Santa Fe Institute Working Paper 11-11-051
arxiv.org:1111.4500 [math.PR]

MCS Codes: 37A50 37B10 60J10

Equivalence of History and Generator ϵ-Machines

Nicholas F. Travers[1, 2, *] and James P. Crutchfield[1, 2, 3, 4, †]

[1]*Complexity Sciences Center*
[2]*Mathematics Department*
[3]*Physics Department*
University of California at Davis,
One Shields Avenue, Davis, CA 95616
[4]*Santa Fe Institute*
1399 Hyde Park Road, Santa Fe, NM 87501

(Dated: December 14, 2012)

ϵ-Machines are minimal, unifilar presentations of stationary stochastic processes. They were originally defined in the history machine sense, as hidden Markov models whose states are the equivalence classes of infinite pasts with the same probability distribution over futures. In analyzing synchronization, though, an alternative generator definition was given: unifilar, edge-emitting hidden Markov models with probabilistically distinct states. The key difference is that history ϵ-machines are defined by a process, whereas generator ϵ-machines define a process. We show here that these two definitions are equivalent in the finite-state case.

PACS numbers:
Keywords: hidden Markov model, history epsilon-machine, generator epsilon-machine, measure theory, synchronization, uniqueness

I. INTRODUCTION

Let $\mathcal{P} = (X_t)_{t\in\mathbb{Z}}$ be a stationary stochastic process. The ϵ-machine $M = M(\mathcal{P})$ for the process $\mathcal{P}$ is the hidden Markov model whose states consist of equivalence classes of infinite past sequences (histories) $\overleftarrow{x} = \ldots x_{-2}x_{-1}$ with the same probability distribution over future sequences $\overrightarrow{x} = x_0x_1\ldots$. The corresponding equivalence relation on the set of pasts $\overleftarrow{x}$ is denoted by $\sim_\epsilon$:

$$\overleftarrow{x} \sim_\epsilon \overleftarrow{x}' \text{ if } \mathbb{P}(\overrightarrow{X}|\overleftarrow{x}) = \mathbb{P}(\overrightarrow{X}|\overleftarrow{x}') \ , \tag{1}$$

where $\overrightarrow{X} = X_0X_1\ldots$ denotes the infinite sequence of future random variables.

These machines were first introduced in [1] as minimal, predictive models to measure the structural complexity of dynamical systems and have subsequently been applied in a number of contexts for nonlinear modeling [2–6]. Important extensions and a more thorough development of the theory were given in [7–10]. However, it was not until quite recently that the first fully formal construction was presented in [11].

Shortly thereafter, in our studies of synchronization [12, 13], we introduced an alternative "generator ϵ-machine" definition, in contrast to the original "history ϵ-machine" construction discussed above. A generator ϵ-machine is defined simply as a unifilar, edge-emitting hidden Markov model with probabilistically distinct states. As opposed to the history ϵ-machine $M_h = M_h(\mathcal{P})$ which is derived from a process $\mathcal{P}$, a generator ϵ-machine M_g itself defines a stationary process $\mathcal{P} = \mathcal{P}(M_g)$. Namely, the stationary output process of the hidden Markov model M_g obtained by choosing the initial state according to the stationary distribution π for states of the underlying Markov chain.

We establish, here, that the history and generator ϵ-machine definitions are equivalent in the finite state case. This has long been assumed, without formally specifying the generator definition. However, our work makes this explicit and gives one of the first formal proofs of equivalence.

The equivalence is also implicit in [11]; in fact, for a more general class of machines, not just finite-state. However, the techniques used there differ substantially from ours and use somewhat more machinery. In particular, the proof of equivalence in the more difficult direction of Theorem 1 (Section IV A) uses a supermartingale argument that, though elegant, relies implicitly on the martingale convergence theorem and is not particularly concrete. By contrast

*Electronic address: ntravers@math.ucdavis.edu
†Electronic address: chaos@ucdavis.edu

our proof of Theorem 1 follows directly from the synchronization results given in [12, 13], which are themselves fairly elementary, using only basic information theory and a large deviation estimate for finite-state Markov chains. Thus, the alternative proof presented here should be useful in providing intuition for the theorem. Also, since the definitions and terminology used in [11] differ significantly from ours, it is not immediately clear that the history-generator equivalence is what is shown there or is a consequence of what is shown. Thus, the exposition here should be helpful in clarifying these issues.

We note also that in order to parallel the generator ϵ-machine definition used in our synchronization studies and apply results from those works, we restrict the range of processes somewhat when defining history ϵ-machines. In particular, we assume when defining history ϵ-machines that the process $\mathcal{P}$ is not only stationary but also ergodic and that the process alphabet is finite. This is required for equivalence, since the output process of a generator ϵ-machine is of this form. However, neither of these assumptions is strictly necessary for history ϵ-machines. Only stationarity is actually needed. The history ϵ-machine definition can be extended to nonergodic stationary processes and countable or even more general alphabets [9, 11].

II. RELATED WORK

Since their introduction in the late 80s, most of the work on ϵ-machines, both theoretical and applied, has come from the physics and information theory perspectives. However, similar concepts have been around for some time in several other disciplines. Among others, there has been substantial work on related topics by both probabilists and automata theorists, as well as those in the symbolic dynamics community. Below, we review some of the most germane developments in these areas. The interested reader is also referred to [9, appendix H] where a very broad overview of such connections is given and to [14] for a recent review of the relation between symbolic dynamics and hidden Markov models in general.

We hope that our review provides context for the study of ϵ-machines and helps elucidate the relationship between ϵ-machines and other related models—both their similarities and their differences. However, an understanding of these relationships will not be necessary for the equivalence results that follow. The reader uninterested in these connections may safely skip to the definitions in Section III.

A. Sofic Shifts and Topological Presentations

Let $\mathcal{X}$ be a finite alphabet, and let $\mathcal{X}^{\mathbb{Z}}$ denote the set of all bi-infinite sequences $\overleftrightarrow{x} = \ldots x_{-1}x_0x_1\ldots$ consisting of symbols in $\mathcal{X}$. A subshift $\Sigma \subset \mathcal{X}^{\mathbb{Z}}$ is said to be *sofic* if it is the image of a subshift of finite type under a k-block factor map. This concept was first introduced in [15], where it also shown that any sofic shift Σ may be presented as a finite, directed graph with edges labeled by symbols in the alphabet $\mathcal{X}$. The allowed sequences $\overleftrightarrow{x} \in \Sigma$ consist of projections (under the edge labeling) of bi-infinite walks on the graph edges.

In the following, we will assume that all vertices in a presenting graph G of a sofic shift are *essential*. That is, each vertex v occurs as the target vertex of some edge e in a bi-infinite walk on the graph edges. If this is not the case, one may restrict to the graph G' consisting of essential vertices in G, along with their outgoing edges, and G' will also be a presenting graph for the sofic shift Σ. Thus, it is only necessary to consider presenting graphs in which all vertices are essential.

The *language* $\mathcal{L}(\Sigma)$ of a subshift Σ is the set of all finite words w occurring in some point $\overleftrightarrow{x} \in \Sigma$. For a sofic shift Σ with presenting graph G one may consider the nondeterministic finite automaton (NFA) M associated with the graph G, in which all states (vertices of G) are both start and accept states. Clearly (under the assumption that all vertices are essential) the language accepted by M is just $\mathcal{L}(\Sigma)$. Thus, the language of any sofic shift is regular. By standard algorithms (see e.g. [16]) one may obtain from M a unique, minimal, deterministic finite automaton (DFA) M' with the fewest number of states of all DFAs accepting the language $\mathcal{L}(\Sigma)$. We call M' the *minimal deterministic automaton* for the sofic shift Σ.

A subshift Σ is said to be *irreducible* if for any two words $w_1, w_2 \in \mathcal{L}(\Sigma)$ there exists $w_3 \in \mathcal{L}(\Sigma)$ such that the word $w_1w_3w_2 \in \mathcal{L}(\Sigma)$. As shown in [17], a sofic shift is irreducible if and only if it has some irreducible (i.e. strongly connected) presenting graph G.

A presenting graph G of a sofic shift Σ is said to be *unifilar* or *right-resolving* if for each vertex $v \in G$ and symbol $x \in \mathcal{X}$, there is at most one outgoing edge e from v labeled with the symbol x. A shown in [18], an irreducible sofic shift Σ always has a unique, minimal, unifilar presenting graph, that, it turns out, is also irreducible. In symbolic dynamics this presentation is often referred to as the (right) *Fischer cover* of Σ.

For an irreducible sofic shift Σ, the graph associated with the minimal deterministic automaton always has a single recurrent, irreducible component. This recurrent component is *isomorphic* to the Fischer cover. That is, there exists

a bijection between vertices in the automaton graph and vertices of the Fischer cover that preserves both edges and edge labels.

A related notion is the Krieger cover based on future sets [19]. For a subshift $\Sigma \subset \mathcal{X}^{\mathbb{Z}}$, let Σ^+ denote the set of allowed future sequences $\overrightarrow{x} = x_0 x_1 \ldots$ and let Σ^- be the set of allowed past sequences $\overleftarrow{x} = \ldots x_{-2} x_{-1}$. That is:

$$\Sigma^+ = \{\overrightarrow{x} : \exists \overleftarrow{x} \text{ with } \overleftarrow{x}\,\overrightarrow{x} \in \Sigma\} \text{ and } \Sigma^- = \{\overleftarrow{x} : \exists \overrightarrow{x} \text{ with } \overleftarrow{x}\,\overrightarrow{x} \in \Sigma\}.$$

Also, for a past $\overleftarrow{x} \in \Sigma^-$, let the *future set* $F(\overleftarrow{x})$ of $\overleftarrow{x}$ be the set of all possible future sequences $\overrightarrow{x}$ that can follow $\overleftarrow{x}$:

$$F(\overleftarrow{x}) = \{\overrightarrow{x} \in \Sigma^+ : \overleftarrow{x}\,\overrightarrow{x} \in \Sigma\}.$$

Define an equivalence relation $\sim_K$ on the set of infinite pasts $\overleftarrow{x} \in \Sigma^-$ by:

$$\overleftarrow{x} \sim_K \overleftarrow{x}' \text{ if } F(\overleftarrow{x}) = F(\overleftarrow{x}'). \tag{2}$$

The Krieger cover of Σ is the (possibly infinite) directed, edge-labeled graph G whose vertices consist of equivalence classes of pasts $\overleftarrow{x}$ under the relation $\sim_K$. There is a directed edge in G from vertex v to vertex v' labeled with symbol x, if for some past $\overleftarrow{x} \in v$ (equivalently all pasts $\overleftarrow{x} \in v$) the past $\overleftarrow{x}' = \overleftarrow{x}\,x \in v'$. By construction, the Krieger cover G is necessarily unifilar. Moreover, it is easily shown that G is a finite graph if and only if the subshift Σ is sofic.

If the subshift Σ is both irreducible and sofic, then the Krieger cover is isomorphic to the subgraph of the minimal deterministic automaton consisting of all states v that are not *finite-time transient* (and their outgoing edges). That is, the subgraph consisting of those states v such that there exist arbitrarily long words $w \in \mathcal{L}(\Sigma)$ on which the automaton transitions from its start state to v. Clearly, any state in the recurrent, irreducible component of the automaton graph is not finite-time transient. Thus, the Krieger cover contains this recurrent component—the Fischer cover.

To summarize, the minimal deterministic automaton, Fischer cover, and Krieger cover are three closely related ways for presenting an irreducible sofic shift that are each, in slightly different senses, minimal unifilar presentations. The Fischer cover is always an irreducible graph. The Krieger cover and graph of the minimal deterministic automaton are not necessarily irreducible, but they each have a single recurrent, irreducible component that is isomorphic to the Fischer cover. The Krieger cover itself is also isomorphic to a subgraph of the minimal deterministic automaton.

ϵ-Machines are a probabilistic extension of these purely topological presentations. More specifically, for a stationary process $\mathcal{P}$ the history ϵ-machine M_h is the probabilistic analog of the Krieger cover G for the subshift consisting of $\text{supp}(\mathcal{P})$. It is the edge-emitting hidden Markov model defined analogously to the Krieger cover, but with states that are equivalence classes of infinite past sequences $\overleftarrow{x}$ with the same probability distribution over future sequences $\overrightarrow{x}$, rather than simply the same set of allowed future sequences. (Compare Equations (1) and (2).)

In some cases the two presentations may be topologically equivalent—e.g., the history ϵ-machine and Krieger cover can be isomorphic as graphs when the transition probabilities are removed from edges of the ϵ-machine. In other cases, however, they are not. For example, for the Even Process (Example 1, Section III E) the Krieger cover (or at least its recurrent component, the Fischer cover) and the history ϵ-machine are topologically equivalent. But this is not so for the ABC process (Example 2, Section III E). In fact, there exist many examples of ergodic processes whose support is an irreducible sofic shift, but for which the history ϵ-machine has an infinite (or even continuum) number of states. See, e.g., Example 4 in Section III E, and Example 3.26 in [11].

B. Semigroup Measures

Semigroup measures are a class of probability measures on sofic shifts that arise from assigning probability transition structures to the right and left covers obtained from the Cayley graphs associated with generating semigroups for the shifts. These measures are studied extensively in [20], where a rich theory is developed and many of their key structural properties are characterized.

In particular, it is shown there that a stationary probability measure $\mathbb{P}$ on a sofic shift Σ is a semigroup measure if and only if it has a finite number of *future measures*—distributions over future sequences $\overrightarrow{x}$—induced by all finite-length past words w. That is, if there exist a finite number of finite length words $w_1, \ldots, w_N$ such that for any word w of positive probability:

$$\mathbb{P}(\overrightarrow{X}|w) = \mathbb{P}(\overrightarrow{X}|w_i) \ ,$$

for some $1 \leq i \leq N$, where $\overrightarrow{X}$ denotes the infinite sequence of future random variables X_t on Σ, defined by the natural projections $X_t(\overleftrightarrow{x}) = x_t$.

By contrast, a process $\mathcal{P}$ (or measure $\mathbb{P}$) has a finite-state history ϵ-machine if there exist a finite number of infinite past sequences $\overleftarrow{x}_1, \ldots, \overleftarrow{x}_N$ such that, for almost every infinite past $\overleftarrow{x}$:

$$\mathbb{P}(\overrightarrow{X}|\overleftarrow{x}) = \mathbb{P}(\overrightarrow{X}|\overleftarrow{x}_i) \ ,$$

for some $1 \leq i \leq N$. The latter condition is strictly more general. The Alternating Biased Coins Process described in Section III E, for instance, has a finite-state (2-state) ϵ-machine, but does not correspond to a semigroup measure.

Thus, unfortunately, though the theory of semigroup measures is quite rich and well developed, much of it does not apply for the measures we study. For this reason, our proof methods are quite different from those previously used for semigroup measures, despite the seeming similarity between the two settings

C. g-Functions and g-Measures

For a finite alphabet $\mathcal{X}$, let $\mathcal{X}^-$ denote the set of all infinite past sequences $\overleftarrow{x} = \ldots x_{-2}x_{-1}$ consisting of symbols in $\mathcal{X}$. A *g-function* for the full shift $\mathcal{X}^{\mathbb{Z}}$ is a map:

$$g : (\mathcal{X}^- \times \mathcal{X}) \to [0, 1] \ ,$$

such that for any $\overleftarrow{x} \in \mathcal{X}^-$:

$$\sum_{x \in \mathcal{X}} g(\overleftarrow{x}, x) = 1 \ .$$

A *g-measure* for a g-function g on the full shift $\mathcal{X}^{\mathbb{Z}}$ is stationary probability measure $\mathbb{P}$ on $\mathcal{X}^{\mathbb{Z}}$ that is consistent with the g-function g in that for $\mathbb{P}$ a.e. $\overleftarrow{x} \in \mathcal{X}^-$:

$$\mathbb{P}(X_0 = x|\overleftarrow{X} = \overleftarrow{x}) = g(\overleftarrow{x}, x), \text{ for each } x \in \mathcal{X}.$$

g-Functions and g-Measures have been studied for some time, though sometimes under different names [21–24]. In particular, many of these studies address when a g-function will or will not have a unique corresponding g-measure. Normally, g is assumed to be continuous (with respect to the natural product topology) and in this case, using fixed point theory, it can be shown that at least one g-measure exists. However, continuity is not enough to ensure uniqueness, even if some natural mixing conditions are required as well [23]. Thus, stronger conditions are often required, such as Hölder continuity.

Of particular relevance to us is the more recent work [25] on g-functions restricted to subshifts. It is shown there, in many instances, how to construct g-functions on subshifts with an infinite or even continuum number of future measures, subject to fairly strong requirements. For example, residual local constancy or a synchronization condition similar to the exactness condition introduced in [12]. Most surprising, perhaps, are the constructions of g-functions for irreducible subshifts, which themselves take only a finite number of values, but have unique associated g-measures with an infinite number of future measures.

The relation to ϵ-machines is the following. Given a g-function g, one may divide the set of infinite past sequences $\overleftarrow{x}$ into equivalence classes, in a manner analogous to that for history ϵ-machines, by the relation $\sim_g$:

$$\overleftarrow{x} \sim_g \overleftarrow{x}' \text{ if } g(\overleftarrow{x}, x) = g(\overleftarrow{x}', x), \text{ for all } x \in \mathcal{X} \ . \tag{3}$$

The equivalence classes induced by the relation $\sim_g$ of Equation 3 are coarser than those induced by the relation $\sim_\epsilon$ of Equation 1. For any g-measure $\mathbb{P}$ of the g-function g, the states of the history ϵ-machine are a refinement or splitting of the $\sim_g$ equivalence classes. Two infinite pasts $\overleftarrow{x}$ and $\overleftarrow{x}'$ that induce different probability distributions over the next symbol x_0 must induce different probability distributions over infinite future sequences $\overrightarrow{x}$, but the converse is not necessarily true. As shown in [25], the splitting may, in fact, be quite "bad" even if "nice" conditions are enforced on the g-function associated with the probability measure $\mathbb{P}$. Concretely, there exist processes with history ϵ-machines that have an infinite or even continuum number of states, but for which the associated "nice" g-function from which the process is derived has only a finite number of equivalence classes.

III. DEFINITIONS

In this section we set up the formal framework for our results and give more complete definitions for our objects of study: stationary processes, hidden Markov models, and ϵ-machines.

A. Processes

There are several ways to define a stochastic process. Perhaps the most traditional is simply as a sequence of random variables (X_t) on some common probability space Ω. However, in the following it will be convenient to use a slightly different, but equivalent, construction in which a process is itself a probability space whose sample space consists of bi-infinite sequences $\overleftrightarrow{x} = \ldots x_{-1}x_0x_1\ldots$. Of course, on this space we have random variables X_t defined by the natural projections $X_t(\overleftrightarrow{x}) = x_t$, which we will employ at times in our proofs. However, for most of our development and, in particular, for defining history ϵ-machines, it will be more convenient to adopt the sequence-space viewpoint.

Throughout, we restrict our attention to processes over a finite alphabet $\mathcal{X}$. We denote by $\mathcal{X}^*$ the set of all words w of finite positive length consisting of symbols in $\mathcal{X}$ and, for a word $w \in \mathcal{X}^*$, we write $|w|$ for its length. Note that we deviate slightly from the standard convention here and explicitly exclude the null word λ from $\mathcal{X}^*$.

Definition 1. *Let $\mathcal{X}$ be a finite set. A process $\mathcal{P}$ over the alphabet $\mathcal{X}$ is a probability space $(\mathcal{X}^{\mathbb{Z}}, \mathbb{X}, \mathbb{P})$ where:*

- *$\mathcal{X}^{\mathbb{Z}}$ is the set of all bi-infinite sequences of symbols in $\mathcal{X}$: $\mathcal{X}^{\mathbb{Z}} = \{\overleftrightarrow{x} = \ldots x_{-1}x_0x_1\ldots : x_t \in \mathcal{X}, \text{ for all } t \in \mathbb{Z}\}$.*
- *$\mathbb{X}$ is the σ-algebra generated by finite cylinder sets of the form $A_{w,t} = \{\overleftrightarrow{x} \in \mathcal{X}^{\mathbb{Z}} : x_t \ldots x_{t+|w|-1} = w\}$.*
- *$\mathbb{P}$ is a probability measure on the measurable space $(\mathcal{X}^{\mathbb{Z}}, \mathbb{X})$.*

For each symbol $x \in \mathcal{X}$, we assume implicitly that $\mathbb{P}(A_{x,t}) > 0$ for some $t \in \mathbb{N}$. Otherwise, the symbol x is useless and the process can be restricted to the alphabet $\mathcal{X}/\{x\}$. In the following, we will be primarily interested in stationary, ergodic processes.

Let $r : \mathcal{X}^{\mathbb{Z}} \to \mathcal{X}^{\mathbb{Z}}$ be the right shift operator. A process $\mathcal{P}$ is *stationary* if the measure $\mathbb{P}$ is shift invariant: $\mathbb{P}(A) = \mathbb{P}(r(A))$ for any measurable set A. A process $\mathcal{P}$ is *ergodic* if every shift invariant event A is trivial. That is, for any measurable event A such that A and $r(A)$ are $\mathbb{P}$ a.s. equal, the probability of A is either 0 or 1. A stationary process $\mathcal{P}$ is defined entirely by the word probabilities $\mathbb{P}(w)$, $w \in \mathcal{X}^*$, where $\mathbb{P}(w) = \mathbb{P}(A_{w,t})$ is the shift invariant probability of cylinder sets for the word w. Ergodicity is equivalent to the almost sure convergence of empirical word probabilities $\widehat{\mathbb{P}}(w)$ in finite sequences $\overrightarrow{x}^t = x_0x_1\ldots x_{t-1}$ to their true values $\mathbb{P}(w)$, as $t \to \infty$.

For a stationary process $\mathcal{P}$ and words $w, v \in \mathcal{X}^*$ with $\mathbb{P}(v) > 0$, we define $\mathbb{P}(w|v)$ as the probability that the word w is followed by the word v in a bi-infinite sequence $\overleftrightarrow{x}$:

$$\begin{aligned}\mathbb{P}(w|v) &\equiv \mathbb{P}(A_{w,0}|A_{v,-|v|}) \\ &= \mathbb{P}(A_{v,-|v|} \cap A_{w,0})/\mathbb{P}(A_{v,-|v|}) \\ &= \mathbb{P}(vw)/\mathbb{P}(v) \ . \end{aligned} \tag{4}$$

The following facts concerning word probabilities and conditional word probabilities for a stationary process come immediately from the definitions. They will be used repeatedly throughout our development, without further mention. For any words $u, v, w \in \mathcal{X}^*$:

1. $\sum_{x\in\mathcal{X}} \mathbb{P}(wx) = \sum_{x\in\mathcal{X}} \mathbb{P}(xw) = \mathbb{P}(w)$;
2. $\mathbb{P}(w) \geq \mathbb{P}(wv)$ and $\mathbb{P}(w) \geq \mathbb{P}(vw)$;
3. If $\mathbb{P}(w) > 0$, $\sum_{x\in\mathcal{X}} \mathbb{P}(x|w) = 1$;
4. If $\mathbb{P}(u) > 0$, $\mathbb{P}(v|u) \geq \mathbb{P}(vw|u)$; and
5. If $\mathbb{P}(u) > 0$ and $\mathbb{P}(uv) > 0$, $\mathbb{P}(vw|u) = \mathbb{P}(v|u) \cdot \mathbb{P}(w|uv)$.

B. Hidden Markov Models

There are two primary types of hidden Markov models: *state-emitting* (or *Moore*) and *edge-emitting* (or *Mealy*). The state-emitting type is the simpler of the two and, also, the more commonly studied and applied [26, 27]. However, we focus on edge-emitting hidden Markov models here, since ϵ-machines are edge-emitting. We also restrict to the case where the hidden Markov model has a finite number of states and output symbols, although generalizations to countably infinite and even uncountable state sets and output alphabets are certainly possible.

Definition 2. *An edge-emitting* hidden Markov model (HMM) *is a triple $(\mathcal{S}, \mathcal{X}, \{T^{(x)}\})$ where:*

- *$\mathcal{S}$ is a finite set of states,*

- $\mathcal{X}$ *is a finite alphabet of output symbols, and*
- $T^{(x)}, x \in \mathcal{X}$ *are symbol-labeled transition matrices.* $T^{(x)}_{\sigma\sigma'} \geq 0$ *represents the probability of transitioning from state* σ *to state* σ' *on symbol* x*.*

In what follows, we normally take the state set to be $\mathcal{S} = \{\sigma_1, \ldots, \sigma_N\}$ and denote $T^{(x)}_{\sigma_i\sigma_j}$ simply as $T^{(x)}_{ij}$. We also denote the overall state-to-state transition matrix for an HMM as T: $T = \sum_{x\in\mathcal{X}} T^{(x)}$. T_{ij} is the overall probability of transitioning from state σ_i to state σ_j, regardless of symbol. The matrix T is stochastic: $\sum_{j=1}^{N} T_{ij} = 1$, for each i.

Pictorially, an HMM can be represented as a directed graph with labeled edges. The vertices are the states $\sigma_1, \ldots, \sigma_N$ and, for each i, j, x with $T^{(x)}_{ij} > 0$, there is a directed edge from state σ_i to state σ_j labeled $p|x$ for the symbol x and transition probability $p = T^{(x)}_{ij}$. The transition probabilities are normalized so that their sum on all outgoing edges from each state σ_i is 1.

Example. *Even Machine*

Figure 1 depicts an HMM for the Even Process. The support for this process consists of all binary sequences in which blocks of uninterrupted 1s are even in length, bounded by 0s. After each even length is reached, there is a probability p of breaking the block of 1s by inserting a 0. The HMM has two states $\{\sigma_1, \sigma_2\}$ and symbol-labeled transitions matrices:

$$T^{(0)} = \begin{pmatrix} p & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad T^{(1)} = \begin{pmatrix} 0 & 1-p \\ 1 & 0 \end{pmatrix}$$

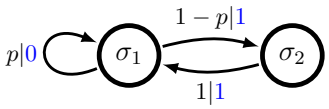

FIG. 1: A hidden Markov model (the ϵ-machine) for the Even Process. The HMM has two internal states $\mathcal{S} = \{\sigma_1, \sigma_2\}$, a two-symbol alphabet $\mathcal{X} = \{0, 1\}$, and a single parameter $p \in (0, 1)$ that controls the transition probabilities.

The operation of an HMM may be thought of as a weighted random walk on the associated directed graph. That is, from the current state σ_i, the next state σ_j is determined by selecting an outgoing edge from σ_i according to their relative probabilities. Having selected a transition, the HMM then moves to the new state and outputs the symbol x labeling this edge. The same procedure is then invoked repeatedly to generate future states and output symbols.

The state sequence determined in such a fashion is simply a Markov chain with transition matrix T. However, we are interested not simply in the HMM's state sequence, but rather the associated sequence of output symbols it generates. We assume that an observer of the HMM has direct access to this sequence of output symbols, but not to the associated sequence of "hidden" states.

Formally, from an initial state σ_i the probability that the HMM next outputs symbol x and transitions to state σ_j is:

$$\mathbf{P}_{\sigma_i}(x, \sigma_j) = T^{(x)}_{ij} \ . \tag{5}$$

And, the probability of longer sequences is computed inductively. Thus, for an initial state $\sigma_i = \sigma_{i_0}$ the probability the HMM outputs a length-l word $w = w_0 \ldots w_{l-1}$ while following the *state path* $s = \sigma_{i_1} \ldots \sigma_{i_l}$ in the next l steps is:

$$\mathbf{P}_{\sigma_i}(w, s) = \prod_{t=0}^{l-1} T^{(w_t)}_{i_t, i_{t+1}} \ . \tag{6}$$

If the initial state is chosen according to some distribution $\rho = (\rho_1, \ldots, \rho_N)$ rather than as a fixed state σ_i, we have by linearity:

$$\mathbf{P}_{\rho}(x, \sigma_j) = \sum_i \rho_i \cdot \mathbf{P}_{\sigma_i}(x, \sigma_j) \text{ and} \tag{7}$$

$$\mathbf{P}_{\rho}(w, s) = \sum_i \rho_i \cdot \mathbf{P}_{\sigma_i}(w, s) \ . \tag{8}$$

The overall probabilities of next generating a symbol x or word $w = w_0 \ldots w_{l-1}$ from a given state σ_i are computed by summing over all possible associated target states or state sequences:

$$\mathbf{P}_{\sigma_i}(x) = \sum_j \mathbf{P}_{\sigma_i}(x, \sigma_j) = \|e_i T^{(x)}\|_1 \text{ and} \tag{9}$$

$$\mathbf{P}_{\sigma_i}(w) = \sum_{\{s:|s|=l\}} \mathbf{P}_{\sigma_i}(w, s) = \|e_i T^{(w)}\|_1 \ , \tag{10}$$

respectively, where $e_i = (0, \ldots, 1, \ldots, 0)$ is the i^{th} standard basis vector in $\mathbb{R}^N$ and

$$T^{(w)} = T^{(w_0 \ldots w_{l-1})} \equiv \prod_{t=0}^{l-1} T^{(w_t)} \ . \tag{11}$$

Finally, the overall probabilities of next generating a symbol x or word $w = w_0 \ldots w_{l-1}$ from an initial state distribution ρ are, respectively:

$$\mathbf{P}_\rho(x) = \sum_i \rho_i \cdot \mathbf{P}_{\sigma_i}(x) = \|\rho T^{(x)}\|_1 \text{ and} \tag{12}$$

$$\mathbf{P}_\rho(w) = \sum_i \rho_i \cdot \mathbf{P}_{\sigma_i}(w) = \|\rho T^{(w)}\|_1 \ . \tag{13}$$

If the graph G associated with a given HMM is strongly connected, then the corresponding Markov chain over states is irreducible and the state-to-state transition matrix T has a unique *stationary distribution* π satisfying $\pi = \pi T$ [28]. In this case, we may define a stationary process $\mathcal{P} = (\mathcal{X}^{\mathbb{Z}}, \mathbb{X}, \mathbb{P})$ by the word probabilities obtained from choosing the initial state according to π. That is, for any word $w \in \mathcal{X}^*$:

$$\mathbb{P}(w) \equiv \mathbf{P}_\pi(w) = \|\pi T^{(w)}\|_1 \ . \tag{14}$$

Strong connectivity also implies the process $\mathcal{P}$ is ergodic, as it is a pointwise function of the irreducible Markov chain over edges, which is itself ergodic [28]. That is, at each time step the symbol labeling the edge is a deterministic function of the edge.

We denote the corresponding (stationary, ergodic) process over bi-infinite symbol-state sequences $(\overleftrightarrow{x}, \overleftrightarrow{s})$ by $\widetilde{\mathcal{P}}$. That is, $\widetilde{\mathcal{P}} = ((\mathcal{X} \text{ x } \mathcal{S})^{\mathbb{Z}}, (\mathbb{X} \text{ x } \mathbb{S}), \widetilde{\mathbb{P}})$ where:

1. $(\mathcal{X} \text{ x } \mathcal{S})^{\mathbb{Z}} = \{(\overleftrightarrow{x}, \overleftrightarrow{s}) \cong (x_t, s_t)_{t \in \mathbb{Z}} : x_t \in \mathcal{X} \text{ and } s_t \in \mathcal{S}, \text{ for all } t \in \mathbb{Z}\}$.
2. $(\mathbb{X} \text{ x } \mathbb{S})$ is the σ-algebra generated by finite cylinder sets on the bi-infinite symbol-state sequences.
3. The (stationary) probability measure $\widetilde{\mathbb{P}}$ on $(\mathbb{X} \text{ x } \mathbb{S})$ is defined by Equation (8) with $\rho = \pi$. Specifically, for any length-l word w and length-l state sequence s we have:

   $$\widetilde{\mathbb{P}}(\{(\overleftrightarrow{x}, \overleftrightarrow{s}) : x_0 \ldots x_{l-1} = w, s_1 \ldots s_l = s\}) = \mathbf{P}_\pi(w, s).$$

   By stationarity, this measure may be extended uniquely to all finite cylinders and, hence, to all $(\mathbb{X} \text{ x } \mathbb{S})$-measurable sets. And, it is consistent with the measure $\mathbb{P}$ in that:

   $$\widetilde{\mathbb{P}}(\{(\overleftrightarrow{x}, \overleftrightarrow{s}) : x_0 \ldots x_{l-1} = w\}) = \mathbb{P}(w) \ ,$$

   for all $w \in \mathcal{X}^*$.

Two HMMs are said to be *isomorphic* if there is a bijection between their state sets that preserves edges, including the symbols and probabilities labeling the edges. Clearly, any two isomorphic, irreducible HMMs generate the same process, but the converse is not true. Nonisomorphic HMMs may also generate equivalent processes. In Section IV we will be concerned with isomorphism between generator and history ϵ-machines.

C. Generator ϵ-Machines

Generator ϵ-machines are irreducible HMMs with two additional important properties: unifilarity and probabilistically distinct states.

Definition 3. *A* generator ϵ-machine M_g *is an HMM with the following properties:*

1. Irreducibility*: The graph G associated with the HMM is strongly connected.*
2. Unifilarity*: For each state $\sigma_i \in \mathcal{S}$ and each symbol $x \in \mathcal{X}$ there is at most one outgoing edge from state σ_i labeled with symbol x.*
3. Probabilistically distinct states*: For each pair of distinct states $\sigma_i, \sigma_j \in \mathcal{S}$ there exists some word $w \in \mathcal{X}^*$ such that $\mathbf{P}_{\sigma_i}(w) \neq \mathbf{P}_{\sigma_j}(w)$.*

Note that all three of these properties may be easily checked for a given HMM. Irreducibility and unifilarity are immediate. The probabilistically distinct states condition can (if necessary) be checked by inductively separating distinct pairs with an algorithm similar to the one used to check for topologically distinct states in [12].

By irreducibility, there is always a unique stationary distribution π over the states of a generator ϵ-machine, so we may associate to each generator ϵ-machine M_g a unique stationary, ergodic process $\mathcal{P} = \mathcal{P}(M_g)$ with word probabilities defined as in Equation (14). We refer to $\mathcal{P}$ as *the process* generated by the generator ϵ-machine M_g. The transition function for a generator ϵ-machine or, more generally, any unifilar HMM is denoted by δ. That is, for i and x with $\mathbf{P}_{\sigma_i}(x) > 0$, $\delta(\sigma_i, x) \equiv \sigma_j$, where σ_j is the (unique) state to which state σ_i transitions on symbol x.

In a unifilar HMM, for any given initial state σ_i and word $w = w_0 \dots w_{l-1} \in \mathcal{X}^*$, there can be at most one associated state path $s = s_1 \dots s_l$ such that the word w may be generated following the state path s from σ_i. Moreover, the probability $\mathbf{P}_{\sigma_i}(w)$ of generating w from σ_i is nonzero if and only if there is such a path s. In this case, the states $s_1, \dots, s_l$ are defined inductively by the relations $s_{t+1} = \delta(s_t, w_t), 0 \leq t \leq l-1$ with $s_0 = \sigma_i$, and the probability $\mathbf{P}_{\sigma_i}(w)$ is simply:

$$\mathbf{P}_{\sigma_i}(w) = \prod_{t=0}^{l-1} \mathbf{P}_{s_t}(w_t). \tag{15}$$

Slightly more generally, Equation (15) holds as long as there is a well defined path $s_1 \dots s_{l-1}$ upon which the subword $w_0 \dots w_{l-2}$ may be generated starting in σ_i. Though, in this case $\mathbf{P}_{\sigma_i}(w)$ may be 0 if state s_{l-1} has no outgoing transition on symbol w_{l-1}. This formula for word probabilities in unifilar HMMs will be useful in establishing the equivalence of generator and history ϵ-machines in Section IV.

D. History ϵ-Machines

The history ϵ-machine M_h for a stationary process $\mathcal{P}$ is, essentially, just the hidden Markov model whose states are the equivalence classes of infinite past sequences defined by the equivalence relation $\sim_\epsilon$ of Equation (1). Two pasts $\overleftarrow{x}$ and $\overleftarrow{x}'$ are considered equivalent if they induce the same probability distribution over future sequences. However, it takes some effort to make this notion precise and specify the transitions. The formal definition itself is quite lengthy, so for clarity verification of many technicalities is deferred to the appendices. We recommend first reading through this section in its entirety without reference to the appendices for an overview and, then, reading through the appendices separately afterward for the details. The appendices are entirely self contained in that, except for the notation introduced here, none of the results derived in the appendices relies on the development in this section. As noted before, our focus is restricted to ergodic, finite-alphabet processes to parallel the generator definition. Although, neither of these requirements is strictly necessary. Only stationarity is actually needed.

Let $\mathcal{P} = (\mathcal{X}^{\mathbb{Z}}, \mathbb{X}, \mathbb{P})$ be a stationary, ergodic process over a finite alphabet $\mathcal{X}$, and let $(\mathcal{X}^-, \mathbb{X}^-, \mathbb{P}^-)$ be the corresponding probability space over past sequences $\overleftarrow{x}$. That is:

- $\mathcal{X}^-$ is the set of infinite past sequences of symbols in $\mathcal{X}$: $\mathcal{X}^- = \{\overleftarrow{x} = \dots x_{-2}x_{-1} : x_t \in \mathcal{X}, t = -1, -2, \dots\}$.
- $\mathbb{X}^-$ is the σ-algebra generated by finite cylinder sets on past sequences: $\mathbb{X}^- = \sigma\left(\bigcup_{t=1}^{\infty} \mathbb{X}_t^-\right)$, where $\mathbb{X}_t^- = \sigma\left(\{A_w^- : |w| = t\}\right)$ and $A_w^- = \{\overleftarrow{x} = \dots x_{-2}x_{-1} : x_{-|w|} \dots x_{-1} = w\}$.
- $\mathbb{P}^-$ is the probability measure on the measurable space $(\mathcal{X}^-, \mathbb{X}^-)$ which is the projection of $\mathbb{P}$ to past sequences: $\mathbb{P}^-(A_w^-) = \mathbb{P}(w)$ for each $w \in \mathcal{X}^*$.

For a given past $\overleftarrow{x} \in \mathcal{X}^-$, we denote the last t symbols of $\overleftarrow{x}$ as $\overleftarrow{x}^t = x_{-t} \dots x_{-1}$. A past $\overleftarrow{x} \in \mathcal{X}^-$ is said to be *trivial* if $\mathbb{P}(\overleftarrow{x}^t) = 0$ for some finite t and *nontrivial* otherwise. If a past $\overleftarrow{x}$ is nontrivial, then for each $w \in \mathcal{X}^*$ $\mathbb{P}(w|\overleftarrow{x}^t)$ is well defined for each t, Equation (4), and one may consider $\lim_{t\to\infty} \mathbb{P}(w|\overleftarrow{x}^t)$. A nontrivial past $\overleftarrow{x}$ is said to be *w-regular* if $\lim_{t\to\infty} \mathbb{P}(w|\overleftarrow{x}^t)$ exists and *regular* if it is w-regular for each $w \in \mathcal{X}^*$. Appendix A shows that

the set of trivial pasts $\mathcal{T}$ is a null set and that the set of regular pasts $\mathcal{R}$ has full measure. That is, $\mathbb{P}^-(\mathcal{T}) = 0$ and $\mathbb{P}^-(\mathcal{R}) = 1$.

For a word $w \in \mathcal{X}^*$ the function $\mathbf{P}(w|\cdot) : \mathcal{R} \to \mathbb{R}$ is defined by:

$$\mathbf{P}(w|\overleftarrow{x}) \equiv \lim_{t\to\infty} \mathbb{P}(w|\overleftarrow{x}^t) \ . \tag{16}$$

Intuitively, $\mathbf{P}(w|\overleftarrow{x})$ is the conditional probability of w given $\overleftarrow{x}$. However, this probability is technically not well defined in the sense of Equation (4), since the probability of each past $\overleftarrow{x}$ is normally 0. And, we do not want to define $\mathbf{P}(w|\overleftarrow{x})$ in terms of a formal conditional expectation, because such a definition is only unique up to a.e. equivalence, while we would like its value on individual pasts to be uniquely determined. Nevertheless, intuitively speaking, $\mathbf{P}(w|\overleftarrow{x})$ is the conditional probability of w given $\overleftarrow{x}$, and this intuition should be kept in mind as it will provide understanding for what follows. Indeed, if one does consider the conditional probability $\mathbb{P}(w|\overleftarrow{X})$ as a formal conditional expectation, any version of it will be equal to $\mathbf{P}(w|\overleftarrow{x})$ for a.e. $\overleftarrow{x}$. So, this intuition is justified.

The central idea in the construction of the history ϵ-machine is the following equivalence relation on the set of regular pasts:

$$\overleftarrow{x} \sim \overleftarrow{x}' \text{ if } \mathbf{P}(w|\overleftarrow{x}) = \mathbf{P}(w|\overleftarrow{x}') \ , \quad \text{for all } w \in \mathcal{X}^* \ . \tag{17}$$

That is, two pasts $\overleftarrow{x}$ and $\overleftarrow{x}'$ are $\sim$ equivalent if their predictions are the same: Conditioning on either past leads to the same probability distribution over future words of all lengths. This is simply a more precise definition of the equivalence relation $\sim_\epsilon$ of Equation (1). (We drop the subscript ϵ, as this is the only equivalence relation we will consider from here on.)

The set of equivalence classes of regular pasts under the relation $\sim$ is denoted as $\mathcal{E} = \{E_\beta, \beta \in B\}$, where B is simply an index set. In general, there may be finitely many, countably many, or uncountably many such equivalence classes. Examples with $\mathcal{E}$ finite and countably infinite are given in Section III E. For uncountable $\mathcal{E}$, see example 3.26 in [11].

For an equivalence class $E_\beta \in \mathcal{E}$ and word $w \in \mathcal{X}^*$ we define the probability of w given E_β as:

$$\mathbf{P}(w|E_\beta) \equiv \mathbf{P}(w|\overleftarrow{x}) \ , \overleftarrow{x} \in E_\beta. \tag{18}$$

By construction of the equivalence classes this definition is independent of the representative $\overleftarrow{x} \in E_\beta$, and Appendix B shows that these probabilities are normalized, so that for each equivalence class E_β:

$$\sum_{x\in\mathcal{X}} \mathbf{P}(x|E_\beta) = 1 \ . \tag{19}$$

Appendix B also shows that the equivalence-class-to-equivalence-class transitions for the relation $\sim$ are well defined in that:

1. For any regular past $\overleftarrow{x}$ and symbol $x \in \mathcal{X}$ with $\mathbf{P}(x|\overleftarrow{x}) > 0$, the past $\overleftarrow{x}\,x$ is also a regular.
2. If $\overleftarrow{x}$ and $\overleftarrow{x}'$ are two regular pasts in the same equivalence class E_β and $\mathbf{P}(x|E_\beta) > 0$, then the two pasts $\overleftarrow{x}\,x$ and $\overleftarrow{x}'x$ must also be in the same equivalence class.

So, for each $E_\beta \in \mathcal{E}$ and $x \in \mathcal{X}$ with $\mathbf{P}(x|E_\beta) > 0$ there is a unique equivalence class $E_\alpha = \delta_h(E_\beta, x)$ to which equivalence class E_β transitions on symbol x.

$$\delta_h(E_\beta, x) \equiv E_\alpha, \text{ where } \overleftarrow{x}\,x \in E_\alpha \text{ for } \overleftarrow{x} \in E_\beta. \tag{20}$$

By point 2 above, this definition is again independent of the representative $\overleftarrow{x} \in E_\beta$.

The subscript h in δ_h indicates that it is a transition function between equivalence classes of pasts, or histories, $\overleftarrow{x}$. Formally, it is to be distinguished from the transition function δ between the states of a unifilar HMM. However, the two are essentially equivalent for a history ϵ-machine.

Appendix C shows that each equivalence class E_β is an $\mathbb{X}^-$ measurable set, so we can meaningfully assign a probability:

$$\begin{aligned} \mathbb{P}(E_\beta) &\equiv \mathbb{P}^-(\{\overleftarrow{x} \in E_\beta\}) \\ &= \mathbb{P}(\{\overleftrightarrow{x} = \overleftarrow{x}\,\overrightarrow{x} : \overleftarrow{x} \in E_\beta\}) \end{aligned} \tag{21}$$

to each equivalence class E_β. We say a process $\mathcal{P}$ is *finitely characterized* if there are a finite number of positive probability equivalence classes $E_1, \ldots, E_N$ that together comprise a set of full measure: $\mathbb{P}(E_i) > 0$ for each $1 \le i \le N$

and $\sum_{i=1}^{N} \mathbb{P}(E_i) = 1$. For a finitely characterized process $\mathcal{P}$ we will also occasionally say, by a slight abuse of terminology, that $\mathcal{E}^+ \equiv \{E_1, \ldots, E_N\}$ *is* the set of equivalence classes of pasts and ignore the remaining measure-zero subset of equivalence classes.

Appendix E shows that for any finitely characterized process $\mathcal{P}$, the transitions from the positive probability equivalence classes $E_i \in \mathcal{E}^+$ all go to other positive probability equivalence classes. That is, if $E_i \in \mathcal{E}^+$ then:

$$\delta_h(E_i, x) \in \mathcal{E}^+, \text{ for all } x \text{ with } \mathbf{P}(x|E_i) > 0. \tag{22}$$

As such, we define symbol-labeled transition matrices $T^{(x)}, x \in \mathcal{X}$ between the equivalence classes $E_i \in \mathcal{E}^+$. A component $T^{(x)}_{ij}$ of the matrix $T^{(x)}$ gives the probability that equivalence class E_i transitions to equivalence class E_j on symbol x:

$$T^{(x)}_{ij} = \mathbf{P}(E_i \xrightarrow{x} E_j) \equiv I(x, i, j) \cdot \mathbf{P}(x|E_i) \ , \tag{23}$$

where $I(x, i, j)$ is the indicator function of the transition from E_i to E_j on symbol x:

$$I(x, i, j) = \begin{cases} 1 & \text{if } \mathbf{P}(x|E_i) > 0 \text{ and } \delta_h(E_i, x) = E_j, \\ 0 & \text{otherwise.} \end{cases} \tag{24}$$

It follows from Equations (19) and (22) that the matrix $T \equiv \sum_{x \in \mathcal{X}} T^{(x)}$ is stochastic. (See also Claim 17 in Appendix E.)

Definition 4. *Let $\mathcal{P} = (\mathcal{X}^{\mathbb{Z}}, \mathbb{X}, \mathbb{P})$ be a finitely characterized, stationary, ergodic, finite-alphabet process. The history ϵ-machine $M_h(\mathcal{P})$ is defined as the triple $(\mathcal{E}^+, \mathcal{X}, \{T^{(x)}\})$.*

Note that M_h is a valid HMM since T is stochastic.

E. Examples

In this section we present several examples of irreducible HMMs and the associated ϵ-machines for the processes that these HMMs generate. This should hopefully provide some useful intuition for the definitions. For the sake of brevity, descriptions of the history ϵ-machine constructions in our examples will be less detailed than in the formal definition given above, but the ideas should be clear. In all cases, the process alphabet is the binary alphabet $\mathcal{X} = \{0, 1\}$.

Example 1. *Even Machine*

The first example we consider, shown in Figure 2, is the generating HMM M for the Even Process previously introduced in Section III B. It is easily seen that this HMM is both irreducible and unifilar and, also, that it has probabilistically distinct states. State σ_1 can generate the symbol 0, whereas state σ_2 cannot. M is therefore a generator ϵ-machine, and by Theorem 1 below the history ϵ-machine M_h for the process $\mathcal{P}$ that M generates is isomorphic to M. The Fischer cover for the sofic shift $\text{supp}(\mathcal{P})$ is also isomorphic to M, if probabilities are removed from the edge labels in M.

FIG. 2: The Even Machine M (left) and associated history ϵ-machine M_h (right) for the process $\mathcal{P}$ generated by M. $p \in (0, 1)$ is a parameter.

More directly, the history ϵ-machine states for $\mathcal{P}$ can be deduced by noting that 0 is a *synchronizing word* for M [12]: It synchronizes the observer to state σ_1. Thus, for any nontrivial past $\overleftarrow{x}$ terminating in $x_{-1} = 0$, the initial state s_0 must be σ_1. By unifilarity, any nontrivial past $\overleftarrow{x}$ terminating in a word of the form 01^n for some $n \geq 0$ also uniquely determines the initial state s_0. For n even, we must have $s_0 = \sigma_1$ and, for n odd, we must have $s_0 = \sigma_2$. Since

a.e. infinite past $\overleftarrow{x}$ generated by M contains at least one 0 and the distributions over future sequences $\overrightarrow{x}$ are distinct for the two states σ_1 and σ_2, the process $\mathcal{P}$ is finitely characterized with exactly two positive probability equivalence classes of infinite pasts: $E_1 = \{\overleftarrow{x} = \ldots 01^n : n \text{ is even }\}$ and $E_2 = \{\overleftarrow{x} = \ldots 01^n : n \text{ is odd }\}$. These correspond to the states σ_1 and σ_2 of M, respectively. More generally, a similar argument holds for any *exact* generator ϵ-machine. That is, any generator ϵ-machine having a finite synchronizing word w [12].

Example 2. *Alternating Biased Coins Machine*

Figure 3 depicts a generating HMM M for the Alternating Biased Coins (ABC) Process. This process may be thought of as being generated by alternately flipping two coins with different biases $p \neq q$. The phase—p-bias on odd flips or p-bias on even flips—is chosen uniformly at random. M is again, by inspection, a generator ϵ-machine: irreducible and unifilar with probabilistically distinct states. Therefore, by Theorem 1 below, the history ϵ-machine M_h for the process $\mathcal{P}$ that M generates is again isomorphic to M. However, the Fischer cover for the sofic shift $\text{supp}(\mathcal{P})$ is not isomorphic to M. The support of $\mathcal{P}$ is the full shift $\mathcal{X}^{\mathbb{Z}}$, so the Fischer cover consists of a single state transitioning to itself on both symbols 0 and 1.

FIG. 3: The Alternating Biased Coins (ABC) Machine M (left) and associated history ϵ-machine M_h (right) for the process $\mathcal{P}$ generated by M. $p, q \in (0, 1)$ are parameters, $p \neq q$.

In this simple example, the history ϵ-machine states can also be deduced directly, despite the fact that the generator M does not have a synchronizing word. If the initial state is $s_0 = \sigma_1$, then by the strong law of large numbers the limiting fraction of 1s at odd time steps in finite-length past blocks $\overleftarrow{x}^t$ converges a.s. to q. Whereas, if the initial state is $s_0 = \sigma_2$, then the limiting fraction of 1s at odd time steps converges a.s. to p. Therefore, the initial state s_0 can be inferred a.s. from the complete past $\overleftarrow{x}$, so the process $\mathcal{P}$ is finitely characterized with two positive probability equivalence classes of infinite pasts E_1 and E_2, corresponding to the two states σ_1 and σ_2. Unlike the exact case, however, arguments like this do not generalize as easily to other nonexact generator ϵ-machines.

Example 3. *Nonminimal Noisy Period-2 Machine*

Figure 4 depicts a nonminimal generating HMM M for the Noisy Period-2 (NP2) Process $\mathcal{P}$ in which 1s alternate with random symbols. M is again unifilar, but it does not have probabilistically distinct states and is, therefore, not a generator ϵ-machine. States σ_1 and σ_3 have the same probability distribution over future output sequences as do states σ_2 and σ_4.

There are two positive probability equivalence classes of pasts $\overleftarrow{x}$ for the process $\mathcal{P}$: Those containing 0s at a subset of the odd time steps, and those containing 0s at a subset of the even time steps. Those with 0s at odd time steps induce distributions over future output equivalent to that from states σ_2 and σ_4. While those with 0s at even time steps induce distributions over future output equivalent to that from states σ_1 and σ_3. Thus, the ϵ-machine for $\mathcal{P}$ consists of just two states $E_1 \sim \{\sigma_1, \sigma_3\}$ and $E_2 \sim \{\sigma_2, \sigma_4\}$. In general, for a unifilar HMM without probabilistically distinct states the ϵ-machine is formed by grouping together equivalent states in a similar fashion.

Example 4. *Simple Nonunifilar Source*

Figure 5 depicts a generating HMM M known as the Simple Nonunifilar Source (SNS) [7]. The output process $\mathcal{P}$ generated by M consists of long sequences of 1s broken by isolated 0s. As its name indicates, M is nonunifilar, so it is not an ϵ-machine.

Symbol 0 is a synchronizing word for M, so all pasts $\overleftarrow{x}$ ending in a 0 induce the same probability distribution over future output sequences $\overrightarrow{x}$: Namely, the distribution over futures given by starting M in the initial state $s_0 = \sigma_1$. However, since M is nonunifilar, an observer does not remain synchronized after seeing a 0. Any nontrivial past of the form $\overleftarrow{x} = \ldots 01^n$ induces the same distribution over the initial state s_0 as any other. However, for $n \geq 1$ there is some possibility of being in both σ_1 and σ_2 at time 0. A direct calculation shows that the distributions over s_0 and,

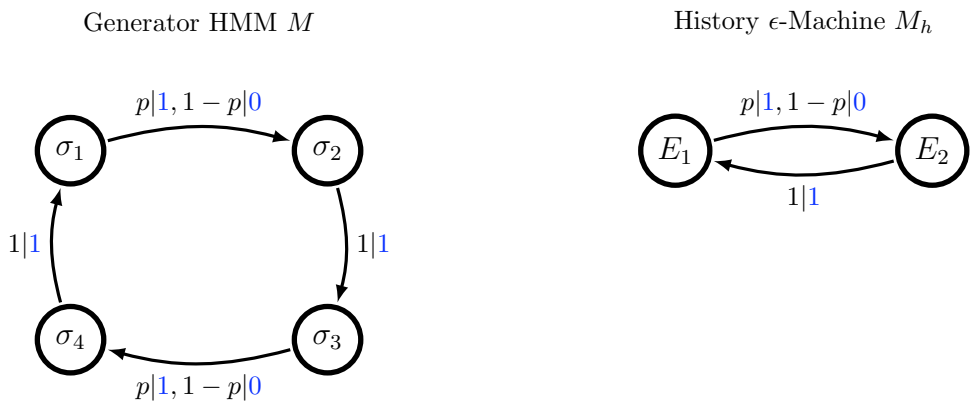

FIG. 4: A nonminimal generating HMM M for the Noisy Period-2 (NP2) Process (left), and the associated history ϵ-machine M_h for this process (right). $p \in (0, 1)$ is a parameter.

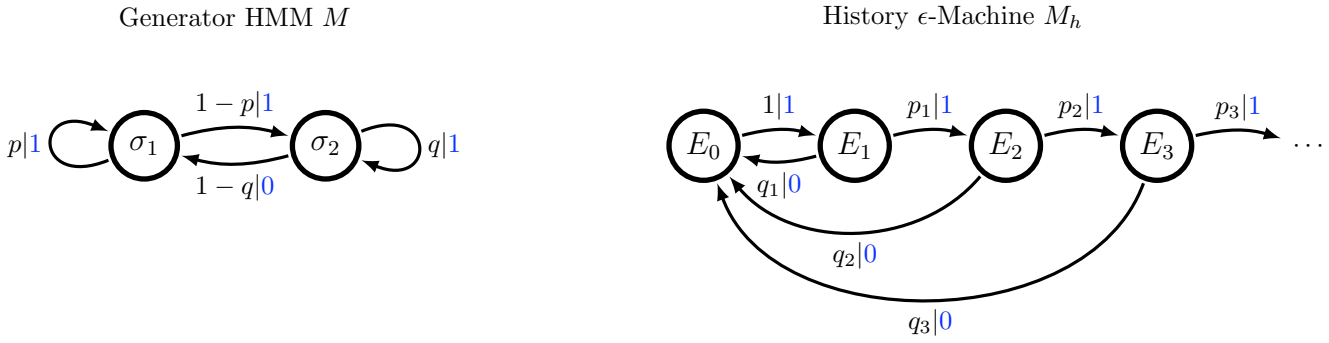

FIG. 5: The Simple Nonunifilar Source (SNS) M (left) and associated history ϵ-machine M_h (right) for the process $\mathcal{P}$ generated by M. In the history ϵ-machine, $q_n + p_n = 1$ for each $n \in \mathbb{N}$ and $(q_n)_{n\in\mathbb{N}}$ is an increasing sequence defined by: $q_n = (1-q) \cdot \left((1-p)\sum_{m=0}^{n-1} p^m q^{n-1-m}\right) / \left(p^n + (1-p)\sum_{m=0}^{n-1} p^m q^{n-1-m}\right)$.

hence, the distributions over future output sequences $\overrightarrow{x}$ are distinct for different values of n. Thus, since a.e. past $\overleftarrow{x}$ contains at least one 0, it follows that the process $\mathcal{P}$ has a countable collection of positive probability equivalence classes of pasts, comprising a set of full measure: $\{E_n : n = 0, 1, 2 \ldots\}$ where $E_n = \{\overleftarrow{x} = \ldots 01^n\}$. This leads to a *countable-state history ϵ-machine* M_h as depicted on the right of Figure 5. We will not address countable-state machines further here, as other technical issues arise in this case. Conceptually, however, it is similar to the finite-state case and may be depicted graphically in an analogous fashion.

IV. EQUIVALENCE

We will show that the two ϵ-machine definitions—history and generator—are equivalent in the following sense:

1. If $\mathcal{P}$ is the process generated by a generator ϵ-machine M_g, then $\mathcal{P}$ is finitely characterized and the history ϵ-machine $M_h(\mathcal{P})$ is isomorphic to M_g as a hidden Markov model.

2. If $\mathcal{P}$ is a finitely characterized, stationary, ergodic, finite-alphabet process, then the history ϵ-machine $M_h(\mathcal{P})$, when considered as a hidden Markov model, is also a generator ϵ-machine. And, the process $\mathcal{P}'$ generated by M_h is the same as the original process $\mathcal{P}$ from which the history machine was derived.

That is, there is a $1-1$ correspondence between finite-state generator ϵ-machines and finite-state history ϵ-machines. Every generator ϵ-machine is also a history ϵ-machine, for the same process $\mathcal{P}$ it generates. Every history ϵ-machine is also a generator ϵ-machine, for the same process $\mathcal{P}$ from which it was derived.

A. Generator ϵ-Machines are History ϵ-Machines

In this section we establish equivalence in the following direction:

Theorem 1. *If $\mathcal{P} = (\mathcal{X}^{\mathbb{Z}}, \mathbb{X}, \mathbb{P})$ is the process generated by a generator ϵ-machine M_g, then $\mathcal{P}$ is finitely characterized and the history ϵ-machine $M_h(\mathcal{P})$ is isomorphic to M_g as a hidden Markov model.*

The key ideas in proving this theorem come from the study of synchronization to generator ϵ-machines [12, 13]. In order to state these ideas precisely, however, we first need to introduce some terminology.

Let M_g be a generator ϵ-machine, and let $\mathcal{P} = (\mathcal{X}^{\mathbb{Z}}, \mathbb{X}, \mathbb{P})$ and $\widetilde{\mathcal{P}} = ((\mathcal{X} \text{ x } \mathcal{S})^{\mathbb{Z}}, (\mathbb{X} \text{ x } \mathbb{S}), \widetilde{\mathbb{P}})$ be the associated symbol and symbol-state processes generated by M_g as in Section III B. Further, let the random variables $X_t : (\mathcal{X} \text{ x } \mathcal{S})^{\mathbb{Z}} \to \mathcal{X}$ and $S_t : (\mathcal{X} \text{ x } \mathcal{S})^{\mathbb{Z}} \to \mathcal{S}$ be the natural projections $X_t(\overleftrightarrow{x}, \overleftrightarrow{s}) = x_t$ and $S_t(\overleftrightarrow{x}, \overleftrightarrow{s}) = s_t$, and let $\overrightarrow{X}^t = X_0 \dots X_{t-1}$ and $\overleftarrow{X}^t = X_{-t} \dots X_{-1}$.

The *process language* $\mathcal{L}(\mathcal{P})$ is the set of words w of positive probability: $\mathcal{L}(\mathcal{P}) = \{w \in \mathcal{X}^* : \mathbb{P}(w) > 0\}$. For a given word $w \in \mathcal{L}(\mathcal{P})$, we define $\phi(w) = \widetilde{\mathbb{P}}(\mathcal{S}|w)$ to be an observer's *belief distribution* as to the machine's current state after observing the word w. Specifically, for a length-t word $w \in \mathcal{L}(\mathcal{P})$, $\phi(w)$ is a probability distribution over the machine states $\{\sigma_1, \dots, \sigma_N\}$ whose i^{th} component is:

$$\begin{aligned} \phi(w)_i &= \widetilde{\mathbb{P}}(S_0 = \sigma_i | \overleftarrow{X}^t = w) \\ &= \widetilde{\mathbb{P}}(S_0 = \sigma_i, \overleftarrow{X}^t = w) / \widetilde{\mathbb{P}}(\overleftarrow{X}^t = w) \ . \end{aligned} \tag{25}$$

For a word $w \notin \mathcal{L}(\mathcal{P})$ we will, by convention, take $\phi(w) = \pi$.

For any word w, $\overline{\sigma}(w)$ is defined to be the most likely machine state at the current time given that the word w was just observed. That is, $\overline{\sigma}(w) = \sigma_{i^*}$, where i^* is defined by the relation $\phi(w)_{i^*} = \max_i \phi(w)_i$. In the case of a tie, i^* is taken to be the lowest value of the index i maximizing the quantity $\phi(w)_i$. Also, $P(w)$ is defined to be the probability of the most likely state after observing w:

$$P(w) \equiv \phi(w)_{i^*} \ . \tag{26}$$

And, $Q(w)$ is defined to be the combined probability of all other states after observing w:

$$Q(w) \equiv \sum_{i \neq i^*} \phi(w)_i = 1 - P(w) \ . \tag{27}$$

So, for example, if $\phi(w) = (0.2, 0.7, 0.1)$ then $\overline{\sigma}(w) = \sigma_2$, $P(w) = 0.7$, and $Q(w) = 0.3$.

The most recent t symbols are described by the block random variable $\overleftarrow{X}^t$, and so we define the corresponding random variables $\Phi_t = \phi(\overleftarrow{X}^t)$, $\overline{S}_t = \overline{\sigma}(\overleftarrow{X}^t)$, $P_t = P(\overleftarrow{X}^t)$, and $Q_t = Q(\overleftarrow{X}^t)$. Although the values depend only on the symbol sequence $\overleftrightarrow{x}$, formally we think of Φ_t, $\overline{S}_t$, P_t, and Q_t as defined on the cross product space $(\mathcal{X} \text{ x } \mathcal{S})^{\mathbb{Z}}$. Their realizations are denoted with lowercase letters ϕ_t, $\overline{s}_t$, p_t, and q_t, so that for a given realization $(\overleftrightarrow{x}, \overleftrightarrow{s}) \in (\mathcal{X} \text{ x } \mathcal{S})^{\mathbb{Z}}$, $\phi_t = \phi(\overleftarrow{x}^t)$, $\overline{s}_t = \overline{\sigma}(\overleftarrow{x}^t)$, $p_t = P(\overleftarrow{x}^t)$, and $q_t = Q(\overleftarrow{x}^t)$. The primary result we use is the following exponential decay bound on the quantity Q_t.

Lemma 1. *For any generator ϵ-machine M_g there exist constants $K > 0$ and $0 < \alpha < 1$ such that:*

$$\widetilde{\mathbb{P}}(Q_t > \alpha^t) \leq K \alpha^t, \ \textit{for all } t \in \mathbb{N} \ . \tag{28}$$

Proof. This follows directly from the Exact Machine Synchronization Theorem of [12] and the Nonexact Machine Synchronization Theorem of [13] by stationarity. (Note that the notation used there differs slightly from that here by a time shift of length t. That is, Q_t there refers to the observer's doubt in S_t given $\overrightarrow{X}^t$, instead of the observer's doubt in S_0 given $\overleftarrow{X}^t$. Also, L is used as a time index rather than t in those works.) □

Essentially, this lemma says that after observing a block of t symbols it is exponentially unlikely that an observer's doubt Q_t in the machine state will be more than exponentially small. Using the lemma we now prove Theorem 1.

Proof. (Theorem 1) Let M_g be a generator ϵ-machine with state set $\mathcal{S} = \{\sigma_1, \dots, \sigma_N\}$ and stationary distribution $\pi = (\pi_1, \dots, \pi_N)$. Let $\mathcal{P}$ and $\widetilde{\mathcal{P}}$ be the associated symbol and symbol-state processes generated by M_g. By Lemma 1

there exist constants $K > 0$ and $0 < \alpha < 1$ such that $\widetilde{\mathbb{P}}(Q_t > \alpha^t) \leq K\alpha^t$, for all $t \in \mathbb{N}$. Let us define sets:

$$\begin{aligned}
V_t &= \{(\overleftrightarrow{x}, \overleftrightarrow{s}) : q_t \leq \alpha^t \ , s_0 = \overline{s}_t\} \ , \\
V_t' &= \{(\overleftrightarrow{x}, \overleftrightarrow{s}) : q_t \leq \alpha^t \ , s_0 \neq \overline{s}_t\} \ , \\
W_t &= \{(\overleftrightarrow{x}, \overleftrightarrow{s}) : q_t > \alpha^t\} \ , \text{ and} \\
U_t &= W_t \cup V_t' \ .
\end{aligned}$$

Then, we have:

$$\begin{aligned}
\widetilde{\mathbb{P}}(U_t) &= \widetilde{\mathbb{P}}(V_t') + \widetilde{\mathbb{P}}(W_t) \\
&\leq \alpha^t + K\alpha^t \\
&= (K+1)\alpha^t \ .
\end{aligned}$$

So:

$$\sum_{t=1}^{\infty} \widetilde{\mathbb{P}}(U_t) \leq \sum_{t=1}^{\infty} (K+1)\alpha^t < \infty \ .$$

Hence, by the Borel-Cantelli Lemma, $\widetilde{\mathbb{P}}(U_t \text{ occurs infinitely often}) = 0$. Or, equivalently, for $\widetilde{\mathbb{P}}$ a.e. $(\overleftrightarrow{x}, \overleftrightarrow{s})$ there exists $t_0 \in \mathbb{N}$ such that $(\overleftrightarrow{x}, \overleftrightarrow{s}) \in V_t$ for all $t \geq t_0$. Now, define:

$$\begin{aligned}
C &= \{(\overleftrightarrow{x}, \overleftrightarrow{s}) : \text{ there exists } t_0 \in \mathbb{N} \text{ such that } (\overleftrightarrow{x}, \overleftrightarrow{s}) \in V_t \text{ for all } t \geq t_0\} \ , \\
D_i &= \{(\overleftrightarrow{x}, \overleftrightarrow{s}) : s_0 = \sigma_i\} \ , \text{ and} \\
C_i &= C \cap D_i \ .
\end{aligned}$$

According to the above discussion $\widetilde{\mathbb{P}}(C) = 1$ and, clearly, $\widetilde{\mathbb{P}}(D_i) = \pi_i$. Thus, $\widetilde{\mathbb{P}}(C_i) = \widetilde{\mathbb{P}}(C \cap D_i) = \pi_i$. Also, by the convention for $\phi(w), w \notin \mathcal{L}(\mathcal{P})$, we know that for every $(\overleftrightarrow{x}, \overleftrightarrow{s}) \in C_i$, the corresponding symbol past $\overleftarrow{x}$ is nontrivial. So, the conditional probabilities $\mathbb{P}(w|\overleftarrow{x}^t)$ are well defined for each t.

Now, given any $(\overleftrightarrow{x}, \overleftrightarrow{s}) \in C_i$ take t_0 sufficiently large so that for all $t \geq t_0$, $(\overleftrightarrow{x}, \overleftrightarrow{s}) \in V_t$. Then, for $t \geq t_0$, $\overline{s}_t = \sigma_i$ and $q_t \leq \alpha^t$. So, for any word $w \in \mathcal{X}^*$ and any $t \geq t_0$, we have:

$$\begin{aligned}
&|\mathbb{P}(w|\overleftarrow{x}^t) - \mathbf{P}_{\sigma_i}(w)| \\
&\quad = \left| \widetilde{\mathbb{P}}(\overrightarrow{X}^{|w|} = w | \overleftarrow{X}^t = \overleftarrow{x}^t) - \widetilde{\mathbb{P}}(\overrightarrow{X}^{|w|} = w | S_0 = \sigma_i) \right| \\
&\quad \overset{(*)}{=} \left| \left\{ \sum_j \widetilde{\mathbb{P}}(\overrightarrow{X}^{|w|} = w | S_0 = \sigma_j) \widetilde{\mathbb{P}}(S_0 = \sigma_j | \overleftarrow{X}^t = \overleftarrow{x}^t) \right\} - \widetilde{\mathbb{P}}(\overrightarrow{X}^{|w|} = w | S_0 = \sigma_i) \right| \\
&\quad = \left| \left\{ \sum_{j \neq i} \widetilde{\mathbb{P}}(\overrightarrow{X}^{|w|} = w | S_0 = \sigma_j) \widetilde{\mathbb{P}}(S_0 = \sigma_j | \overleftarrow{X}^t = \overleftarrow{x}^t) \right\} - \left(1 - \widetilde{\mathbb{P}}(S_0 = \sigma_i | \overleftarrow{X}^t = \overleftarrow{x}^t)\right) \widetilde{\mathbb{P}}(\overrightarrow{X}^{|w|} = w | S_0 = \sigma_i) \right| \\
&\quad \leq \left\{ \sum_{j \neq i} \widetilde{\mathbb{P}}(\overrightarrow{X}^{|w|} = w | S_0 = \sigma_j) \widetilde{\mathbb{P}}(S_0 = \sigma_j | \overleftarrow{X}^t = \overleftarrow{x}^t) \right\} + \left(1 - \widetilde{\mathbb{P}}(S_0 = \sigma_i | \overleftarrow{X}^t = \overleftarrow{x}^t)\right) \widetilde{\mathbb{P}}(\overrightarrow{X}^{|w|} = w | S_0 = \sigma_i) \\
&\quad \leq \left\{ \sum_{j \neq i} \widetilde{\mathbb{P}}(S_0 = \sigma_j | \overleftarrow{X}^t = \overleftarrow{x}^t) \right\} + \left(1 - \widetilde{\mathbb{P}}(S_0 = \sigma_i | \overleftarrow{X}^t = \overleftarrow{x}^t)\right) \\
&\quad = 2q_t \\
&\quad \leq 2\alpha^t \ .
\end{aligned}$$

Step (*) follows from the fact that $\overleftarrow{X}^m$ and $\overrightarrow{X}^n$ are conditionally independent given S_0 for any $m, n \in \mathbb{N}$, by construction of the measure $\widetilde{\mathbb{P}}$. Since $|\mathbb{P}(w|\overleftarrow{x}^t) - \mathbf{P}_{\sigma_i}(w)| \leq 2\alpha^t$ for all $t \geq t_0$, we know $\lim_{t\to\infty} \mathbb{P}(w|\overleftarrow{x}^t) = \mathbf{P}_{\sigma_i}(w)$ exists. Since this holds for all $w \in \mathcal{X}^*$, we know $\overleftarrow{x}$ is regular and $\mathbf{P}(w|\overleftarrow{x}) = \mathbf{P}_{\sigma_i}(w)$ for all $w \in \mathcal{X}^*$.

Now, let us define equivalence classes E_i, $i = 1, \ldots, N$, by:

$$E_i = \{\overleftarrow{x} : \overleftarrow{x} \text{ is regular and } \mathbf{P}(w|\overleftarrow{x}) = \mathbf{P}_{\sigma_i}(w) \text{ for all } w \in \mathcal{X}^*\} \ .$$

And, also, for each $i = 1, \ldots, N$ let:

$$\widetilde{E}_i = \{(\overleftrightarrow{x}, \overleftrightarrow{s}) : \overleftarrow{x} \in E_i\} \ .$$

By results from Appendix C we know that each equivalence class E_i is measurable, so each set $\widetilde{E}_i$ is also measurable with $\widetilde{\mathbb{P}}(\widetilde{E}_i) = \mathbb{P}(E_i)$. And, for each i, $C_i \subseteq \widetilde{E}_i$, so $\mathbb{P}(E_i) = \widetilde{\mathbb{P}}(\widetilde{E}_i) \geq \widetilde{\mathbb{P}}(C_i) = \pi_i$. Since $\sum_{i=1}^N \pi_i = 1$ and the equivalence classes $E_i, i = 1, \ldots, N$, are all disjoint, it follows that $\mathbb{P}(E_i) = \pi_i$ for each i, and $\sum_{i=1}^N \mathbb{P}(E_i) = \sum_{i=1}^N \pi_i = 1$. Hence, the process $\mathcal{P}$ is finitely characterized with positive probability equivalences classes $\mathcal{E}^+ = \{E_1, \ldots, E_N\}$.

Moreover, the equivalence classes $\{E_1, \ldots, E_N\}$—the history ϵ-machine states—have a natural one-to-one correspondence with the states of the generating ϵ-machine: $E_i \sim \sigma_i, i = 1, \ldots, N$. It remains only to verify that this bijection is also edge preserving and, thus, an isomorphism. Specifically, we must show that:

1. For each $i = 1, \ldots, N$ and $x \in \mathcal{X}$, $\mathbf{P}(x|E_i) = \mathbf{P}_{\sigma_i}(x)$, and
2. For all i and x with $\mathbf{P}(x|E_i) = \mathbf{P}_{\sigma_i}(x) > 0$, $\delta_h(E_i, x) \cong \delta(\sigma_i, x)$. That is, if $\delta_h(E_i, x) = E_j$ and $\delta(\sigma_i, x) = \sigma_{j'}$, then $j = j'$.

Point 1 follows directly from the definition of E_i. To show Point 2, take any i and x with $\mathbf{P}(x|E_i) = \mathbf{P}_{\sigma_i}(x) > 0$ and let $\delta_h(E_i, x) = E_j$ and $\delta(\sigma_i, x) = \sigma_{j'}$. Then, for any word $w \in \mathcal{X}^*$, we have:

(i) $\mathbf{P}(xw|E_i) = \mathbf{P}_{\sigma_i}(xw)$, by definition of the equivalence class E_i,

(ii) $\mathbf{P}(xw|E_i) = \mathbf{P}(x|E_i) \cdot \mathbf{P}(w|E_j)$, by Claim 11 in Appendix D, and

(iii) $\mathbf{P}_{\sigma_i}(xw) = \mathbf{P}_{\sigma_i}(x) \cdot \mathbf{P}_{\sigma_{j'}}(w)$, by Equation (10) applied to a unifilar HMM.

Since $\mathbf{P}(x|E_i) = \mathbf{P}_{\sigma_i}(x) > 0$, it follows that $\mathbf{P}(w|E_j) = \mathbf{P}_{\sigma_{j'}}(w)$. Since this holds for all $w \in \mathcal{X}^*$ and the states of the generator are probabilistically distinct, by assumption, it follows that $j = j'$.

□

Corollary 1. *Generator ϵ-machines are unique: Two generator ϵ-machines M_{g_1} and M_{g_2} that generate the same process $\mathcal{P}$ are isomorphic.*

Proof. By Theorem 1 the two generator ϵ-machines are both isomorphic to the process's history ϵ-machine $M_h(\mathcal{P})$ and, hence, isomorphic to each other. □

Remark. *Unlike history ϵ-machines that are unique by construction, generator ϵ-machines are not by definition unique. And, it is not a priori clear that they must be. Indeed, general HMMs are not unique. There are infinitely many nonisomorphic HMMs for any given process $\mathcal{P}$ generated by some HMM. Moreover, if either the unifilarity or probabilistically distinct states condition is removed from the definition of generator ϵ-machines, then uniqueness no longer holds. It is only when both of these properties are required together that one obtains uniqueness.*

B. History ϵ-Machines are Generator ϵ-Machines

In this section we establish equivalence in the reverse direction:

Theorem 2. *If $\mathcal{P}$ is a finitely characterized, stationary, ergodic, finite-alphabet process, then the history ϵ-machine $M_h(\mathcal{P})$, when considered as a hidden Markov model, is also a generator ϵ-machine. And, the process $\mathcal{P}'$ generated by M_h is the same as the original process $\mathcal{P}$ from which the history machine was derived.*

Note that by Claim 17 in Appendix E we know that for any finitely characterized, stationary, ergodic, finite-alphabet process the history ϵ-machine $M_h(\mathcal{P}) = (\mathcal{E}^+, \mathcal{X}, \{T^{(x)}\})$ is a valid hidden Markov model. So, we need only show that this HMM has the three properties of a generator ϵ-machine—strongly connected graph, unifilar transitions, and probabilistically distinct states—and that the process $\mathcal{P}'$ generated by this HMM is the same as $\mathcal{P}$. Unifilarity is immediate from the construction, but the other claims take more work and require several lemmas to establish. Throughout $\mu = (\mu_1, \ldots, \mu_N) \equiv (\mathbb{P}(E_1), \ldots, \mathbb{P}(E_N))$, where $\mathcal{E}^+ = \{E_1, \ldots, E_n\}$ is the set of positive probability equivalence classes for the process $\mathcal{P}$.

Lemma 2. *The distribution μ over equivalence-class states is stationary for the transition matrix $T = \sum_{x \in \mathcal{X}} T^{(x)}$. That is, for any $1 \leq j \leq N$, $\mu_j = \sum_{i=1}^N \mu_i \cdot T_{ij}$.*

Proof. This follows directly from Claim 15 in Appendix E and the definition of the $T^{(x)}$ matrices. □

Lemma 3. *The graph G associated with the HMM $M_h = (\mathcal{E}^+, \mathcal{X}, \{T^{(x)}\})$ consists entirely of disjoint strongly connected components. Each connected component of G is strongly connected.*

Proof. It is equivalent to show that the graphical representation of the associated Markov chain with state set $\mathcal{E}^+$ and transition matrix T consists entirely of disjoint strongly connected components. But this follows directly from the existence of a stationary distribution μ with $\mu_i = \mathbb{P}(E_i) > 0$ for all i [28]. □

Lemma 4. *For any $E_i \in \mathcal{E}^+$ and $w \in \mathcal{X}^*$, $\mathbf{P}(w|E_i) = \mathbf{P}_{E_i}(w)$, where $\mathbf{P}_{E_i}(w) \equiv \|e_i T^{(w)}\|_1$ is the probability of generating the word w starting in state E_i of the HMM $M_h = (\mathcal{E}^+, \mathcal{X}, \{T^{(x)}\})$ as defined in Section III B.*

Proof. By construction M_h is a unifilar HMM, and its transition function δ, as defined in Section III C, is the same as the transition function δ_h between equivalence classes of histories as defined in Equation (20). Moreover, we have by construction that for each $x \in \mathcal{X}$ and state E_i, $\mathbf{P}_{E_i}(x) = \mathbf{P}(x|E_i)$. The lemma follows essentially from these facts. We consider separately the two cases $\mathbf{P}(w|E_i) > 0$ and $\mathbf{P}(w|E_i) = 0$.

- Case (i) - $\mathbf{P}(w|E_i) > 0$. Let $w = w_0 \dots w_{l-1}$ be a word of length $l \geq 1$ with $\mathbf{P}(w|E_i) > 0$. By Claim 12 in Appendix D and the ensuing remark we know that the equivalence classes $s_0 = E_i$, $s_1 = \delta_h(s_0, w_0), \dots, s_l = \delta_h(s_{l-1}, w_{l-1})$ are well defined and:

  $$\mathbf{P}(w|E_i) = \prod_{t=0}^{l-1} \mathbf{P}(w_t|s_t) \ .$$

  Since $\delta_h \cong \delta$ we see that there is an allowed state path s in the HMM M_h—namely, $s = s_1, \dots, s_l$—such that the word w can be generated following s from the initial state E_i. It follows that $\mathbf{P}_{E_i}(w) > 0$ and given by Equation (15):

  $$\mathbf{P}_{E_i}(w) = \prod_{t=0}^{l-1} \mathbf{P}_{s_t}(w_t) = \prod_{t=0}^{l-1} \mathbf{P}(w_t|s_t) \ .$$

- Case (ii) - $\mathbf{P}(w|E_i) = 0$. Let $w = w_0 \dots w_{l-1}$ be a word of length $l \geq 1$ with $\mathbf{P}(w|E_i) = 0$. For $0 \leq m \leq l-1$, define $w^m = w_0 \dots w_{m-1}$ (w^0 is the null word λ). Take the largest integer $m \in \{0, \dots, l-1\}$ such that $\mathbf{P}(w^m|E_i) > 0$. By convention we take $\mathbf{P}(\lambda|E_i) = 1$ for all i, so there is always some such m. A similar analysis to above then shows that the equivalence classes $s_0, \dots, s_m$ defined by $s_0 = E_i$, $s_{t+1} = \delta_h(s_t, w_t)$ are well defined and:

  $$\mathbf{P}(w^{m+1}|E_i) = \prod_{t=0}^{m} \mathbf{P}(w_t|s_t) = \mathbf{P}_{E_i}(w^{m+1}).$$

  By our choice of m, $\mathbf{P}(w^{m+1}|E_i) = 0$, so $\mathbf{P}_{E_i}(w^{m+1}) = 0$ as well. It follows that $\mathbf{P}_{E_i}(w) = 0$, since w^{m+1} is a prefix of w.

□

Lemma 5. *For any $w \in \mathcal{X}^*$, $\mathbb{P}(w) = \|\mu T^{(w)}\|_1$.*

Proof. Let $E_{i,w} \equiv \{\overleftrightarrow{x} : \overrightarrow{x}^{|w|} = w, \overleftarrow{x} \in E_i\}$. Claim 14 of Appendix D shows that each $E_{i,w}$ is an $\mathbb{X}$-measurable set with $\mathbb{P}(E_{i,w}) = \mathbb{P}(E_i) \cdot \mathbf{P}(w|E_i)$. Since the E_is are disjoint sets with probabilities summing to 1, it follows that $\mathbb{P}(w) = \sum_{i=1}^{N} \mathbb{P}(E_{i,w})$ for each $w \in \mathcal{X}^*$. Thus, applying Lemma 4, for any $w \in \mathcal{X}^*$ we have:

$$\begin{aligned}
\mathbb{P}(w) &= \sum_{i=1}^{N} \mathbb{P}(E_{i,w}) \\
&= \sum_{i=1}^{N} \mathbb{P}(E_i) \cdot \mathbf{P}(w|E_i) \\
&= \sum_{i=1}^{N} \mu_i \|e_i T^{(w)}\|_1 \\
&= \|\mu T^{(w)}\|_1 \ .
\end{aligned}$$

□

Proof. (Theorem 2)

1. *Unifilarity*: As mentioned above, this is immediate from the history ϵ-machine construction.

2. *Probabilistically Distinct States*: Take any i and j with $i \neq j$. By construction of the equivalence classes there exists some word $w \in \mathcal{X}^*$ such that $\mathbf{P}(w|E_i) \neq \mathbf{P}(w|E_j)$. But by Lemma 4, $\mathbf{P}(w|E_i) = \mathbf{P}_{E_i}(w)$ and $\mathbf{P}(w|E_j) = \mathbf{P}_{E_j}(w)$. Hence, $\mathbf{P}_{E_i}(w) \neq \mathbf{P}_{E_j}(w)$, so the states E_i and E_j of the HMM $M_h = (\mathcal{E}^+, \mathcal{X}, \{T^{(x)}\})$ are probabilistically distinct. Since this holds for all $i \neq j$, M_h has probabilistically distinct states.

3. *Strongly Connected Graph*: By Lemma 3, we know the graph G associated with the HMM M_h consists of one or more connected components $C_1, \ldots, C_n$, each of which is strongly connected. Assume that there is more than one of these strongly connected components: $n \geq 2$. By Points 1 and 2 above we know that each component C_k defines a generator ϵ-machine. If two of these components C_k and C_j were isomorphic via a function $f : C_k \text{ states} \to C_j \text{ states}$, then for states $E_i \in C_k$ and $E_l \in C_j$ with $f(E_i) = E_l$, we would have $\mathbf{P}_{E_i}(w) = \mathbf{P}_{E_l}(w)$ for all $w \in \mathcal{X}^*$. By Lemma 4, however, this implies $\mathbf{P}(w|E_i) = \mathbf{P}(w|E_l)$ for all $w \in \mathcal{X}^*$ as well, which contradicts the fact that E_i and E_l are distinct equivalence classes. Hence, no two of the components $C_k, k = 1, \ldots, n$, can be isomorphic. By Corollary 1, this implies that the stationary processes $\mathcal{P}^k, k = 1, \ldots, n$, generated by each of the generator ϵ-machine components are all distinct. But, by a block diagonalization argument, it follows from Lemma 5 that $\mathcal{P} = \sum_{k=1}^{n} \mu^k \cdot \mathcal{P}^k$, where $\mu^k = \sum_{\{i:E_i \in C_k\}} \mu_i$. That is, for any word $w \in \mathcal{X}^*$, we have:

$$\begin{aligned} \mathbb{P}(w) &= \sum_{k=1}^{n} \mu^k \cdot \mathbb{P}^k(w) \\ &= \sum_{k=1}^{n} \mu^k \cdot \|\rho^k T^{k,(w)}\|_1 \ , \end{aligned}$$

   where ρ^k and $T^{k,(w)}$ are, respectively, the stationary state distribution and w-transition matrix for the generator ϵ-machine of component C_k. Since the $\mathcal{P}^k$s are all distinct, this implies that the process $\mathcal{P}$ cannot be ergodic, which is a contradiction. Hence, there can only be one strongly connected component C_1—the whole graph is strongly connected.

4. *Equivalence of $\mathcal{P}$ and $\mathcal{P}'$*: Since the graph of the HMM $M_h = (\mathcal{E}^+, \mathcal{X}, \{T^{(x)}\})$ is strongly connected there is a unique stationary distribution π over the states satisfying $\pi = \pi T$. However, we already know the distribution μ is stationary. Hence, $\pi = \mu$. By definition, the word probabilities $\mathbb{P}'(w)$ for the process $\mathcal{P}'$ generated by this HMM are $\mathbb{P}'(w) = \|\pi T^{(w)}\|_1, w \in \mathcal{X}^*$. But, by Lemma 5, we have also $\mathbb{P}(w) = \|\mu T^{(w)}\|_1 = \|\pi T^{(w)}\|_1$ for each $w \in \mathcal{X}^*$. Hence, $\mathbb{P}(w) = \mathbb{P}'(w)$ for all $w \in \mathcal{X}^*$, so $\mathcal{P}$ and $\mathcal{P}'$ are the same process.

□

V. CONCLUSION

We have demonstrated the equivalence of finite-state history and generator ϵ-machines. This is not a new idea. However, a formal treatment was absent until quite recently. While the rigorous development of ϵ-machines in [11] also implies equivalence, the proofs given here, especially for Theorem 1, are more direct and provide improved intuition.

The key step in proving the equivalence, at least the new approach used for Theorem 1, comes directly from recent bounds on synchronization rates for finite-state generator ϵ-machines. To generalize the equivalence to larger model classes, such as machines with a countably infinite number of states, it therefore seems reasonable that one should first deduce and apply similar synchronization results for countable-state generators. Unfortunately, for countable-state generators synchronization can be much more difficult and exponential decay rates as in Lemma 1 no longer always hold. Thus, it is unclear whether equivalence in the countable-state case always holds either. Though, the results in [11] do indicate equivalence holds for countable-state machines if the entropy in the stationary distribution $H[\pi]$ is finite, which it often is.

Acknowledgments

This work was partially support by ARO award W911NF-12-1-0234-0. NT was partially supported by an NSF VIGRE fellowship.

Appendix A: Regular Pasts and Trivial Pasts

We establish that the set of trivial pasts $\mathcal{T}$ is a null set and the set of regular pasts $\mathcal{R}$ has full measure. Throughout this section $\mathcal{P} = (\mathcal{X}^{\mathbb{Z}}, \mathbb{X}, \mathbb{P})$ is a stationary, ergodic process over a finite alphabet $\mathcal{X}$, and $(\mathcal{X}^-, \mathbb{X}^-, \mathbb{P}^-)$ is the corresponding probability space over past sequences $\overleftarrow{x}$. Other notation is used as in Section III.

Claim 1. *$\mathbb{P}^-$ a.e. $\overleftarrow{x}$ is nontrivial. That is, $\mathcal{T}$ is an $\mathbb{X}^-$ measurable set with $\mathbb{P}^-(\mathcal{T}) = 0$.*

Proof. For any fixed t, $\mathcal{T}_t \equiv \{\overleftarrow{x} : \mathbb{P}(\overleftarrow{x}^t) = 0\}$ is $\mathbb{X}^-$ measurable, since it is $\mathbb{X}_t^-$ measurable, and $\mathbb{P}^-(\mathcal{T}_t) = 0$. Hence, $\mathcal{T} = \bigcup_{t=1}^{\infty} \mathcal{T}_t^-$ is also $\mathbb{X}^-$ measurable with $\mathbb{P}^-(\mathcal{T}) = 0$. □

Claim 2. *For any $w \in \mathcal{X}^*$, $\mathbb{P}^-$ a.e. $\overleftarrow{x}$ is w-regular. That is:*

$$\mathcal{R}_w \equiv \{\overleftarrow{x} : \mathbb{P}(\overleftarrow{x}^t) > 0, \text{ for all } t \text{ and } \lim_{t\to\infty} \mathbb{P}(w|\overleftarrow{x}^t) \text{ exists}\}$$

is an $\mathbb{X}^-$ measurable set with $\mathbb{P}^-(\mathcal{R}_w) = 1$.

Proof. Fix $w \in \mathcal{X}^*$. Let $Y_{w,t} : \mathcal{X}^- \to \mathbb{R}$ be defined by:

$$Y_{w,t}(\overleftarrow{x}) = \begin{cases} \mathbb{P}(w|\overleftarrow{x}^t) & \text{if } \mathbb{P}(\overleftarrow{x}^t) > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Then, the sequence $(Y_{w,t})$ is a martingale with respect to the filtration $(\mathbb{X}_t^-)$ and $\mathbb{E}(Y_{w,t}) \leq 1$ for all t. Hence, by the Martingale Converge Theorem $Y_{w,t} \xrightarrow{a.s.} Y_w$ for some $\mathbb{X}^-$ measurable random variable Y_w. In particular, $\lim_{t\to\infty} Y_{w,t}(\overleftarrow{x})$ exists for $\mathbb{P}^-$ a.e. $\overleftarrow{x}$.

Let $\widehat{\mathcal{R}}_w \equiv \{\overleftarrow{x} : \lim_{t\to\infty} Y_{w,t}(\overleftarrow{x}) \text{ exists}\}$. Then, as just shown, $\widehat{\mathcal{R}}_w$ is $\mathbb{X}^-$ measurable with $\mathbb{P}^-(\widehat{\mathcal{R}}_w) = 1$, and from Claim 1, we know $\mathcal{T}$ is $\mathbb{X}^-$ measurable with $\mathbb{P}^-(\mathcal{T}) = 0$. Hence, $\mathcal{R}_w = \widehat{\mathcal{R}}_w \cap \mathcal{T}^c$ is also $\mathbb{X}^-$ measurable with $\mathbb{P}^-(\mathcal{R}_w) = 1$. □

Claim 3. *$\mathbb{P}^-$ a.e. $\overleftarrow{x}$ is regular. That is, $\mathcal{R}$ is an $\mathbb{X}^-$ measurable set with $\mathbb{P}^-(\mathcal{R}) = 1$.*

Proof. $\mathcal{R} = \bigcap_{w\in\mathcal{X}^*} \mathcal{R}_w$. By Claim 2, each $\mathcal{R}_w$ is $\mathbb{X}^-$ measurable with $\mathbb{P}^-(\mathcal{R}_w) = 1$. Since there are only countably many finite length words $w \in \mathcal{X}^*$, it follows that $\mathcal{R}$ is also $\mathbb{P}^-$ measurable with $\mathbb{P}^-(\mathcal{R}) = 1$. □

Appendix B: Well Definedness of Equivalence Class Transitions

We establish that the equivalence-class-to-equivalence-class transitions are well defined and normalized for the equivalence classes $E_\beta \in \mathcal{E}$. Throughout this section $\mathcal{P} = (\mathcal{X}^{\mathbb{Z}}, \mathbb{X}, \mathbb{P})$ is a stationary, ergodic process over a finite alphabet $\mathcal{X}$ and $(\mathcal{X}^-, \mathbb{X}^-, \mathbb{P}^-)$ is the corresponding probability space over past sequences $\overleftarrow{x}$. Other notation is used as in Section III. Recall that, by definition, for any regular past $\overleftarrow{x}$, $\mathbb{P}(\overleftarrow{x}^t) > 0$ for each $t \in \mathbb{N}$. This fact is used implicitly in the proofs of the following claims several times to ensure that various quantities are well defined.

Claim 4. *For any regular past $\overleftarrow{x} \in \mathcal{X}^-$ and word $w \in \mathcal{X}^*$ with $\mathbf{P}(w|\overleftarrow{x}) > 0$:*

(i) $\mathbb{P}(\overleftarrow{x}^t w) > 0$ for each $t \in \mathbb{N}$ and

(ii) $\mathbb{P}(w|\overleftarrow{x}^t) > 0$ for each $t \in \mathbb{N}$.

Proof. Fix any regular past $\overleftarrow{x} \in \mathcal{X}^-$ and word $w \in \mathcal{X}^*$ with $\mathbf{P}(w|\overleftarrow{x}) > 0$. Assume there exists $t \in \mathbb{N}$ such that $\mathbb{P}(\overleftarrow{x}^t w) = 0$. Then $\mathbb{P}(\overleftarrow{x}^n w) = 0$ for all $n \geq t$ and, thus, $\mathbb{P}(w|\overleftarrow{x}^n) = \mathbb{P}(\overleftarrow{x}^n w)/\mathbb{P}(\overleftarrow{x}^n) = 0$ for all $n \geq t$ as well. Taking the limit gives $\mathbf{P}(w|\overleftarrow{x}) = \lim_{n\to\infty} \mathbb{P}(w|\overleftarrow{x}^n) = 0$, which is a contradiction. Hence, we must have $\mathbb{P}(\overleftarrow{x}^t w) > 0$ for each t, proving (i). (ii) follows since $\mathbb{P}(w|\overleftarrow{x}^t) = \mathbb{P}(\overleftarrow{x}^t w)/\mathbb{P}(\overleftarrow{x}^t)$ is greater than zero as long as $\mathbb{P}(\overleftarrow{x}^t w) > 0$. □

Claim 5. *For any regular past $\overleftarrow{x} \in \mathcal{X}^-$ and any symbol $x \in \mathcal{X}$ with $\mathbf{P}(x|\overleftarrow{x}) > 0$, the past $\overleftarrow{x}x$ is regular.*

Proof. Fix any regular past $\overleftarrow{x} \in \mathcal{X}^-$ and symbol $x \in \mathcal{X}$ with $\mathbf{P}(x|\overleftarrow{x}) > 0$. By Claim 4, $\mathbb{P}(\overleftarrow{x}^t x)$ and $\mathbb{P}(x|\overleftarrow{x}^t)$ are both nonzero for each $t \in \mathbb{N}$. Thus, the past $\overleftarrow{x}x$ is nontrivial, and the conditional probability $\mathbb{P}(w|\overleftarrow{x}^t x)$ is well defined for each $w \in \mathcal{X}^*, t \in \mathbb{N}$ and given by:

$$\mathbb{P}(w|\overleftarrow{x}^t x) = \frac{\mathbb{P}(xw|\overleftarrow{x}^t)}{\mathbb{P}(x|\overleftarrow{x}^t)} \ .$$

Taking the limit gives:

$$\begin{aligned}\lim_{t\to\infty}\mathbb{P}(w|(\overleftarrow{x}\,x)^t) &= \lim_{t\to\infty}\mathbb{P}(w|\overleftarrow{x}^t x)\\ &= \lim_{t\to\infty}\frac{\mathbb{P}(xw|\overleftarrow{x}^t)}{\mathbb{P}(x|\overleftarrow{x}^t)}\\ &= \frac{\lim_{t\to\infty}\mathbb{P}(xw|\overleftarrow{x}^t)}{\lim_{t\to\infty}\mathbb{P}(x|\overleftarrow{x}^t)}\\ &= \frac{\mathbf{P}(xw|\overleftarrow{x})}{\mathbf{P}(x|\overleftarrow{x})}\ .\end{aligned}$$

In particular, $\lim_{t\to\infty}\mathbb{P}(w|(\overleftarrow{x}\,x)^t) = \mathbf{P}(xw|\overleftarrow{x})/\mathbf{P}(x|\overleftarrow{x})$ exists. Since this holds for all $w\in\mathcal{X}^*$, the past $\overleftarrow{x}\,x$ is regular. □

Claim 6. *If $\overleftarrow{x}$ and $\overleftarrow{x}'$ are two regular pasts in the same equivalence class $E_\beta\in\mathcal{E}$ then, for any symbol $x\in\mathcal{X}$ with $\mathbf{P}(x|E_\beta)>0$, the regular pasts $\overleftarrow{x}\,x$ and $\overleftarrow{x}'x$ are also in the same equivalence class.*

Proof. Let $E_\beta\in\mathcal{E}$ and fix any $\overleftarrow{x},\overleftarrow{x}'\in E_\beta$ and $x\in\mathcal{X}$ with $\mathbf{P}(x|E_\beta)=\mathbf{P}(x|\overleftarrow{x})=\mathbf{P}(x|\overleftarrow{x}')>0$. By Claim 5, $\overleftarrow{x}\,x$ and $\overleftarrow{x}'x$ are both regular. And, just as in the proof of Claim 5, for any $w\in\mathcal{X}^*$ we have:

$$\mathbf{P}(w|\overleftarrow{x}\,x) = \lim_{t\to\infty}\mathbb{P}(w|(\overleftarrow{x}\,x)^t) = \frac{\mathbf{P}(xw|\overleftarrow{x})}{\mathbf{P}(x|\overleftarrow{x})} = \frac{\mathbf{P}(xw|E_\beta)}{\mathbf{P}(x|E_\beta)}\ .$$

Also, similarly, for any $w\in\mathcal{X}^*$:

$$\mathbf{P}(w|\overleftarrow{x}'x) = \lim_{t\to\infty}\mathbb{P}(w|(\overleftarrow{x}'x)^t) = \frac{\mathbf{P}(xw|\overleftarrow{x}')}{\mathbf{P}(x|\overleftarrow{x}')} = \frac{\mathbf{P}(xw|E_\beta)}{\mathbf{P}(x|E_\beta)}\ .$$

Since this holds for all $w\in\mathcal{X}^*$, it follows that $\overleftarrow{x}\,x$ and $\overleftarrow{x}'x$ are both in the same equivalence class. □

Claim 7. *For any equivalence class E_β, $\sum_{x\in\mathcal{X}}\mathbf{P}(x|E_\beta)=1$.*

Proof. Fix $\overleftarrow{x}\in E_\beta$. Then:

$$\begin{aligned}\sum_{x\in\mathcal{X}}\mathbf{P}(x|E_\beta) &= \sum_{x\in\mathcal{X}}\mathbf{P}(x|\overleftarrow{x})\\ &= \sum_{x\in\mathcal{X}}\lim_{t\to\infty}\mathbb{P}(x|\overleftarrow{x}^t)\\ &= \lim_{t\to\infty}\sum_{x\in\mathcal{X}}\mathbb{P}(x|\overleftarrow{x}^t)\\ &= \lim_{t\to\infty}1\\ &= 1.\end{aligned}$$

□

Appendix C: Measurability of Equivalence Classes

We establish that the equivalence classes $E_\beta,\beta\in B$, are measurable sets. Throughout this section $\mathcal{P}=(\mathcal{X}^{\mathbb{Z}},\mathbb{X},\mathbb{P})$ is a stationary, ergodic process over a finite alphabet $\mathcal{X}$ and $(\mathcal{X}^-,\mathbb{X}^-,\mathbb{P}^-)$ is the corresponding probability space over past sequences $\overleftarrow{x}$. Other notation is used as in Section III.

Claim 8. *Let $\mathcal{A}_{w,p}\equiv\{\overleftarrow{x}:\mathbb{P}(\overleftarrow{x}^t)>0,\ \text{for all } t \text{ and } \lim_{t\to\infty}\mathbb{P}(w|\overleftarrow{x}^t)=p\}$. Then $\mathcal{A}_{w,p}$ is $\mathbb{X}^-$ measurable for each $w\in\mathcal{X}^*$ and $p\in[0,1]$.*

Proof. We proceed in steps through a series of intermediate sets.

- Let $\mathcal{A}^+_{w,p,\epsilon,t} \equiv \{\overleftarrow{x} : \mathbb{P}(\overleftarrow{x}^t) > 0,\ \mathbb{P}(w|\overleftarrow{x}^t) \leq p+\epsilon\}$ and $\mathcal{A}^-_{w,p,\epsilon,t} \equiv \{\overleftarrow{x} : \mathbb{P}(\overleftarrow{x}^t) > 0,\ \mathbb{P}(w|\overleftarrow{x}^t) \geq p-\epsilon\}$. $\mathcal{A}^+_{w,p,\epsilon,t}$ and $\mathcal{A}^-_{w,p,\epsilon,t}$ are both $\mathbb{X}^-$ measurable, since they are both $\mathbb{X}^-_t$ measurable.
- Let $\mathcal{A}^+_{w,p,\epsilon} \equiv \bigcup_{n=1}^{\infty}\bigcap_{t=n}^{\infty}\mathcal{A}^+_{w,p,\epsilon,t} = \{\overleftarrow{x} : \mathbb{P}(\overleftarrow{x}^t) > 0, \forall t \text{ and } \exists n \in \mathbb{N} \text{ such that } \mathbb{P}(w|\overleftarrow{x}^t) \leq p+\epsilon, \text{ for } t \geq n\}$, and $\mathcal{A}^-_{w,p,\epsilon} \equiv \bigcup_{n=1}^{\infty}\bigcap_{t=n}^{\infty}\mathcal{A}^-_{w,p,\epsilon,t} = \{\overleftarrow{x} : \mathbb{P}(\overleftarrow{x}^t) > 0, \forall t \text{ and } \exists n \in \mathbb{N} \text{ such that } \mathbb{P}(w|\overleftarrow{x}^t) \geq p-\epsilon, \text{ for } t \geq n\}$. Then $\mathcal{A}^+_{w,p,\epsilon}$ and $\mathcal{A}^-_{w,p,\epsilon}$ are each $\mathbb{X}^-$ measurable since they are countable unions of countable intersections of $\mathbb{X}^-$ measurable sets.
- Let $\mathcal{A}_{w,p,\epsilon} \equiv \mathcal{A}^+_{w,p,\epsilon} \cap \mathcal{A}^-_{w,p,\epsilon} = \left\{\overleftarrow{x} : \mathbb{P}(\overleftarrow{x}^t) > 0, \forall t \text{ and } \exists n \in \mathbb{N} \text{ such that } \left|\mathbb{P}(w|\overleftarrow{x}^t) - p\right| \leq \epsilon, \text{ for } t \geq n\right\}$. $\mathcal{A}_{w,p,\epsilon}$ is $\mathbb{X}^-$ measurable since it is the intersection of two $\mathbb{X}^-$ measurable sets.
- Finally, note that $\mathcal{A}_{w,p} = \bigcap_{m=1}^{\infty}\mathcal{A}_{w,p,\epsilon_m}$ where $\epsilon_m = 1/m$. And, hence, $\mathcal{A}_{w,p}$ is $\mathbb{X}^-$ measurable as it is a countable intersection of $\mathbb{X}^-$ measurable sets.

□

Claim 9. *Any equivalence class $E_\beta \in \mathcal{E}$ is an $\mathbb{X}^-$ measurable set.*

Proof. Fix any equivalence class $E_\beta \in \mathcal{E}$ and, for $w \in \mathcal{X}^*$, let $p_w = \mathbf{P}(w|E_\beta)$. By definition $E_\beta = \bigcap_{w\in\mathcal{X}^*}\mathcal{A}_{w,p_w}$ and, by Claim 8, each $\mathcal{A}_{w,p_w}$ is $\mathbb{X}^-$ is measurable. Thus, since there are only countably many finite length words $w \in \mathcal{X}^*$, E_β must also be $\mathbb{X}^-$ measurable. □

Appendix D: Probabilistic Consistency of Equivalence Class Transitions

We establish that the probability of word generation from each equivalence class is consistent in the sense of Claims 12 and 14. Claim 14 is used in the proof of Claim 15 in Appendix E, and Claim 12 is used in the proof of Theorem 2. Throughout this section we assume $\mathcal{P} = (\mathcal{X}^{\mathbb{Z}}, \mathbb{X}, \mathbb{P})$ is a stationary, ergodic process over a finite alphabet $\mathcal{X}$ and denote the corresponding probability space over past sequences as $(\mathcal{X}^-, \mathbb{X}^-, \mathbb{P}^-)$, with other notation is as in Section III. We define also the *history σ-algebra* $\mathbb{H}$ for a process $\mathcal{P} = (\mathcal{X}^{\mathbb{Z}}, \mathbb{X}, \mathbb{P})$ as the σ-algebra generated by cylinder sets of all finite length histories. That is,

$$\mathbb{H} = \sigma\left(\bigcup_{t=1}^{\infty}\mathbb{H}_t\right) \quad \text{where} \quad \mathbb{H}_t = \sigma\left(\{A_{w,-|w|} : |w| = t\}\right),$$

with $A_{w,t} = \{\overleftrightarrow{x} : x_t \ldots x_{t+|w|-1} = w\}$ as in Section III. $\mathbb{H}$ is the projection onto $\mathcal{X}^{\mathbb{Z}}$ of the σ-algebra $\mathbb{X}^-$ on the space $\mathcal{X}^-$.

Claim 10. *For any $E_\beta \in \mathcal{E}$ and $w, v \in \mathcal{X}^*$, $\mathbf{P}(wv|E_\beta) \leq \mathbf{P}(w|E_\beta)$.*

Proof. Fix $\overleftarrow{x} \in E_\beta$. Since $\mathbb{P}(wv|\overleftarrow{x}^t) \leq \mathbb{P}(w|\overleftarrow{x}^t)$ for each t:

$$\mathbf{P}(wv|E_\beta) = \mathbf{P}(wv|\overleftarrow{x}) = \lim_{t\to\infty}\mathbb{P}(wv|\overleftarrow{x}^t) \leq \lim_{t\to\infty}\mathbb{P}(w|\overleftarrow{x}^t) = \mathbf{P}(w|\overleftarrow{x}) = \mathbf{P}(w|E_\beta).$$

□

Claim 11. *Let $E_\beta \in \mathcal{E}$, $x \in \mathcal{X}$ with $\mathbf{P}(x|E_\beta) > 0$, and let $E_\alpha = \delta_h(E_\beta, x)$. Then, $\mathbf{P}(xw|E_\beta) = \mathbf{P}(x|E_\beta)\cdot\mathbf{P}(w|E_\alpha)$ for any word $w \in \mathcal{X}^*$.*

Proof. Fix $\overleftarrow{x} \in E_\beta$. Then $\overleftarrow{x}\,x \in E_\alpha$ is regular, so $\mathbb{P}(\overleftarrow{x}^t x) > 0$ for all t and we have:

$$\begin{aligned}
\mathbf{P}(xw|E_\beta) &= \mathbf{P}(xw|\overleftarrow{x}) \\
&= \lim_{t\to\infty}\mathbb{P}(xw|\overleftarrow{x}^t) \\
&= \lim_{t\to\infty}\mathbb{P}(x|\overleftarrow{x}^t)\cdot\mathbb{P}(w|\overleftarrow{x}^t x) \\
&= \lim_{t\to\infty}\mathbb{P}(x|\overleftarrow{x}^t)\cdot\lim_{t\to\infty}\mathbb{P}(w|\overleftarrow{x}^t x) \\
&= \mathbf{P}(x|\overleftarrow{x})\cdot\mathbf{P}(w|\overleftarrow{x}\,x) \\
&= \mathbf{P}(x|E_\beta)\cdot\mathbf{P}(w|E_\alpha)\ .
\end{aligned}$$

□

Claim 12. *Let $w = w_0 \dots w_{l-1} \in \mathcal{X}^*$ be a word of length $l \geq 1$, and let $w^m = w_0 \dots w_{m-1}$ for $0 \leq m \leq l$. Assume that $\mathbf{P}(w^{l-1}|E_\beta) > 0$ for some $E_\beta \in \mathcal{E}$. Then the equivalence classes s_t, $0 \leq t \leq l-1$, defined inductively by the relations $s_0 = E_\beta$ and $s_t = \delta_h(s_{t-1}, w_{t-1})$ for $1 \leq t \leq l-1$, are well defined. That is, $\mathbf{P}(w_{t-1}|s_{t-1}) > 0$ for each $1 \leq t \leq l-1$. Further, the probability $\mathbf{P}(w|E_\beta)$ may be expressed as:*

$$\mathbf{P}(w|E_\beta) = \prod_{t=0}^{l-1} \mathbf{P}(w_t|s_t).$$

In the above, $w^0 = \lambda$ is the null word and, for any equivalence class E_β, $\mathbf{P}(\lambda|E_\beta) \equiv 1$.

Proof. For $|w| = 1$ the statement is immediate and, for $|w| = 2$, it reduces to Claim 11. For $|w| \geq 3$, it can proved by induction on the length of w using Claim 11 and the consistency bound provided by Claim 10 which guarantees that $\mathbf{P}(w_0|E_\beta) > 0$ if $\mathbf{P}(w^{l-1}|E_\beta) > 0$. □

Remark. *If $\mathbf{P}(w|E_\beta) > 0$, then by Claim 10 we know $\mathbf{P}(w^{l-1}|E_\beta) > 0$, so the formula above holds for any word w with $\mathbf{P}(w|E_\beta) > 0$. Moreover, in this case, $\mathbf{P}(w_{l-1}|s_{l-1})$ must be nonzero in order to ensure $\mathbf{P}(w|E_\beta)$ is nonzero. Thus, the equivalence class $s_l = \delta_h(s_{l-1}, w_{l-1})$ is also well defined.*

The following theorem from [29, Chapter 4, Theorem 5.7] is needed in the proof of Claim 13. It is an application of the Martingale Convergence Theorem.

Theorem 3. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \mathcal{F}_3 \dots$ be an increasing sequence of σ-algebras on Ω with $\mathcal{F}_\infty = \sigma(\bigcup_{n=1}^{\infty} \mathcal{F}_n) \subseteq \mathcal{F}$. Suppose $X : \Omega \to \mathbb{R}$ is an $\mathcal{F}$-measurable random variable (with $\mathbb{E}|X| < \infty$). Then, for (any versions of) the conditional expectations $\mathbb{E}(X|\mathcal{F}_n)$ and $\mathbb{E}(X|\mathcal{F}_\infty)$, we have:*

$$\mathbb{E}(X|\mathcal{F}_n) \to \mathbb{E}(X|\mathcal{F}_\infty) \text{ a.s. and in } L^1.$$

Claim 13. *For any $w \in \mathcal{X}^*$, $\mathbf{P}_w(\overleftrightarrow{x})$ is (a version of) the conditional expectation $\mathbb{E}\left(\mathbb{1}_{A_{w,0}}|\mathbb{H}\right)(\overleftrightarrow{x})$, where $\mathbf{P}_w : \mathcal{X}^{\mathbb{Z}} \to [0,1]$ is defined by:*

$$\mathbf{P}_w(\overleftrightarrow{x}) = \begin{cases} \mathbf{P}(w|\overleftarrow{x}) & \text{if } \overleftarrow{x} \text{ is regular, where } \overleftrightarrow{x} = \overleftarrow{x}\,\overrightarrow{x}, \\ 0 & \text{otherwise.} \end{cases}$$

Proof. Fix $w \in \mathcal{X}^*$, and let $\mathbb{E}_w$ be any fixed version of the conditional expectation $\mathbb{E}\left(\mathbb{1}_{A_{w,0}}|\mathbb{H}\right)$. Since the function $\mathbf{P}_{w,t} : \mathcal{X}^{\mathbb{Z}} \to [0,1]$ defined by:

$$\mathbf{P}_{w,t}(\overleftrightarrow{x}) = \begin{cases} \mathbb{P}(w|\overleftarrow{x}^t) & \text{if } \mathbb{P}(\overleftarrow{x}^t) > 0, \\ 0 & \text{otherwise} \end{cases}$$

is a version of the conditional expectation $\mathbb{E}(\mathbb{1}_{A_{w,0}}|\mathbb{H}_t)$, Theorem 3 implies that $\mathbf{P}_{w,t}(\overleftrightarrow{x}) \to \mathbb{E}_w(\overleftrightarrow{x})$ for $\mathbb{P}$ a.e. $\overleftrightarrow{x}$. Now, define:

$$V_w = \{\overleftrightarrow{x} : \mathbf{P}_{w,t}(\overleftrightarrow{x}) \to \mathbb{E}_w(\overleftrightarrow{x})\},$$
$$W_w = \{\overleftrightarrow{x} \in V_w : \overleftarrow{x} \text{ is regular}\}\ .$$

By the above $\mathbb{P}(V_w) = 1$ and, by Claim 3, the regular pasts have probability 1. Hence, $\mathbb{P}(W_w) = 1$.

However, for each $\overleftrightarrow{x} \in W_w$ we have:

$$\mathbf{P}_w(\overleftrightarrow{x}) = \mathbf{P}(w|\overleftarrow{x}) = \mathbb{E}_w(\overleftrightarrow{x})\ .$$

Thus, $\mathbf{P}_w(\overleftrightarrow{x}) = \mathbb{E}_w(\overleftrightarrow{x})$ for $\mathbb{P}$ a.e. $\overleftrightarrow{x}$. So, for any $\mathbb{H}$-measurable set H, $\int_H \mathbf{P}_w \, d\mathbb{P} = \int_H \mathbb{E}_w \, d\mathbb{P}$. Furthermore, $\mathbf{P}_w$ is $\mathbb{H}$-measurable since $\mathbf{P}_{w,t} \xrightarrow{a.s.} \mathbf{P}_w$ and each $\mathbf{P}_{w,t}$ is $\mathbb{H}$-measurable. It follows that $\mathbf{P}_w(\overleftrightarrow{x})$ is a version of the conditional expectation $\mathbb{E}\left(\mathbb{1}_{A_{w,0}}|\mathbb{H}\right)$. □

Claim 14. *For any equivalence class $E_\beta \in \mathcal{E}$ and word $w \in \mathcal{X}^*$, the set $E_{\beta,w} \equiv \{\overleftrightarrow{x} : \overleftarrow{x} \in E_\beta, \overrightarrow{x}^{|w|} = w\}$ is $\mathbb{X}$-measurable with $\mathbb{P}(E_{\beta,w}) = \mathbb{P}(E_\beta) \cdot \mathbf{P}(w|E_\beta)$.*

Proof. Let $\widehat{E}_\beta = \{\overleftrightarrow{x} : \overleftarrow{x} \in E_\beta\}$. Then $\widehat{E}_\beta$ and $A_{w,0}$ are both $\mathbb{X}$-measurable, so their intersection $E_{\beta,w}$ is as well. And, we have:

$$\begin{aligned}
\mathbb{P}(E_{\beta,w}) &= \int_{\widehat{E}_\beta} \mathbb{1}_{A_{w,0}}(\overleftrightarrow{x})\ d\mathbb{P} \\
&\overset{(a)}{=} \int_{\widehat{E}_\beta} \mathbb{E}(\mathbb{1}_{A_{w,0}}|\mathbb{H})(\overleftrightarrow{x})\ d\mathbb{P} \\
&\overset{(b)}{=} \int_{\widehat{E}_\beta} \mathbf{P}_w(\overleftrightarrow{x})\ d\mathbb{P} \\
&= \int_{\widehat{E}_\beta} \mathbf{P}(w|E_\beta)\ d\mathbb{P} \\
&= \mathbb{P}(E_\beta) \cdot \mathbf{P}(w|E_\beta)\ ,
\end{aligned}$$

where (a) follows from the fact that $\widehat{E}_\beta$ is $\mathbb{H}$-measurable and (b) follows from Claim 13. □

Appendix E: Finitely Characterized Processes

We establish several results concerning finitely characterized processes. In particular, we show (Claim 17) that the history ϵ-machine $M_h(\mathcal{P})$ is, in fact, a well defined HMM. Throughout, we assume $\mathcal{P} = (\mathcal{X}^{\mathbb{Z}}, \mathbb{X}, \mathbb{P})$ is a stationary, ergodic, finitely characterized process over a finite alphabet $\mathcal{X}$ and denote the corresponding probability space over past sequences as $(\mathcal{X}^-, \mathbb{X}^-, \mathbb{P}^-)$. The set of positive probability equivalences is denoted $\mathcal{E}^+ = \{E_1, \ldots, E_N\}$ and the set of all equivalence classes as $\mathcal{E} = \{E_\beta, \beta \in B\}$. For equivalence classes $E_\beta, E_\alpha \in \mathcal{E}$ and symbol $x \in \mathcal{X}$, $I(x, \alpha, \beta)$ is the indicator of the transition from class E_α to class E_β on symbol x.

$$I(x, \alpha, \beta) = \begin{cases} 1 & \text{if } \mathbf{P}(x|E_\alpha) > 0 \text{ and } \delta_h(E_\alpha, x) = E_\beta, \\ 0 & \text{otherwise.} \end{cases}$$

Finally, the symbol-labeled transition matrices $T^{(x)}, x \in \mathcal{X}$ between equivalence classes $E_1, \ldots, E_N$ are defined by $T^{(x)}_{ij} = \mathbf{P}(x|E_i) \cdot I(x, i, j)$. The overall transition matrix between these equivalence classes is denoted by T, $T = \sum_{x \in \mathcal{X}} T^{(x)}$.

Claim 15. *For any equivalence class $E_\beta \in \mathcal{E}$:*

$$\mathbb{P}(E_\beta) = \sum_{i=1}^{N} \sum_{x \in \mathcal{X}} \mathbb{P}(E_i) \cdot \mathbf{P}(x|E_i) \cdot I(x, i, \beta)\ .$$

Proof. We have:

$$\begin{aligned}
\mathbb{P}(E_\beta) &\equiv \mathbb{P}(\{\overleftrightarrow{x} : \overleftarrow{x} \in E_\beta\}) \\
&\overset{(a)}{=} \mathbb{P}(\{\overleftrightarrow{x} : \overleftarrow{x}\, x_0 \in E_\beta\}) \\
&\overset{(b)}{=} \sum_{i=1}^{N} \mathbb{P}(\{\overleftrightarrow{x} : \overleftarrow{x}\, x_0 \in E_\beta, \overleftarrow{x} \in E_i\}) \\
&= \sum_{i=1}^{N} \sum_{x \in \mathcal{X}} \mathbb{P}(\{\overleftrightarrow{x} : \overleftarrow{x}\, x_0 \in E_\beta, \overleftarrow{x} \in E_i, x_0 = x\}) \\
&= \sum_{i=1}^{N} \sum_{x \in \mathcal{X}} \mathbb{P}(\{\overleftrightarrow{x} : \overleftarrow{x} \in E_i, x_0 = x\}) \cdot I(x, i, \beta) \\
&= \sum_{i=1}^{N} \sum_{x \in \mathcal{X}} \mathbb{P}(E_{i,x}) \cdot I(x, i, \beta) \\
&\overset{(c)}{=} \sum_{i=1}^{N} \sum_{x \in \mathcal{X}} \mathbb{P}(E_i) \cdot \mathbf{P}(x|E_i) \cdot I(x, i, \beta)\ ,
\end{aligned}$$

where (a) follows from stationarity, (b) from the fact that $\sum_{i=1}^{N} \mathbb{P}(E_i) = 1$, and (c) from Claim 14. □

Claim 16. *For any $E_i \in \mathcal{E}^+$ and symbol x with $\mathbf{P}(x|E_i) > 0$, $\delta_h(E_i, x) \in \mathcal{E}^+$.*

Proof. Fix $E_i \in \mathcal{E}^+$ and $x \in \mathcal{X}$ with $\mathbf{P}(x|E_i) > 0$. By Claim 15, $\mathbb{P}(\delta_h(E_i, x)) \geq \mathbb{P}(E_i) \cdot \mathbf{P}(x|E_i) > 0$. Hence, $\delta_h(E_i, x) \in \mathcal{E}^+$. □

Claim 17. *The transition matrix $T = \sum_{x \in X} T^{(x)}$ is stochastic: $\sum_{j=1}^{N} T_{ij} = 1$, for each $1 \leq i \leq N$. Hence, the HMM $M_h(\mathcal{P}) = (\mathcal{E}^+, \mathcal{X}, \{T^{(x)}\})$ is well defined.*

Proof. This follows directly from Claims 7 and 16. □

[1] J. P. Crutchfield and K. Young. Inferring statistical complexity. *Phys. Rev. Let.*, 63:105–108, 1989.
[2] J. P. Crutchfield and D. P. Feldman. Statistical complexity of simple one-dimensional spin systems. *Phys. Rev. E*, 55(2):1239R–1243R, 1997.
[3] D. P. Varn, G. S. Canright, and J.P. Crutchfield. Discovering planar disorder in close-packed structures from X-ray diffraction: Beyond the fault model. *Phys. Rev. B*, 66, 2002.
[4] D. P. Varn and J. P. Crutchfield. From finite to infinite range order via annealing: The causal architecture of deformation faulting in annealed close-packed crystals. *Phys. Lett. A*, 234(4):299–307, 2004.
[5] S. Still, J. P. Crutchfield, and C. J. Ellison. Optimal causal inference: Estimating stored information and approximating causal architecture. *CHAOS*, 20(3):037111, 2010.
[6] C.-B. Li, H. Yang, and T. Komatsuzaki. Multiscale complex network of protein conformational fluctuations in single-molecule time series. *Proc. Natl. Acad. Sci. USA*, 105:536–541, 2008.
[7] J. P. Crutchfield. The calculi of emergence: Computation, dynamics, and induction. *Physica D*, 75:11–54, 1994.
[8] D. R. Upper. *Theory and Algorithms for Hidden Markov Models and Generalized Hidden Markov Models.* PhD thesis, University of California, Berkeley, 1997. Published by University Microfilms Intl, Ann Arbor, Michigan.
[9] C. R. Shalizi and J. P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *J. Stat. Phys.*, 104:817–879, 2001.
[10] N. Ay and J. P. Crutchfield. Reductions of hidden information sources. *J. Stat. Phys.*, 210(3-4):659–684, 2005.
[11] W. Löhr. *Models of Discrete Time Stochastic Processes and Associated Complexity Measures.* PhD thesis, Max Planck Institute for Mathematics in the Sciences, Leipzig, 2010.
[12] N. F. Travers and J. P. Crutchfield. Exact synchronization for finite-state sources. *J. Stat. Phys.*, 145(5):1181–1201, 2011.
[13] N. F. Travers and J. P. Crutchfield. Asymptotic synchronization for finite-state sources. *J. Stat. Phys.*, 145(5):1202–1223, 2011.
[14] M. Boyle and K. Petersen. Hidden Markov processes in the context of symbolic dynamics. *http://arxiv.org/*, 0907.1858, 2009.
[15] B. Weiss. Subshifts of finite type and sofic systems. *Monatsh. Math.*, 77:462–474, 1973.
[16] J. E. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation.* Addison-Wesley, Reading, 1979.
[17] R. Fischer. Sofic systems and graphs. *Monatsh. Math.*, 80:179–186, 1975.
[18] M. Boyle, B. Kitchens, and B. Marcus. A note on minimal covers for sofic systems. *Proc. AMS*, 95(3):403–411, 1985.
[19] D. Lind and B. Marcus. *An Introduction to Symbolic Dynamics and Coding.* Cambridge University Press, New York, 1995.
[20] B. Kitchens and S. Tuncel. Finitary measures for subshifts of finite type and sofic systems. *Mem. AMS*, 58:no. 338, 1985.
[21] T. E. Harris. On chains of infinite order. *Pacific J. Math.*, 5:707–724, 1955.
[22] M. Keane. Strongly mixing g-measures. *Inventiones Math.*, 16:309–324, 1972.
[23] M. Bramson and S. Kalikow. Nonuniqueness in g-functions. *Israel J. Math.*, 84:153–160, 1993.
[24] Ö. Stenflo. Uniqueness in g-measures. *Nonlinearity*, 16:403–410, 2003.
[25] W. Krieger and B. Weiss. On g measures in symbolic dynamics. *Israel J. Math*, 176:1–27, 2010.
[26] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *IEEE Proc.*, 77:257, 1989.
[27] Y. Ephraim and N. Merhav. Hidden Markov processes. *IEEE Trans. Info. Th.*, 48(6):1518–1569, 2002.
[28] D. Levin, Y. Peres, and E.L. Wilmer. *Markov chains and mixing times.* American Mathematical Society, Providence, Rhode Island, 2006.
[29] R. Durrett. *Probability: Theory and examples.* Wadsworth Publishing Company, Pacific Grove, California, second edition, 1995.