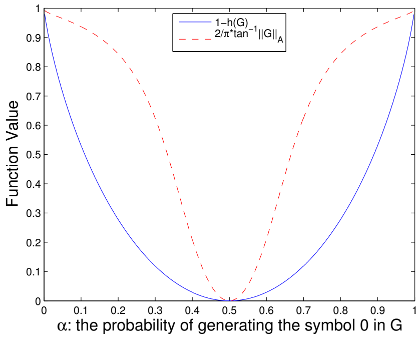


Fig. 1. Comparison of $(1 - h(G))$ and $\|G\|_A$ of an i.i.d process.

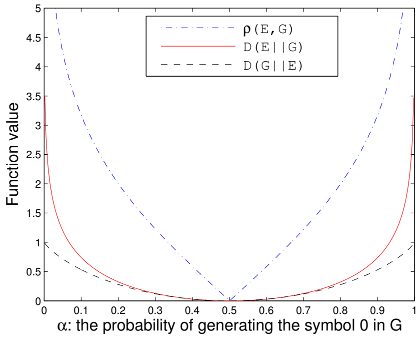


Fig. 2. Comparison of $\rho(E, G_2)$, $\mathbb{D}(E\|G_2)$, and $\mathbb{D}(G_2\|E)$ versus parameter $\alpha$, where $E$ is the *Symbolic White Noise*.

In contrast to the entropy (rate) which is a measure of the uncertainty, the ideal deterministic symbolic system should have the maximum norm while the completely random process should have a zero norm. For an independent and identically distributed (i.i.d.) process, namely, a single-state PFSA $G$, over the binary alphabet $\Sigma = \{0, 1\}$, let the probabilities of generating the symbol 0 and the symbol 1 be $\alpha$ and $(1 - \alpha)$, respectively, with $\alpha \in (0, 1)$.

Figure 1 compares $(1 - h(G))$ (solid line) and $\frac{2}{\pi}\tan^{-1}(\|G\|_A)$ (dashed line), where the entropy rate $h(G)$ has the range $[0, 1]$ and the norm curve is transformed from $[0, \infty)$ to $[0, 1]$. It is observed that the profiles for $(1-h(G))$ and $\frac{2}{\pi}\tan^{-1}(\|G\|_A)$ are qualitatively similar. Hence, it is possible to interpret the norm in Eq. (28) as a measure of certainty or information contained in $G$.

The K-L divergence [14] of two i.i.d. processes with probability mass functions $P_1$ and $P_2$ is

$$\mathcal{D}(P_1\|P_2) \triangleq \sum_i P_1(i) \log \frac{P_1(i)}{P_2(i)} \tag{29}$$

Let $E$ be the *Symbolic White Noise*, i.e., the i.i.d. process with uniform distribution over the symbols. Then, Figure 2 compares the induced distance $\rho(E, G_2) \triangleq \|E - G_2\|_A$ (dash-dot line), the K-L divergence $\mathcal{D}(E\|G_2)$ (solid line), and the K-L divergence $\mathcal{D}(G_2\|E)$ (dashed line) versus the probability parameter $\alpha$. It is seen that these three curves are qualitatively very similar as all of them approach infinity when $\alpha$ approaches 0 or 1 and achieve the minimum at 0 if $\alpha = 0.5$. An advantage of the proposed measure is that $\rho$ is a metric but K-L divergence is not.

## VI. Orthogonal Projection and Model Order Reduction

The space $\mathscr{A}$ of PFSA is not complete because a Cauchy sequence of PFSA with an increasingly number of states will have the limit point that is not a finite-state machine. However, in practice, a finite-dimensional subspace could be adequate for feature extraction from symbolic sequences. For example, finite-order $D$-Markov machines [7] have been used for anomaly detection. Since all finite-dimensional vector spaces over the real field $\mathbb{R}$ are guaranteed to be complete, any closed finite subspace of $\mathscr{A}$ is a well-defined Hilbert space. Therefore, an inner product in the space $\mathscr{A}$ admits the orthogonal projection whose existence and uniqueness is guaranteed on the Hilbert space.

Let $\mathbb{P}_{\mathscr{A}_2} : \mathscr{A}_1 \to \mathscr{A}_2$ denote the orthogonal projection from a closed subspace $\mathscr{A}_1 \subset \mathscr{A}$ onto another smaller closed subspace of $\mathscr{A}_2 \subset \mathscr{A}$. Then, if $\{V_i\}_{i=1}^n$ is an orthonormal basis for the space $\mathscr{A}_2$, where $n = \dim(\mathscr{A}_2)$, it follows that

$$\mathbb{P}_{\mathscr{A}_2}(G) = \sum_{i=1}^{n} \langle G, V_i \rangle_A \cdot G \tag{30}$$

The error due to projection onto the smaller dimensional space $\mathscr{A}_2$ is obtained as $\|G - \mathbb{P}_{\mathscr{A}_2}(G)\|_A$. In this setting, the model reduction problem is formulated as follows.

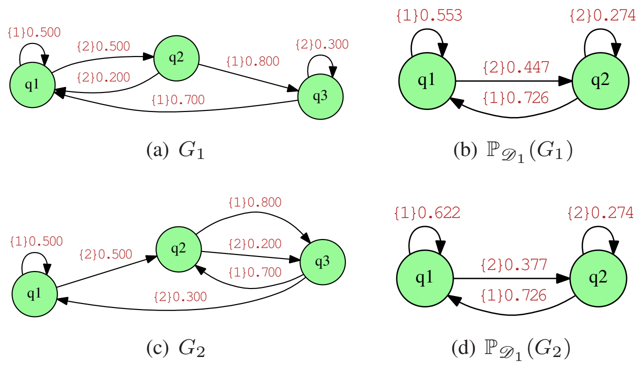
Let a PFSA $G$ on a space $\widetilde{\mathscr{S}} \subset \mathscr{A}$ represent the semantic model of a symbolic system. Let a desired sequence of lower dimensional closed subspaces be chosen in $\mathscr{A}$, which forms a projection chain $\mathscr{S}_1 \subset \mathscr{S}_2 \subset \ldots \subset \mathscr{S}_n \subset \widetilde{\mathscr{S}}$.

Let a cost functional $f : \mathscr{A} \to [0, \infty)$ be defined for $G$ as

$$f_G(\mathscr{S}) = \|G - \mathbb{P}_{\mathscr{S}}(G)\|_A + g(\mathbb{P}_{\mathscr{S}}(G)) \tag{31}$$

where the first term on the right hand is the projection error, interpreted as some form of information loss due to the projection and the second term is a user-selected cost functional of the projected model, which signifies the complexity of the projected model. For example, it can be taken as proportional to the number of states in $\mathbb{P}_{\mathscr{S}}(G)$. The objective is to minimize the cost functional $f$ over the projection chain; this problem can be numerically efficiently solved in the framework of the proposed orthogonal projection.

Now we present a numerical example using D-Markov machines [7] that have been used for system identification and anomaly detection. In this formulation, all D-Markov machines with positive morph matrices form a subspace of $\mathscr{A}$. Let this D-Markov subspace with depth $d$ be denoted by $\mathscr{D}_d$. (Note that $d$ is a positive integer.) Figure 3 presents two PFSA, $G_1$ and $G_2$, which are not D-Markov machines, and their projections onto the space $\mathscr{D}_1$ are $\mathbb{P}_{\mathscr{D}_1}(G_1)$ and $\mathbb{P}_{\mathscr{D}_1}(G_2)$, respectively. Figure 4 displays the projection errors of PFSA $G_1$ (solid line) and $G_2$ (dashed line) onto the projection chain $\mathscr{D}_1 \subset \mathscr{D}_2 \subset \ldots \subset \mathscr{D}_8$, respectively. It is observed that, for both $G_1$ and $G_2$, the projection errors



(a) $G_1$ (b) $\mathbb{P}_{\mathscr{D}_1}(G_1)$

(c) $G_2$ (d) $\mathbb{P}_{\mathscr{D}_1}(G_2)$

Fig. 3. Projection of PFSA $G_1$ and $G_2$ on $\mathscr{D}_1$.

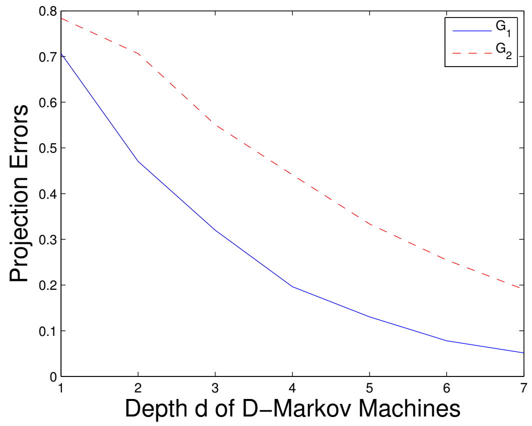


Fig. 4. Projection errors of $G_1$ and $G_2$ on D-Markov subspaces.

decrease as the order $d$ becomes larger. To interpret the meaning of the projection, a symbol sequence of length $10,000$ is generated by simulating the PFSA $G_1$ and then the symbol sequence is used to obtain a D-Markov machine with depth $d = 1$. The resulting output is shown in Figure 5, which is very close to the analytically derived projection $\mathbb{P}_{\mathscr{D}_1}(G_1)$ in Figure 3(b). That is, the lower order model captured by the D-Markov algorithm from the simulated symbolic sequence is very close to the optimal projection point in the proposed Hilbert space setting.

## VII. Conclusions and Future Work

This second part of the two-part paper introduces a family of inner products on the vector space of PFSA that is constructed in the first part [13]. The objective here is to develop a numerically efficient tool of model order reduction for symbolic system identification. From this perspective, the norm induced by the inner product is interpreted as a measure of the information contained in the PFSA. The process of order reduction in PFSA models is presented as an orthogonal projection in the Hilbert space setting. A numerical example is presented to illustrate the procedure. While there are numerous research issues that need to addressed before commercial codes of model identification and order reduction can be made available, a few research topics are presented below.

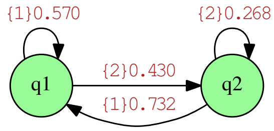


Fig. 5. 1D-Markov Machine based on simulated sequences from $G_1$.

- For model reduction, the performance of different state merging algorithms for PFSA model identification need to be quantitatively evaluated by the induced metric.
- For pattern classification, PFSA models of symbolic systems need to be chosen as feature vectors for pattern classification in the space $\mathscr{A}$.
- For optimization involving PFSA models, square of the norm $\| \cdot \|_A^2$, induced by the inner product, could be used as a mathematical structure of the cost functional; further theoretical research is necessary in this direction.

## References

[1] T. Breiten and T. Damm, “Krylov subspace methods for model order reduction of bilinear control systems,” *System and Control Leeters*, vol. 59, 2010.

[2] R. Badii and A. Politi, *Complexity: Hierarchical Structures and Scaling in Physics*. Cambridge, U.K: Cambridge University Press, 1997.

[3] D. Lind and B. Marcus, *An Introduction to Symbolic Dynamics and Coding*. Cambridge, U.K: Cambridge University Press, 1995.

[4] M. Sipser, *Introduction to the Theory of Computation*. Boston, MA, USA: PWS Publishing, 1997.

[5] J. E. Hopcroft, R. Motwani, and J. D. Ullman, *Introduction to Automata Theory, Languages, and Computation, 2nd edition*. Boston, MA, USA: Addison-Wesley, 2001.

[6] L.Ljung, *System Identification*. Upper Saddle River, NJ, USA: Prentice Hall, 1995.

[7] A. Ray, “Symbolic dynamic analysis of complex systems for anomaly detection,” *Signal Processing*, vol. 84, no. 7, pp. 1115–1130, 2004.

[8] V. Rajagopalan and A. Ray, “Symbolic time series analysis via wavelet-based partitioning,” *Signal Processing*, vol. 86, no. 11, pp. 3309–3320, Nov 2006.

[9] G. Kotsalis and M. Dahleh, “Model reduction of irreducible markov chains,” in *the 42nd IEEE Conference on Decision and Control*, 2003.

[10] P. M. K. Deng, Y. Sun and S. Meyn, “An information-theoretic framework to aggregate a markov chain,” in *American Control Conference*, 2009.

[11] Z. Rached, F. Alalaji, and L. Campbell, “The kullback-leibler divergence rate between markov sources,” *IEEE Transactions on Information Theory*, vol. 50, no. 5, pp. 917–921, 2004.

[12] I. Chattopadhyay and A. Ray, “Structural transformations of probabilistic finite state machines,” *International Journal of Control*, vol. 81, no. 5, pp. 820–835, 2008.

[13] Y. Wen, A. Ray, I. Chattopadhyay, and S. Phoha, “Modeling of symbolic systems: Part I - vector space representation of probabilistic finite state automata,” in *2011 American Control Conference, San Francisco, CA, USA*, June-July.

[14] T. Cover and J. Thomas, *Elements of Information Theory*. New York, NY, USA: Wileay-interscience, 2 ed., 2006.