1 Formalization Of Probabilistic Generators For Stochastic Dynamical Systems

Throughout this paper, Σ denotes a finite alphabet of symbols. The set of all finite but possibly unbounded strings on Σ is denoted by $\Sigma^\star$, the Kleene $\star$ operation [8]. The set of finite strings over Σ form a concatenative monoid, with the empty word λ as identity. Concatenation of two strings $x, y \in \Sigma^\star$ is written as xy. Thus $xy \quad x\lambda y \quad xy\lambda \quad \lambda xy$. The set of strictly infinite strings on Σ is denoted as Σ^ω, where ω denotes the first transfinite cardinal. For a string $x \in \Sigma^\star$, $|x|$ denotes the length of x and for a set A, $|A|$ denotes its cardinality.

1.1 Quantized Stochastic Processes

Definition 1 (QSP). *A QSP $\mathcal{H}$ is a discrete time Σ-valued strictly stationary, ergodic stochastic process, i.e.*

$$\mathcal{H} \quad \{X_t \quad X_t \text{ is a } \Sigma\text{-valued random variable for } t \in \mathbb{N} \cup \{0\}\}$$

A stochastic process is ergodic if moments can be calculated from a single, sufficiently long realization, and strictly stationary if moments are not functions of time.

We next formalize the connection of QSPs to PFSA generators.

Definition 2 (σ-Algebra On Infinite Strings). *For the set of infinite strings on Σ, we define $\mathbb{B}$ to be the smallest σ-algebra generated by $\{x\Sigma^\omega \quad x \in \Sigma^\star\}$.*

Lemma 1. *Any QSP induces a probability space $\Sigma^\omega, \mathbb{B}, \mu$.*

Proof. Using stationarity of QSP $\mathcal{H}$ we can construct a probability measure $\mu \quad \mathbb{B} \to 0, 1$ by defining for any sequence $x \in \Sigma^\star \setminus \{\lambda\}$, and a sufficiently large number of realizations N_R of $\mathcal{H}$, with fixed initial conditions:

$$\mu x\Sigma^\omega \quad \lim_{N_R \to} \frac{\left(\#\text{ of initial occurrences of } x\right)}{\left(\begin{array}{l}\#\text{ of initial occurrences of}\\ \text{all sequences of length } |x|\end{array}\right)}$$

and extending the measure to elements of $\mathbb{B}\backslash B$ via at most countable sums. Note that $\mu\Sigma^\omega \quad \sum_{x\in\Sigma^\star} \mu x\Sigma^\omega \quad 1$, and for the null word $\mu\lambda\Sigma^\omega \quad \mu\Sigma^\omega \quad 1$. □

For notational brevity, we denote $\mu x\Sigma^\omega$ as Prx.

Classically, states are induced via the Nerode equivalence, which defines two strings to be equivalent if and only if any finite extension of the strings is either both accepted, or both rejected [8] by the language under consideration. We use a probabilistic extension [5].

Definition 3 (Probabilistic Nerode Relation). *$\Sigma^\omega, \mathbb{B}, \mu$ induces an equivalence relation $\sim_N$ on the set of finite strings $\Sigma^\star$ as:*

$$\forall x, y \in \Sigma^\star, x \sim_N y \iff \forall z \in \Sigma^\star\big(\big(Prxz \quad Pryz \quad 0\big) \bigvee |Prxz/Prx - Pryz/Pry| \quad 0\big) \tag{1}$$

For $x \in \Sigma^\star$, the equivalence class of x is denoted as x.

It follows that $\sim_N$ is right invariant, *i.e.*

$$x \sim_N y \Rightarrow \forall z \in \Sigma^\star, xz \sim_N yz \tag{2}$$

A right-invariant equivalence on $\Sigma^\star$ always induces an automaton structure.

Definition 4 (Initial-Marked PFSA). *An Initial-Marked PFSA is a 5-tuple $Q, \Sigma, \delta, \widetilde{\pi}, q_0$, where Q is a finite state set, Σ is the alphabet, $\delta \quad Q \times \Sigma \to Q$ is the state transition function, and $\widetilde{\pi} \quad Q \times \Sigma \to 0, 1$ specifies the conditional symbol-generation probabilities. δ and $\widetilde{\pi}$ are recursively extended to arbitrary $y \quad \sigma x \in \Sigma^\star$ as $\delta q, \sigma x \quad \delta\delta q, \sigma, x$, and $\widetilde{\pi} q, \sigma x \quad \widetilde{\pi} q, \sigma\widetilde{\pi}\delta q, \sigma, x$. $q_0 \in Q$ is the initial state. If the next symbol is specified, our resultant state is fixed; similar to Probabilistic Deterministic Automata [7]. However, unlike the latter, we lack final states. Additionally, we assume our graphs to be strongly connected.*

Definition 4 has no notion of a final state, and later we will remove initial state dependence using ergodicity. First we formalize how the notion of a PFSA arises from a QSP.

Lemma 2 (From QSP To A PFSA). *If the probabilistic Nerode relation has a finite index, then there exists an initial-marked PFSA generator.*

Proof. Every QSP represented as a probability space $\Sigma^\omega, \mathbb{B}, \mu$ induces a probabilistic automaton $Q, \Sigma, \delta, \tilde{\pi}, q_0$, where Q is the set of equivalence classes of the probabilistic Nerode relation (Definition 3), Σ is the alphabet, and:

$$\delta x, \sigma \quad x\sigma \tag{3}$$

$$\tilde{\pi} x, \sigma \quad \frac{Prx'\sigma}{Prx'} \text{for any choice of } x' \in x \tag{4}$$

q_0 is identified with λ, and finite index of $\sim_N$ implies $|Q| < \ $. □

The above construction yields a *minimal realization* unique up to state renaming.

Corollary 1 (To Lemma 2: Null-word Probability). *For the PFSA $Q, \Sigma, \delta, \tilde{\pi}$ induced from a QSP $\mathcal{H}$:*

$$\forall q \in Q, \tilde{\pi} q, \lambda \quad 1 \tag{5}$$

Proof. For $q \in Q$ let $x \in \Sigma^\star$ such that $x \ q$. From Eq. (4), we have:

$$\tilde{\pi} q, \lambda \quad \frac{Prx'\lambda}{Prx'} \text{ for } x' \in x \tag{6}$$

$$\frac{Prx'}{Prx'} \quad 1 \tag{7}$$

□

While many reported approaches *define* the probability of the null-word to be unity, we can derive it from our formulation.

1.2 Canonical Representations

Next we define canonical representations to remove initial-state dependence. We use $\tilde{\Pi}$ to denote the matrix representation of $\tilde{\pi}$, *i.e.*, $\tilde{\Pi}_{ij} \ \tilde{\pi} q_i, \sigma_j, \ q_i \in Q, \sigma_j \in \Sigma$. We need the notion of transformation matrices Γ_σ.

Definition 5 (Transformation Matrices). *For an initial-marked PFSA $G \ Q, \Sigma, \delta, \tilde{\pi}, q_0$, the symbol-specific transformation matrices $\Gamma_\sigma \in \{0, 1\}^{|Q| \times |Q|}$ are:*

$$\Gamma_\sigma|_{ij} \ \begin{cases} \tilde{\pi} q_i, \sigma, & \text{if } \delta q_i, \sigma \ q_j \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

Transformation matrices have a single non-zero entry per row, reflecting our generation rule that given a state and a generated symbol, the next state is fixed. States in the canonical representation (denoted as $\wp_x$) are identified with probability distributions over states of the initial-marked PFSA. Here x denotes the string in $\Sigma^\star$ realizing this distribution, beginning from the stationary distribution on the states of the initial-marked representation. $\wp_x$ is an equivalence class, and hence x is not unique.

Definition 6 (Canonical Representations). *An initial-marked PFSA $G \ Q, \Sigma, \delta, \tilde{\pi}, q_0$ uniquely induces a canonical representation $Q^C, \Sigma, \delta^C, \tilde{\pi}^C$, with Q^C being the set of probability distributions over Q, $\delta^C \ Q^C \times \Sigma \to Q^C$, and $\tilde{\pi}^C \ Q^C \times \Sigma \to 0, 1$, as follows:*

1. *Construct the stationary distribution on Q using the transition probabilities of the Markov Chain induced by G, and include this as the first element $\wp_\lambda$ of Q^C. The transition matrix for this induced chain is the row-stochastic matrix $M \in 0, 1^{|Q| \times |Q|}$, with $M_{ij} \ \sum_{\sigma \delta q_i, \sigma q_j} \tilde{\pi} q_i, \sigma$.*

2. Define δ^C and $\widetilde{\pi}^C$ recursively:

$$\delta^C(\wp_x, \sigma) \triangleq \frac{1}{||\wp_x \Gamma_\sigma||_1} \wp_x \Gamma_\sigma \triangleq \wp_{x\sigma} \tag{9}$$

$$\widetilde{\pi}^C(\wp_x, \sigma) \triangleq \wp_x \widetilde{\Pi} \tag{10}$$

For a QSP $\mathcal{H}$, the canonical representation is denoted as $\mathcal{C}_\mathcal{H}$.

Ergodicity of QSPs, which makes $\wp_\lambda$ independent of the initial state in the initial-marked PFSA, implies that the canonical representation is initial state independent (See Figure **??**), and subsumes the initial-marked representation in the following sense:

Let $\mathcal{E} \triangleq \{e^i \in 0\ 1^{|Q|}, i \triangleq 1, \cdots, |Q|\}$ denote the set of distributions satisfying:

$$e^i|_j \triangleq \begin{cases} 1, & \text{if } i \triangleq j \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

Then we note:

1. If we execute the canonical construction with an initial distribution from $\mathcal{E}$, then we get the initial-marked PFSA (with the initial marking missing).
2. If during the construction we encounter $\wp_x \in \mathcal{E}$ for some x, then we stay within the graph of the initial-marked PFSA for all right extensions of x. This thus eliminates the need of knowing the initial state explicitly (See Figure **??**), provided we find string x, which takes us within or close to $\mathcal{E}$.

Consequently, we denote the initial-marked PFSA induced by a QSP $\mathcal{H}$, with the initial marking removed, as $\mathcal{P}_\mathcal{H}$, and refer to it simply as a "PFSA" (dropping the qualifier "initial-marked"). States in $\mathcal{P}_\mathcal{H}$ are representable as states in $\mathcal{C}_\mathcal{H}$ as elements of $\mathcal{E}$. Next we establish a key result: we always encounter a state arbitrarily close to some element in $\mathcal{E}$ in the canonical construction starting from the stationary distribution $\wp_\lambda$.

Theorem 1 (ϵ-Synchronization of Probabilistic Automata). *For any QSP $\mathcal{H}$ over Σ, the PFSA $\mathcal{P}_\mathcal{H}$ satisfies:*

$$\forall \epsilon' > 0, \exists x \in \Sigma^\star, \exists \vartheta \in \mathcal{E}, ||\wp_x - \vartheta|| \leqq \epsilon' \tag{12}$$

where the norm used is unimportant.

Proof. We show that all PFSA are at least approximately synchronizable [3, 9], which is not true for deterministic automata. If the graph of $\mathcal{P}_\mathcal{H}$ (*i.e.*, the deterministic automaton obtained by removing the arc probabilities) is synchronizable, then Eq. (12) trivially holds true for $\epsilon' \triangleq 0$ for any synchronizing string x. Thus, we assume the graph of $\mathcal{P}_\mathcal{H}$ to be non-synchronizable. From definition of non-synchronizability, it follows:

$$\forall q_i, q_j \in Q, \text{with } q_i \neq q_j, \forall x \in \Sigma^\star, \delta q_i, x \neq \delta q_j, x \tag{13}$$

If the PFSA has a single state, then every string satisfies the condition in Eq. (12). Hence, we assume that the PFSA has more than one state. Now if we have:

$$\forall x \in \Sigma^\star, \frac{Prx'x}{Prx'} \triangleq \frac{Prx''x}{Prx''} \text{ where } x' \triangleq q_i, x'' \triangleq q_j \tag{14}$$

then, by the Definition 3 , we have a contradiction $q_i \triangleq q_j$. Hence $\exists x_0$ such that

$$\frac{Prx'x_0}{Prx'} \neq \frac{Prx''x_0}{Prx''} \text{ where } x' \triangleq q_i, x'' \triangleq q_j \tag{15}$$

Since:

$$\sum_{x \in \Sigma^\star} \frac{Prx'x}{Prx'} \triangleq 1, \text{ for any } x' \text{ where } x' \triangleq q_i \tag{16}$$

we conclude without loss of generality that $\forall q_i, q_j \in Q$, with $q_i \neq q_j$:

$$\exists x^{ij} \in \Sigma^\star, \frac{Prx'x^{ij}}{Prx'} > \frac{Prx''x^{ij}}{Prx''} \text{ where } x' \triangleq q_i, x'' \triangleq q_j$$

It follows from induction that if we start with a distribution $\wp$ on Q such that $\wp_i \;\; \wp_j \;\; 0.5$, then for any $\epsilon' > 0$ we can construct a finite string x_0^{ij} such that if $\delta q_i, x_0^{ij} \;\; q_r, \delta q_j, x_0^{ij} \;\; q_s$, then for the new distribution $\wp'$ after execution of x_0^{ij} will satisfy $\wp'_s > 1 - \epsilon'$. Recalling that $\mathcal{P}_\mathcal{H}$ is strongly connected, Now, for any $q_t \in Q$, there exists a string $y \in \Sigma^\star$, such that $\delta q_s, y \;\; q_t$. Setting $x_\star^{i,j\to t} \;\; x_0^{ij} y$, we can ensure that the distribution $\wp''$ obtained after execution of $x_\star^{ij}$ satisfies $\wp''_t > 1 - \epsilon'$ for any q_t of our choice. For arbitrary initial distributions $\wp^A$ on Q, we consider contributions arising from simultaneously executing $x_\star^{i,j\to t}$ from states other than just q_i and q_j. Nevertheless, it is easy to see that executing $x_\star^{i,j\to t}$ implies that in the new distribution $\wp^{A'}$, we have $\wp_t^{A'} > \wp_i^A \;\; \wp_j^A - \epsilon'$. It immediately follows that executing of the string $x^{1,2\to|Q|} x^{3,4\to|Q|} \ldots x^{n-1,n\to|Q|}$, where

$$n \;\; \begin{cases} |Q| & \text{if } |Q| \text{ is even} \\ |Q| - 1 & \text{otherwise} \end{cases} \tag{17}$$

would result in a final distribution $\wp^{A''}$ which satisfies $\wp_{|Q|}^{A''} > 1 - \frac{1}{2} n\epsilon'$. Appropriate scaling of ϵ' then completes the proof. □

Theorem 1 induces the notion of ϵ-synchronizing strings, and guarantees their existence for arbitrary PFSA.

Definition 7 (ϵ-synchronizing Strings). *An ϵ-synchronizing string $x \in \Sigma^\star$ for a PFSA is one that satisfies:*

$$\exists \vartheta \in \mathcal{E}, ||\wp_x - \vartheta|| \leqq \epsilon \tag{18}$$

The norm used is unimportant.

Theorem 1 does not yield an algorithm for computing synchronizing strings (See Theorem 3). It simply shows that one always exists. As a corollary, we estimate an asymptotic upper bound on the effort required to find it.

Corollary 2 (To Theorem 1). *At most $O 1/\epsilon$ strings need to be analyzed to find an ϵ-synchronizing string.*

Proof. Theorem 1 works by multiplying entries from the $\widetilde{\Pi}$ matrix, which cannot be all identical (otherwise the states would collapse). Let the minimum difference between two unequal entries be η. Then, following the construction in Theorem 1, the length ℓ of the synchronizing string, up to linear scaling, satisfies: $\eta^\ell \;\; O\epsilon$, implying $\ell \;\; O \log 1/\epsilon$. The number of strings to be analyzed therefore is at most all strings of length ℓ, which is given by

$$|\Sigma|^\ell \;\; |\Sigma|^{O \log 1/\epsilon} \;\; O 1/\epsilon \tag{19}$$

□

Next, we describe the basic principle of our inference algorithm. PFSA states are not observable; we observe symbols generated from hidden states. This leads us to the notion of *symbolic derivatives*, which are computable from observations.

We denote the set of probability distributions over a set of cardinality k as $\mathscr{D}k$. First, we specify a count function.

Definition 8 (Symbolic Count Function). *For a string s over Σ, the count function $\#^s \;\; \Sigma^\star \to \mathbb{N} \cup \{0\}$, counts the number of times a particular substring occurs in s. The count is overlapping, i.e., in a string s 0001, we count the number of occurrences of 00s as $\underline{000}1$ and $0\underline{00}1$, implying $\#^s 00$ 2.*

Definition 9 (Symbolic Derivative). *For a string s generated by a QSP over Σ, the symbolic derivative is a function $\phi^s \;\; \Sigma^\star \to \mathscr{D}|\Sigma| - 1$ as:*

$$\phi^s x|_i \;\; \frac{\#^s x\sigma_i}{\sum_{\sigma_i \in \Sigma} \#^s x\sigma_i} \tag{20}$$

Thus, $\forall x \in \Sigma^\star, \phi^s x$ is a probability distribution over Σ. $\phi^s x$ is referred to as the symbolic derivative at x.

For $q_i \in Q$, $\widetilde{\pi}$ induces a distribution over Σ as $\widetilde{\pi}q_i, \sigma_1, \cdots, \widetilde{\pi}q_i, \sigma_{|\Sigma|}$. We denote this as $\widetilde{\pi}q_i, \cdot$. We show that the symbolic derivative at x can be used to estimate this distribution for $q_i \quad x$, provided x is ϵ-synchronizing.

Theorem 2 (ϵ-Convergence). *If $x \in \Sigma^\star$ is ϵ-synchronizing, then:*

$$\forall \epsilon > 0, \lim_{|s|\to} ||\phi^s x - \widetilde{\pi}x, \cdot|| \leqq_{a.s.} \epsilon \tag{21}$$

Proof. The proof follows from the Glivenko-Cantelli theorem [15] on uniform convergence of empirical distributions. Since x is ϵ-synchronizing, we have:

$$\forall \epsilon > 0, \exists \vartheta \in \mathcal{E}, ||\wp_x - \vartheta|| \leqq \epsilon \tag{22}$$

Let x ϵ-synchronize to q. Thus, every time we encounter x while reading s, we are guaranteed to be distributed over Q, and at most ϵ distance from the element of $\mathcal{E}$ corresponding to q. Assuming that we encounter $n|s|$ occurrences of x within s, we note that $\phi^s x$ is an approximate empirical distribution for $\widetilde{\pi}q, \cdot$. Denoting $F_{n|s|}$ as the perfect empirical distributions (*i.e.* ones that would be obtained for $\epsilon \ 0$), we have:

$$\begin{aligned}
&\lim_{|s|\to} ||\phi^s x - \widetilde{\pi}q, \cdot|| \\
&\qquad \lim_{|s|\to} ||\phi^s x - F_{n|s|} \ F_{n|s|} - \widetilde{\pi}q, \cdot|| \\
&\qquad \leqq \lim_{|s|\to} ||\phi^s x - F_{n|s|}|| \overbrace{\lim_{|s|\to} ||F_{n|s|} - \widetilde{\pi}q, \cdot||}^{\text{a.s. 0 by Glivenko-Cantelli}} \leqq_{a.s.} \epsilon
\end{aligned}$$

We use $n|s| \to$ as $|s| \to$, implied by the strong connectivity of our PFSA. □

Next we describe identification of ϵ-synchronizing strings given a sufficiently long observed string s. Theorem 1 guarantees existence, and Corollary 2 establishes that $O1/\epsilon$ substrings need to be analyzed till we encounter an ϵ-synchronizing string. These do not provide an executable algorithm, which arises from an inspection of the geometric structure of the set of probability vectors over Σ, obtained by constructing $\phi^s x$ for different choices of the candidate string x.

Definition 10 (Derivative Heap). *Given a string s generated by a QSP, a derivative heap $\mathcal{D}^s \ 2^{\Sigma^\star} \to \mathscr{D}|\Sigma| - 1$ is the set of probability distributions over Σ calculated for a given subset of strings $L \subset \Sigma^\star$ as follows:*

$$\mathcal{D}^s L \ \{\phi^s x \ \ x \in L \subset \Sigma^\star\} \tag{23}$$

Lemma 3 (Limiting Geometry). *Let $\mathcal{D} \ \lim_{|s|\to} \lim_{L\to\Sigma^\star} \mathcal{D}^s L$, and $\mathscr{U}$ be the convex hull of $\mathcal{D}$. If u is a vertex of $\mathscr{U}$, then*

$$\exists q \in Q, \textit{such that } u \ \widetilde{\pi}q, \cdot \tag{24}$$

Proof. Recalling Theorem 2, the result follows from noting that any element of $\mathcal{D}$ is a convex combination of elements from the set $\{\widetilde{\pi}q_1, \cdot, \cdots, \widetilde{\pi}q_{|Q|}, \cdot\}$. □

Lemma 3 does not claim that the number of vertices of the convex hull of $\mathbb{D}$ equals the number of states, but that every vertex corresponds to a state. We cannot generate $\mathcal{D}$ in practice since we have a finite observed string s, and we can only calculate $\phi^s x$ for a finite number of x. Instead, we show that choosing a string corresponding to the vertex of the convex hull of the heap, constructed by considering $O1/\epsilon$ strings, gives us an ϵ-synchronizing string with high probability.

Theorem 3 (Derivative Heap Approximation). *For s generated by a QSP, let $\mathcal{D}^s L$ be computed with $L \ \Sigma^{O\log 1/\epsilon}$. If for some $x_0 \in \Sigma^{O\log 1/\epsilon}$, $\phi^s x_0$ is a vertex of the convex hull of $\mathcal{D}^s L$, then*

$$Prob x_0 \textit{ is not } \epsilon\textit{-synchronizing} \leqq e^{-|s|\epsilon O1} \tag{25}$$

Proof. The result follows from Sanov's Theorem [6] for convex set of probability distributions. Note that if $|s| \to$, then x_0 is guaranteed to be ϵ-synchronizing (Theorem 1, and Corollary 2). Denoting the number of times we encounter x_0 in s as $n|s|$, and since $\mathcal{D}$ is a convex set of distributions (allowing us to drop the polynomial factor in Sanov's bound), we apply Sanov's Theorem to the case of finite s:

$$Prob\big(KL\big(\phi^s x_0 || \wp_{x_0} \widetilde{\Pi}\big) > \epsilon\big) \leqq e^{-n|s|\epsilon} \tag{26}$$

where $KL\cdot||\cdot$ denotes the Kullback-Leibler divergence [12]. From the bound [16]:

$$\frac{1}{4}||\phi^s x_0 - \wp_{x_0} \widetilde{\Pi}||^2 \leqq KL\big(\phi^s x_0 || \wp_{x_0} \widetilde{\Pi}\big) \tag{27}$$

and $n|s| \to |s|\alpha \;\; |s| O_1$, where $\alpha > 0$ is the stationary probability of encountering x_0 in s, we conclude:

$$Prob\big(||\phi^s x_0 - \wp_{x_0} \widetilde{\Pi}|| > \epsilon\big) \leqq 2e^{-\frac{1}{2}|s|\epsilon O_1} \;\; e^{-|s|\epsilon O_1} \tag{28}$$

□

Figure ?? illustrates PFSAs, derivative heaps, and ϵ-synchronizing strings. Next, we present our inference algorithm. Next, we use the preceding theoretical development to construct an effective procedure to infer $PFSA\ \mathcal{P}_{\mathcal{H}}$ from a sufficiently long run from a QSP $\mathcal{H}$, and a pre-specified $\epsilon > 0$.

2 Algorithm GenESeSS

2.1 Implementation Steps

We call our algorithm "Generator Extraction Using Self-similar Semantics", or GenESeSS which for an observed sequence s, consists of three steps:

1. *Identification of ϵ-synchronizing string x_0:* Construct a derivative heap $\mathcal{D}^s L$ using the observed trace s (Definition 10), and set L consisting of all strings up to a sufficiently large, but finite, depth. We suggest as initial choice of L as $\log_{|\Sigma|} 1/\epsilon$. In L is sufficiently large, then the inferred model structure will not change for larger values. We then identify a vertex of the convex hull for $\mathcal{D}$, via any standard algorithm for computing the hull [2]. Choose x_0 as the string mapping to this vertex.

2. *Identification of the structure of $\mathcal{P}_{\mathcal{H}}$, i.e., transition function δ:* We generate δ as follows: For each state q, we associate a string identifier $x_q^{id} \in x_0 \Sigma^\star$, and a probability distribution h_q on Σ (which is an approximation of the $\widetilde{\Pi}$-row corresponding to state q). We extend the structure recursively:

   (a) Initialize the set Q as $Q \;\; \{q_0\}$, and set $x_{q_0}^{id} \;\; x_0$, $h_q \;\; \phi^s x_0$.

   (b) For each state $q \in Q$, compute for each symbol $\sigma \in \Sigma$, find symbolic derivative $\phi^s x_q^{id} \sigma$. If $||\phi^s x_q^{id} \sigma - h_{q'\sigma}|| \leqq \epsilon$ for some $q' \in Q$, then define $\delta q, \sigma \;\; q'$. If, on the other hand, no such q' can be found in Q, then add a new state q' to Q, and define $x_{q'}^{id} \;\; x_q^{id}\sigma$, $h_{q'} \;\; \phi^s x_q^{id}\sigma$.

   The process terminates when every $q \in Q$ has a target state, for each $\sigma \in \Sigma$. Then, if necessary, we ensure strong connectivity using [14].

3. *Identification of arc probabilities, i.e., function $\widetilde{\pi}$:*

   (a) Choose an arbitrary initial state $q \in Q$.

   (b) Run sequence s through the identified graph, as directed by δ, *i.e.*, if current state is q, and the next symbol read from s is σ, then move to $\delta q, \sigma$. Count arc traversals, *i.e*, generate numbers N_j^i where $q_i \overset{\sigma_j}{\underset{N_j^i}{\to}} q_k$.

   (c) Generate $\widetilde{\Pi}$ by row normalization, *i.e.*, $\widetilde{\Pi}_{ij} \;\; N_j^i / \sum_j N_j^i$

[7, 4] use similar recursive structure extension. However, with no notion of ϵ-synchronization, they are restricted to inferring only synchronizable or short-memory models, or large approximations for long-memory ones (See Figure ??).

2.2 Complexity Analysis & PAC Learnability

While h_q (in step 2) approximates $\widetilde{\Pi}$ rows, we find the arc probabilities via normalization of traversal count. h_q only uses sequences in $x_0\Sigma^\star$, while traversal counting uses the entire sequence s, and is more accurate.

GenESeSS has no upper bound on the number of states; which is a function of the complexity of the process itself.

Theorem 4 (Time Complexity). *Asymptotic time complexity of* GenESeSS *is:*

$$\mathscr{T}\ O\big((1/\epsilon\ |Q|) \times |s| \times |\Sigma|\big) \tag{29}$$

Proof. GenESeSS performs the following computations:

C1 Computation of a derivative heap by computing $\phi^s x$ for $O1/\epsilon$ strings (Corollary 2), each of which involves reading the input s and normalization to distributions over Σ, thus contributing $O1/\epsilon \times |s| \times |\Sigma|$.

C2 Finding a vertex of the convex hull of the heap, which, at worst, involves inspecting $O\epsilon$ points (encoded by strings generating the heap), contributing $O1/\epsilon \times |\Sigma|$, where each inspection is done in $O|\Sigma|$ time.

C3 Finding δ, involving computing derivatives at string-identifiers (Step 2), thus contributing $O|Q| \times |s| \times |\Sigma|$. Strong connectivity [14] requires $O|Q|\ |\Sigma|$, and hence is unimportant.

C4 Identification of arc probabilities using traversal counts and normalization, done in time linear in the number of arcs, *i.e* $O|Q| \times |\Sigma|$.

Summing the contributions, and using $|s| > |\Sigma|$,

$$\mathscr{T}\ O1/\epsilon \times |s| \times |\Sigma|\ \ 1/\epsilon \times |\Sigma|\ \ |Q| \times |s| \times |\Sigma|\ \ |Q| \times |\Sigma|$$
$$O\big((1/\epsilon\ |Q|) \times |s| \times |\Sigma|\big)$$

□

Theorem 4 shows that GenESeSS is polynomial in $O1/\epsilon$, size of input s, model size $|Q|$, and alphabet size $|\Sigma|$. In practice, $|Q| \ll 1/\epsilon$, implying that

$$\mathscr{T}\ O\left(\frac{|s||\Sigma|}{\epsilon}\right) \tag{30}$$

An identification method is said to identify a target language $L_\star$ in the Probably Approximately Correct (PAC) sense [17, 1, 10], if it always halts and outputs L such that:

$$\exists \epsilon, \delta > 0, PdL_\star, L \leqq \epsilon \geqq 1 - \delta \tag{31}$$

where $d\cdot,\cdot$ is a metric on the space of target languages. A class of languages is efficiently PAC-learnable if there exists an algorithm that PAC-identifies every language in the class, and runs in time polynomial in $1/\epsilon$, $1/\delta$, length of sample input, and inferred model size. We prove PAC-learnability of QSPs, by first establishing a metric on the space of probabilistic automata over Σ.

2.3 PAC Identifiability Of QSPs

Lemma 4 (Metric For Probabilistic Automata). *For two strongly connected PFSAs G_1, G_2, denote the symbolic derivative at $x \in \Sigma^\star$ as $\phi^s_{G_1} x$ and $\phi^s_{G_2} x$ respectively. Then,*

$$\Theta G_1, G_2 \ \sup_{x \in \Sigma^\star} \left\{ Jx \ \lim_{|s_1| \to} \ \lim_{|s_2| \to} ||\phi^{s_1}_{G_1} x - \phi^{s_2}_{G_2} x|| \right\}$$

defines a metric on the space of probabilistic automata on Σ.

Proof. Non-negativity and symmetry follows immediately. Triangular inequality follows from noting that $||\phi^{s_1}_{G_1} x - \phi^{s_2}_{G_2} x||$ is upper bounded by 1, and therefore for any chosen order of the strings in $\Sigma^\star$, we have two ℓ sequences, which would satisfy the triangular inequality under the sup norm. The metric is well-defined since for any sufficiently long s_1, s_2, the symbolic derivatives at arbitrary x are uniformly convergent to some linear combination of the rows of the corresponding $\widetilde{\Pi}$ matrices. □

Theorem 5 (PAC-Learnability). *QSPs for which the probabilistic Nerode equivalence has a finite index is PAC-learnable using PFSAs, i.e., for $\epsilon, \eta > 0$, and for every sufficiently long sequence s generated by QSP $\mathcal{H}$, we can compute $\mathcal{P}'_{\mathcal{H}}$ as an estimate for $\mathcal{P}_{\mathcal{H}}$ such that:*

$$Prob\big(\Theta\mathcal{P}_{\mathcal{H}}, \mathcal{P}'_{\mathcal{H}} \leqq \epsilon\big) \geqq 1 - \eta \tag{32}$$

The algorithm runs in time polynomial in $1/\epsilon, 1/\eta$, input length $|s|$ and model size.

Proof. GenESeSS construction implies that, once the initial ϵ-synchronizing string x_0 is identified, there is no scope of the model error to be more than ϵ. Hence:

$$Prob\big(\Theta\mathcal{P}_{\mathcal{H}}, \mathcal{P}'_{\mathcal{H}} \leqq \epsilon\big) \;\; 1 - Prob\big(||\phi^s x_0 - \wp_{x_0}\widetilde{\Pi}|| > \epsilon\big)$$
$$\Rightarrow Prob\big(\Theta\mathcal{P}_{\mathcal{H}}, \mathcal{P}'_{\mathcal{H}} \leqq \epsilon\big) \geqq 1 - e^{-|s|\epsilon O_1} \quad \text{(Using Eq. (28))}$$

Thus, for any $\eta > 0$, if we have $|s| \;\; O1/\epsilon \log 1/\eta$, then the required condition of Eq. (32) is met. The polynomial runtimes are established in Theorem 4. □

2.3.1 Remark On Kearns' Hardness Result

We are immune to Kearns' hardness result [11], since $\epsilon > 0$ enforces state distinguishability [13].

3 Summary & Conclusion

We establish the notion of causal generators for stationary ergodic quantized stochastic processes on a rigorous measure-theoretic foundation, and propose a new inference algorithm GenESeSS for identification of probabilistic automata models from sufficiently long traces. We show that our approach can learn processes with long range dependencies, which yield non-synchronizable automata. Additionally, we establish that GenESeSS is computationally efficient in the PAC sense. The results are validated on synthetic and real datasets. Future research will investigate the prediction problem in greater detail in the context of new applications.

References

[1] Dana Angluin. Computational learning theory: survey and selected bibliography. In *Proceedings of the twenty-fourth annual ACM symposium on Theory of computing*, STOC '92, pages 351–369, New York, NY, USA, 1992. ACM.

[2] C. Bradford Barber, David P. Dobkin, and Hannu Huhdanpaa. The quickhull algorithm for convex hulls. *ACM Trans. Math. Softw.*, 22(4):469–483, dec 1996.

[3] S. Bogdanovic, B. Imreh, M. Ciric, and T. Petkovic. Directable automata and their generalizations - a survey. *Novi Sad Journal of Mathematics*, 29(2):31–74, 1999.

[4] Jorge Castro and Ricard Gavaldà. Towards feasible pac-learning of probabilistic deterministic finite automata. In Alexander Clark, François Coste, and Laurent Miclet, editors, *ICGI*, volume 5278 of *Lecture Notes in Computer Science*, pages 163–174. Springer, 2008.

[5] I. Chattopadhyay and A. Ray. Structural transformations of probabilistic finite state machines. *International Journal of Control*, 81(5):820–835, May 2008.

[6] I. Csiszár. Sanov property, generalized I-projection and a conditional limit theorem. *Ann. Probab.*, 12:768–793, 1984.

[7] Ricard Gavaldà, Philipp W. Keller, Joelle Pineau, and Doina Precup. Pac-learning of markov models with hidden state. In Johannes Fürnkranz, Tobias Scheffer, and Myra Spiliopoulou, editors, *ECML*, volume 4212 of *Lecture Notes in Computer Science*, pages 150–161. Springer, 2006.

[8] J. E. Hopcroft, R. Motwani, and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation, 2nd ed.* Addison-Wesley, 2001.

[9] M. Ito and Jürgen Duske. On cofinal and definite automata. *Acta Cybern.*, 6:181–189, 1984.

[10] M. J. Kearns and U. V. Vazirani. *An Introduction to Computational Learning Theory.* MIT Press, Cambridge, MA, USA, 1994.

[11] Michael Kearns, Yishay Mansour, Dana Ron, Ronitt Rubinfeld, Robert E. Schapire, and Linda Sellie. On the learnability of discrete distributions. In *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, STOC '94, pages 273–282, New York, NY, USA, 1994. ACM.

[12] E. L. Lehmann and Joseph P. Romano. *Testing statistical hypotheses.* Springer Texts in Statistics. Springer, New York, third edition, 2005.

[13] Dana Ron, Yoram Singer, and Naftali Tishby. Learning probabilistic automata with variable memory length. In *Computational Learing Theory*, pages 35–46, 1994.

[14] Robert Tarjan. Depth-first search and linear graph algorithms. *SIAM Journal on Computing*, 1(2):146–160, 1972.

[15] Flemming Topsøe. On the glivenko-cantelli theorem. *Probability Theory and Related Fields*, 14:239–250, 1970. 10.1007/BF01111419.

[16] A.B. Tsybakov. *Introduction to nonparametric estimation.* Springer series in statistics. Springer, 2009.

[17] L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27:1134–1142, November 1984.