Equivalence of History and Generator Epsilon-Machines

Nicholas F. Travers
James P. Crutchfield

SFI WORKING PAPER: 2011-11-051

SANTA FE INSTITUTE

Santa Fe Institute Working Paper 11-12-XXX
arxiv.org:1112.XXXX [math.PR]

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: November 18, 2011)

ϵ-Machines are minimal, unifilar representations of stationary stochastic processes. They were originally defined in the history machine sense—as machines whose states are the equivalence classes of infinite histories with the same probability distribution over futures. In analyzing synchronization, though, an alternative generator definition was given: unifilar edge-label 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.

PACS numbers: 02.50.-r 89.70.+c 05.45.Tp 02.50.Ey
Keywords: hidden Markov model, history epsilon-machine, generator epsilon-machine, measure theory, synchronization, uniqueness

I. INTRODUCTION

The ϵ-machine M of a stationary stochastic process $\mathcal{P} = (X_L)$ is its minimal, unifilar presentation. Originally, these machines were introduced as predictive models in the context of the dynamical systems [1]. Specifically, the machine states there were defined as equivalence classes of infinite past sequences $\overleftarrow{x} = \ldots x_{-2}x_{-1}$ that lead to the same predictions over future sequences $\overrightarrow{x} = x_0x_1\ldots$. This notion has subsequently been expanded upon and formalized in Refs. [2–5]. Independently, a similar formulation has also been given in Refs. [6] and [7], wherein the equivalence classes, or ϵ-machine states, are referred to as future measures.

In a recent study of synchronization [8, 9], however, an alternative formulation was given for ϵ-machines as process generators. Specifically, ϵ-machines were defined there as irreducible, edge-label hidden Markov models with unifilar transitions and probabilistically distinct states. Rather than being defined by a process, a generator ϵ-machine defines a process—the process it generates. The generator formulation is often easier to work with than the history machine formulation, since it may be difficult to determine the equivalence classes of histories directly from some other description of the process—such as the equations of motion of a dynamical system from which the process is derived.

Here, we establish the equivalence of the two formulations in the finite-state case: the original history machine definition introduced in Refs. [1, 3] and the generator machine definition used in Refs. [8, 9]. It has long been assumed that the two are equivalent, without formally specifying the generator definition, but our results make this explicit and they are developed rigorously.

Reference [5] also gives a rigorous formulation of the history ϵ-machines. Its results imply equivalence for a more general class of machines, not just finite-state. However, the statements given there are in a different language and, moreover, it is not initially clear that equivalence is their subject. Furthermore, though its proofs, especially those related to Theorem 1 below, are shorter and, perhaps, cleaner than ours, they require more machinery and are not as direct. Therefore, we feel that our demonstration of equivalence in the finite-state case with its more elementary proofs is useful and provides good intuition.

To parallel the generator machine definition, when defining history ϵ-machines here we assume that the process is not only stationary but also ergodic, and that the process alphabet is finite. Although, neither of these assumptions is strictly necessary. Only stationarity is actually needed. The history ϵ-machine definition can easily be extended to non-ergodic stationary processes and countable alphabets [3, 5].

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

II. BACKGROUND

A. Processes

There are several ways to define a stochastic process. Perhaps the most traditional is simply as a sequence of random variables (X_L) on a 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 $\overleftrightarrow{x} = \ldots x_{-1}x_0x_1\ldots$. Throughout, we restrict our attention to processes over a finite alphabet of symbols $\mathcal{X}$. 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 $\mathcal{X}$: $\mathcal{X}^{\mathbb{Z}} = \{\overleftrightarrow{x} = \ldots x_{-1}x_0x_1\ldots : x_L \in \mathcal{X},$ for all $L \in \mathbb{Z}\}$.*
- *$\mathbb{X}$ is the σ-algebra generated by finite cylinder sets of the form $A_{w,L} = \{\overleftrightarrow{x} \in \mathcal{X}^{\mathbb{Z}} : x_L \ldots x_{L+|w|-1} = w\}$; i.e., $\mathbb{X} = \sigma\left(\{A_{w,L} : w \in \mathcal{X}^*, L \in \mathbb{Z}\}\right)$.*
- *$\mathbb{P}$ is a probability measure on the measurable space $(\mathcal{X}^{\mathbb{Z}}, \mathbb{X})$.*

Remark. *We implicitly assume here that for each symbol $x \in \mathcal{X}$, $\mathbb{P}(A_{x,L}) > 0$ for some $L \in \mathbb{N}$. Otherwise, the symbol x is useless and the process can be restricted to the alphabet $\mathcal{X}/\{x\}$.*

Here, we are primarily concerned with *stationary, ergodic* processes. We recall the definitions below.

Definition 2. *A process $\mathcal{P}$ is* stationary *if for each $w \in \mathcal{X}^*$, $\mathbb{P}(A_{w,L}) = \mathbb{P}(A_{w,0})$ for all $L \in \mathbb{Z}$. In this case, we simply write $\mathbb{P}(w)$ for the probability of the word w with respect to the process: $\mathbb{P}(w) = \mathbb{P}(A_{w,0})$. A stationary process is defined entirely by the word probabilities $\mathbb{P}(w), w \in \mathcal{X}^*$.*

Definition 3. *For a length-L sequence $v^L = v_0v_1\ldots v_{L-1}$ of symbols in $\mathcal{X}$ and word $w \in \mathcal{X}^*$ with $|w| < L$, let $p(w|v^L)$ be the* empirical probability *of the word w in the sequence v^L:*

$$\begin{aligned} p(w|v^L) &\equiv \frac{\#\ \textit{occurrences of } w \textit{ in } v^L}{\#\ \textit{length-}|w| \textit{ words in } v^L} \\ &= \frac{n(w|v^L)}{L - |w| + 1} \ . \end{aligned}$$

A stationary process $\mathcal{P}$ is ergodic *if, for $\mathbb{P}$ a.e.$\overleftrightarrow{x} \in \mathcal{X}^{\mathbb{Z}}$:*

$$\lim_{L\to\infty} p(w|\overrightarrow{x}^L) = \mathbb{P}(w) \textit{ for each } w \in \mathcal{X}^* \ , \tag{1}$$

where $\overrightarrow{x}^L = x_0x_1\ldots x_{L-1}$ denotes the length-L future of a bi-infinite sequence $\overleftrightarrow{x} = \ldots x_{-1}x_0x_1\ldots$.

Remark. *One could consider, instead, limits of empirical word probabilities on finite length pasts $\overleftarrow{x}^L = x_{-L}\ldots x_{-1}$ or finite length past-futures $\overleftrightarrow{x}^L = \overleftarrow{x}^L\overrightarrow{x}^L$, but these formulations are all equivalent for stationary processes.*

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 v is followed by the word w in a bi-infinite sequence $\overleftrightarrow{x}$:

$$\begin{aligned} \mathbb{P}(w|v) &\equiv \mathbb{P}(A_{w,0}|A_{v,-|v|}) \\ &= \mathbb{P}(A_{v,-|v|} \cap A_{w,0})/\mathbb{P}(A_{v,-|v|}) \\ &= \mathbb{P}(vw)/\mathbb{P}(v) \ . \end{aligned} \tag{2}$$

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 $\mathbb{P}(uv) > 0$, $\mathbb{P}(vw|u) = \mathbb{P}(v|u) \cdot \mathbb{P}(w|uv)$.

Finally, the *history σ-algebra* $\mathbb{H}$ for a process $\mathcal{P} = (\mathcal{X}^{\mathbb{Z}}, \mathbb{X}, \mathbb{P})$ is defined to be the σ-algebra generated by cylinder sets of all finite length histories. That is,

$$\mathbb{H} = \sigma \left(\bigcup_{L=1}^{\infty} \mathbb{H}_L \right) , \tag{3}$$

where

$$\mathbb{H}_L = \sigma \left(\{ A_{w,-|w|} : |w| = L \} \right) . \tag{4}$$

$\mathbb{H}$ will be important in the construction of the history ϵ-machine in Sec. II C.

B. Generator ϵ-Machines

Generator ϵ-machines are essentially minimal, unifilar edge-label hidden Markov models. We review this definition in more detail below, beginning with general hidden Markov models and then restricting to ϵ-machines.

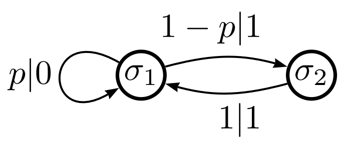
Definition 4. *An* edge-label hidden Markov model *or* (HMM) *is a triple* $(\mathcal{S}, \mathcal{X}, \{T^{(x)}\})$ *where:*

- $\mathcal{S} = \{\sigma_1, \ldots, \sigma_N\}$ *is a finite set of states,*
- $\mathcal{X}$ *is a finite alphabet of symbols, and*
- $T^{(x)}, x \in \mathcal{X}$*, are symbol-labeled transition matrices.* $T^{(x)}_{ij} \geq 0$ *represents the probability of transitioning from state* σ_i *to state* σ_j *on symbol* x*.*

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.

An HMM can also be represented graphically, 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_k is 1.

Example. *Figure 1 depicts an HMM for the Even Process. The support for this process consists of all binary sequences in which blocks of uninterrupted* 1*s are even in length, bounded by* 0*s. After each even length is reached, there is a probability* p *of breaking the block of* 1*s by inserting a* 0*.*



$$T^{(0)} = \begin{pmatrix} p & 0 \\ 0 & 0 \end{pmatrix}$$

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

FIG. 1: A hidden Markov model (the ϵ-machine) for the Even Process. The machine 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 graphical representation is presented on the left, with the corresponding transition matrices on the right. In the graphical representation, transitions denote the probability p of generating symbol x as $p|x$.

The operation of an HMM may be thought of as a weighted random walk on the associated 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 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 may directly observe this sequence of output symbols, but not the associated sequence of 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 that the HMM outputs a word $w = w_0 \ldots w_{L-1}$, $w_l \in \mathcal{X}$, while following the *state path* $s = \sigma_{i_1} \ldots \sigma_{i_L}$ in the next L steps is:

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

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

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

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 $\pi = \pi T$ [10]. 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 [10]. 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, $\widetilde{\mathcal{P}} = ((\mathcal{X} \times \mathcal{S})^{\mathbb{Z}}, (\mathbb{X} \times \mathbb{S}), \widetilde{\mathbb{P}})$ where:

1. $(\mathcal{X} \times \mathcal{S})^{\mathbb{Z}} = \{(x_L, s_L)_{L \in \mathbb{Z}} \cong ((x_L)_{L \in \mathbb{Z}}, (s_L)_{L \in \mathbb{Z}}) = (\overleftrightarrow{x}, \overleftrightarrow{s}) : x_L \in \mathcal{X}$ and $s_L \in \mathcal{S}, \forall L \in \mathbb{Z}\}$.
2. $(\mathbb{X} \times \mathbb{S})$ 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 $(\mathbb{X} \times \mathbb{S})$ is defined by Eq. (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 $(\mathbb{X} \times \mathbb{S})$ 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 $w \in \mathcal{X}^*$.

Definition 5. *A* generator ϵ-machine *is a HMM with the following properties:*

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

Since any generator ϵ-machine has a strongly connected graph, we can 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 Eq. (14). We refer to $\mathcal{P}$ as *the process* generated by the generator ϵ-machine M_G.

Example. *The Even Process machine of Fig. 1 is also an ϵ-machine: the graph is strongly connected, the transitions are unifilar, and the states are probabilistically distinct—state σ_1 can generate the symbol 0, but state σ_2 cannot.*

Finally, for any unifilar HMM we denote the state-symbol-state transition function as δ. That is, for k, x with $\mathbf{P}_{\sigma_k}(x) > 0$, we define $\delta(\sigma_k, x) = \sigma_j$, where σ_j is the (unique) state to which state σ_k transitions on symbol x. It follows immediately from Eq. (10) and the definition of matrix multiplication that for any unifilar HMM and any word $w = w_0 \ldots w_{n-1}$ with $\mathbf{P}_{\sigma_k}(w_0 \ldots w_{n-2}) > 0$:

$$\mathbf{P}_{\sigma_k}(w) = \prod_{m=0}^{n-1} \mathbf{P}_{\sigma_k^m}(w_m) \ , \tag{15}$$

where $\sigma_k^0 = \sigma_k$ and $\sigma_k^m = \delta(\sigma_k^{m-1}, w_{m-1})$, for $1 \leq m \leq n-1$. (Here, if $|w| = 1$, then $w_0 \ldots w_{n-2} = \lambda$, the null word, and $\mathbf{P}_{\sigma_k}(\lambda) \equiv 1$ for all k.)

C. History ϵ-Machines

The history ϵ-machine for a stationary process $\mathcal{P}$ is, roughly speaking, the hidden Markov model whose states are the equivalence classes of infinite past sequences $\overleftarrow{x} = \ldots x_{-2}x_{-1}$ with the same probability distributions over future sequences $\overrightarrow{x} = x_0x_1, \ldots$. However, it takes some effort to make this notion precise. The formal definition itself is quite lengthy, so for clarity of presentation the verification of many technicalities has been deferred to 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 definition 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 $\mathcal{X}$: $\mathcal{X}^- = \{\overleftarrow{x} = \ldots x_{-2}x_{-1} : x_L \in \mathcal{X}, L = -1, -2, \ldots\}$.

- $\mathbb{X}^-$ is the σ-algebra generated by finite cylinder sets on past sequences: $\mathbb{X}^- = \sigma\left(\bigcup_{L=1}^{\infty} \mathbb{X}_L^-\right)$, where $\mathbb{X}_L^- = \sigma\left(\{A_w^- : |w| = L\}\right)$ and $A_w^- = \{\overleftarrow{x} = \ldots x_{-2}x_{-1} : x_{-|w|} \ldots x_{-1} = w\}$.
- $\mathbb{P}^-$ 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 most recent L symbols of $\overleftarrow{x}$ as $\overleftarrow{x}^L = x_{-L} \ldots x_{-1}$. A past $\overleftarrow{x} \in \mathcal{X}^-$ is said to be *trivial* if $\mathbb{P}(\overleftarrow{x}^L) = 0$ for some finite L and *nontrivial* otherwise. If a past $\overleftarrow{x}$ is nontrivial, then for each $w \in \mathcal{X}^*$ $\mathbb{P}(w|\overleftarrow{x}^L)$ is well defined for each L (Eq. (2)) and one may consider $\lim_{L\to\infty} \mathbb{P}(w|\overleftarrow{x}^L)$. A nontrivial past $\overleftarrow{x}$ is said to be *w-regular* if $\lim_{L\to\infty} \mathbb{P}(w|\overleftarrow{x}^L)$ 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:

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

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 Eq. (2), since the probability of each $\overleftarrow{x}$ is normally 0. And, we do not wish to interpret $\mathbf{P}(w|\overleftarrow{x})$ in the sense of a formal conditional expectation (at this point in time), because such a definition is only unique up to a.e. equivalence, while we are concerned with its value on individual pasts. 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.

The central idea in the construction of the history ϵ-machine is the definition of the following equivalence relation $\sim$ 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.

The set of equivalence classes of regular pasts under the relation $\sim$ is denoted as $\mathcal{E} = \{E_\beta, \beta \in B\}$. In general, there may be finitely many, countably many, or uncountably many such equivalence classes. Examples are shown in Figs. 14, 15, and 17 of Ref. [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:

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

Also, App. B shows that the equivalence-class-to-equivalence-class transitions for the relation $\sim$ are well defined. That is:

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 $\overleftarrow{x}, \overleftarrow{x}'$ 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 $E_\alpha = \delta(E_\beta, x)$ to which equivalence class E_β transitions on symbol x. That is, $\delta(E_\beta, x)$ is defined by the relation:

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

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

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 exists a finite number of positive-probability equivalence classes $E_1, \ldots, E_N$ that together comprise a set of full measure: $\mathbb{P}(E_k) > 0$ for each $1 \leq k \leq N$ and $\sum_{k=1}^{N} \mathbb{P}(E_k) = 1$. For a finitely characterized process $\mathcal{P}$ we also sometimes 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 $E_i \in \mathcal{E}^+$ and $\mathbf{P}(x|E_i) > 0$, then:

$$\delta(E_i, x) \in \mathcal{E}^+ \ . \tag{22}$$

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 $T_{ij}^{(x)}$ of the matrix $T^{(x)}$ gives the probability that equivalence class E_i transitions to equivalence class E_j on symbol x:

$$T_{ij}^{(x)} = \mathbf{P}(E_i \overset{x}{\to} E_j) \equiv I(x, i, j) \cdot \mathbf{P}(x|E_i) \ , \tag{23}$$

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

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

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

Definition 6. *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 without referring to the original process $\mathcal{P}$, the state set $\mathcal{E}^+$, alphabet $\mathcal{X}$, and transition matrices $\{T^{(x)}\}$ of Def. 6 together define a valid edge-label hidden Markov model, since T is stochastic. This is critical in establishing the equivalence of history and generator ϵ-machines, since a history ϵ-machine, when viewed as a hidden Markov model, is also a process generator.

III. 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—i.e., it has a strongly connected graph, unifilar transitions, and probabilistically distinct states. And, the process $\mathcal{P}'$ it generates 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 is also a history machine (for the same process $\mathcal{P}$ it generates), and every history machine is also a generator (for the same process $\mathcal{P}$ from which it was derived).

A. Generator ϵ-Machines are History ϵ-Machines

The purpose of this section is to establish the following:

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 finite-state generator ϵ-machines [8, 9]. 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 $\widetilde{\mathcal{P}} = ((\mathcal{X} \times \mathcal{S})^{\mathbb{Z}}, (\mathbb{X} \times \mathbb{S}), \widetilde{\mathbb{P}})$ be the associated symbol and symbol-state processes generated by M_G as in Sec. II B above. Furthermore, let the random variables

$X_L : (\mathcal{X} \times \mathcal{S})^{\mathbb{Z}} \to \mathcal{X}$ and $S_L : (\mathcal{X} \times \mathcal{S})^{\mathbb{Z}} \to \mathcal{S}$ be the natural projections $X_L(\overleftrightarrow{x}, \overleftrightarrow{s}) = x_L$ and $S_L(\overleftrightarrow{x}, \overleftrightarrow{s}) = s_L$, and let $\overrightarrow{X}^L = X_0 \ldots X_{L-1}$ and $\overleftarrow{X}^L = X_{-L} \ldots X_{-1}$.

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-L word $w \in \mathcal{L}(\mathcal{P})$, $\phi(w)$ is a probability distribution over the machine states $\{\sigma_1, \ldots, \sigma_N\}$ whose k_{th} component is:

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

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, $\overline{\sigma}(w) = \sigma_{k^*}$, where k^* is defined by the relation $\phi(w)_{k^*} = \max_k \phi(w)_k$. In the case of a tie, k^* is taken to be the lowest value of the index k maximizing the quantity $\phi(w)_k$. Also, $P(w)$ is defined to be the probability of the most likely state after observing w and $Q(w)$ is defined to be the combined probability of all other states after observing w:

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

and

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

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 L symbols are described by the block random variable $\overleftarrow{X}^L$ and so we define the corresponding random variables $\Phi_L = \phi(\overleftarrow{X}^L)$, $\overline{S}_L = \overline{\sigma}(\overleftarrow{X}^L)$, $P_L = P(\overleftarrow{X}^L)$, and $Q_L = Q(\overleftarrow{X}^L)$. Although the values depend only on the symbol sequence $\overleftrightarrow{x}$, formally we think of Φ_L, $\overline{S}_L$, P_L, and Q_L as defined on the cross product space $(\mathcal{X} \times \mathcal{S})^{\mathbb{Z}}$. Their realizations are denoted with lowercase letters ϕ_L, $\overline{s}_L$, p_L, and q_L, so that for a given realization $(\overleftrightarrow{x}, \overleftrightarrow{s}) \in (\mathcal{X} \times \mathcal{S})^{\mathbb{Z}}$, $\phi_L = \phi(\overleftarrow{x}^L)$, $\overline{s}_L = \overline{\sigma}(\overleftarrow{x}^L)$, $p_L = P(\overleftarrow{x}^L)$, and $q_L = Q(\overleftarrow{x}^L)$. The primary result we use is the following exponential decay bound on the quantity Q_L.

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

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

Proof. This follows directly from the Exact Machine Synchronization Theorem of Ref. [8] and the Nonexact Machine Synchronization Theorem of Ref. [9] by stationarity. (Note that the notation used in those papers differs slightly from that here, by a time shift of length L. That is, Q_L in those papers refers to the observer's doubt in S_L given $\overrightarrow{X}^L$, instead of the observer's doubt in S_0 given $\overleftarrow{X}^L$). □

Essentially this lemma says that after observing a block of L symbols it is exponentially unlikely that an observer's "doubt" Q_L in the machine state will be more than exponentially small. Using this lemma we now prove Thm. 1.

Proof. (Theorem 1) Let M_G be a generating ϵ-machine with state set $\mathcal{S} = \{\sigma_1, \ldots, \sigma_N\}$ and stationary distribution $\pi = (\pi_1, \ldots, \pi_N)$. 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 $\widetilde{\mathbb{P}}(Q_L > \alpha^L) \leq K\alpha^L$, for all $L \in \mathbb{N}$. Let us define sets:

$$\begin{aligned} &V_L = \{(\overleftrightarrow{x}, \overleftrightarrow{s}) : q_L \leq \alpha^L \ , s_0 = \overline{s}_L\} \ , \\ &V'_L = \{(\overleftrightarrow{x}, \overleftrightarrow{s}) : q_L \leq \alpha^L \ , s_0 \neq \overline{s}_L\} \ , \\ &W_L = \{(\overleftrightarrow{x}, \overleftrightarrow{s}) : q_L > \alpha^L\} \ , \text{ and} \\ &U_L = W_L \cup V'_L \ . \end{aligned}$$

Then, we have:

$$\begin{aligned} \widetilde{\mathbb{P}}(U_L) &= \widetilde{\mathbb{P}}(V'_L) + \widetilde{\mathbb{P}}(W_L) \\ &\leq \alpha^L + K\alpha^L \\ &= (K+1)\alpha^L \ . \end{aligned}$$

So:

$$\sum_{L=1}^{\infty} \widetilde{\mathbb{P}}(U_L) \le \sum_{L=1}^{\infty} (K+1)\alpha^L < \infty \ .$$

Hence, by the Borel-Cantelli Lemma, $\widetilde{\mathbb{P}}(U_L \text{ occurs infinitely often}) = 0$. Or, equivalently, for $\widetilde{\mathbb{P}}$ a.e. $(\overleftrightarrow{x}, \overleftrightarrow{s})$ there exists $L_0 \in \mathbb{N}$ such that $(\overleftrightarrow{x}, \overleftrightarrow{s}) \in V_L$ for all $L \ge L_0$. Now define:

$$\begin{aligned} C &= \{(\overleftrightarrow{x}, \overleftrightarrow{s}) : \text{ there exists } L_0 \in \mathbb{N} \text{ such that } (\overleftrightarrow{x}, \overleftrightarrow{s}) \in V_L \text{ for all } L \ge L_0\} \ , \\ D_k &= \{(\overleftrightarrow{x}, \overleftrightarrow{s}) : s_0 = \sigma_k\} \ , \text{ and} \\ C_k &= C \cap D_k \ . \end{aligned}$$

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

Now, given any $(\overleftrightarrow{x}, \overleftrightarrow{s}) \in C_k$ take L_0 sufficiently large so that for all $L \ge L_0$, $(\overleftrightarrow{x}, \overleftrightarrow{s}) \in V_L$. Then, for $L \ge L_0$, $\overline{s}_L = \sigma_k$ and $q_L \le \alpha^L$. So, for any word $w \in \mathcal{X}^*$ and any $L \ge L_0$, we have:

$$\begin{aligned} & |\mathbb{P}(w|\overleftarrow{x}^L) - \mathbf{P}_{\sigma_k}(w)| \\ &\quad = \left| \widetilde{\mathbb{P}}(\overrightarrow{X}^{|w|} = w|\overleftarrow{X}^L = \overleftarrow{x}^L) - \widetilde{\mathbb{P}}(\overrightarrow{X}^{|w|} = w|S_0 = \sigma_k) \right| \\ &\quad \overset{(*)}{=} \left| \left\{ \sum_j \widetilde{\mathbb{P}}(\overrightarrow{X}^{|w|} = w|S_0 = \sigma_j)\widetilde{\mathbb{P}}(S_0 = \sigma_j|\overleftarrow{X}^L = \overleftarrow{x}^L) \right\} - \widetilde{\mathbb{P}}(\overrightarrow{X}^{|w|} = w|S_0 = \sigma_k) \right| \\ &\quad = \left| \left\{ \sum_{j \ne k} \widetilde{\mathbb{P}}(\overrightarrow{X}^{|w|} = w|S_0 = \sigma_j)\widetilde{\mathbb{P}}(S_0 = \sigma_j|\overleftarrow{X}^L = \overleftarrow{x}^L) \right\} - \left(1 - \widetilde{\mathbb{P}}(S_0 = \sigma_k|\overleftarrow{X}^L = \overleftarrow{x}^L)\right) \widetilde{\mathbb{P}}(\overrightarrow{X}^{|w|} = w|S_0 = \sigma_k) \right| \\ &\quad \le \left\{ \sum_{j \ne k} \widetilde{\mathbb{P}}(\overrightarrow{X}^{|w|} = w|S_0 = \sigma_j)\widetilde{\mathbb{P}}(S_0 = \sigma_j|\overleftarrow{X}^L = \overleftarrow{x}^L) \right\} + \left(1 - \widetilde{\mathbb{P}}(S_0 = \sigma_k|\overleftarrow{X}^L = \overleftarrow{x}^L)\right) \widetilde{\mathbb{P}}(\overrightarrow{X}^{|w|} = w|S_0 = \sigma_k) \\ &\quad \le \left\{ \sum_{j \ne k} \widetilde{\mathbb{P}}(S_0 = \sigma_j|\overleftarrow{X}^L = \overleftarrow{x}^L) \right\} + \left(1 - \widetilde{\mathbb{P}}(S_0 = \sigma_k|\overleftarrow{X}^L = \overleftarrow{x}^L)\right) \\ &\quad = 2\widetilde{q}_L \\ &\quad \le 2\alpha^L \ . \end{aligned}$$

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 $|\mathbb{P}(w|\overleftarrow{x}^L) - \mathbf{P}_{\sigma_k}(w)| \le 2\alpha^L$ for all $L \ge L_0$, we know $\lim_{L\to\infty} \mathbb{P}(w|\overleftarrow{x}^L) = \mathbf{P}_{\sigma_k}(w)$ exists. Since this holds for all $w \in \mathcal{X}^*$, we know $\overleftarrow{x}$ is regular and $\mathbf{P}(w|\overleftarrow{x}) = \mathbf{P}_{\sigma_k}(w)$ for all $w \in \mathcal{X}^*$.

Now, let us define equivalence classes E_k, $k = 1, \ldots, N$, by:

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

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

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

By results from App. C we know that each equivalence class E_k is measurable, so each set $\widetilde{E}_k$ is also measurable with $\widetilde{\mathbb{P}}(\widetilde{E}_k) = \mathbb{P}(E_k)$. And, for each k, $C_k \subseteq \widetilde{E}_k$, so $\mathbb{P}(E_k) = \widetilde{\mathbb{P}}(\widetilde{E}_k) \ge \widetilde{\mathbb{P}}(C_k) = \pi_k$. Since $\sum_{k=1}^N \pi_k = 1$ and the equivalence classes $E_k, k = 1, ..., N$, are all disjoint, it follows that $\mathbb{P}(E_k) = \pi_k$ for each k and $\sum_{k=1}^N \mathbb{P}(E_k) = \sum_{k=1}^N \pi_k = 1$. Hence, the process $\mathcal{P}$ is finitely characterized with 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: $E_k \sim \sigma_k, k = 1, \ldots, N$. It remains only to verify that this bijection is also edge preserving and, thus, an isomorphism. That is, we must show that:

1. For each $k = 1, \ldots, N$ and $x \in \mathcal{X}$, $\mathbf{P}(x|E_k) = \mathbf{P}_{\sigma_k}(x)$.

2. For all k, x with $\mathbf{P}(x|E_k) = \mathbf{P}_{\sigma_k}(x) > 0$, $\delta(E_k, x) \cong \delta(\sigma_k, x)$. That is, if $\delta(E_k, x) = E_j$ and $\delta(\sigma_k, x) = \sigma_{j'}$, then $j = j'$.

Here, $\delta(E_k, x)$ is the equivalence-class-to-equivalence-class transition function δ as defined in Eq. (20) and $\delta(\sigma_k, x)$ is the unifilar HMM transition function δ as defined in Sec. II B.

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

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

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

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

Since $\mathbf{P}(x|E_k) = \mathbf{P}_{\sigma_k}(x) > 0$, 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 Thm. 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

The purpose of this section is to establish the following:

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 $\mathcal{P}'$ it generates is the same as the original process $\mathcal{P}$ from which the history machine was derived.*

Note that by Claim 17 in App. 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 $\mathcal{P}$. To do so requires several lemmas. Throughout $\mu = (\mu_1, \ldots, \mu_N) \equiv (\mathbb{P}(E_1), \ldots, \mathbb{P}(E_N))$.

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 App. E and the definition of the $T^{(x)}$ matrices. □

Lemma 3. *The graph G associated with the HMM $(\mathcal{E}^+, \mathcal{X}, \{T^{(x)}\})$ 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 existence of a stationary distribution μ with $\mu_k = \mathbb{P}(E_k) > 0$ for all k [10]. □

Lemma 4. *For any $E_k \in \mathcal{E}^+$ and $w \in \mathcal{X}^*$, $\mathbf{P}(w|E_k) = \mathbf{P}_{E_K}(w)$, where $\mathbf{P}_{E_k}(w) = \|e_k T^{(w)}\|_1$ is the probability of generating the word w starting in state E_k of the HMM $(\mathcal{E}^+, \mathcal{X}, \{T^{(x)}\})$ as defined in Sec. II B.*

Proof. As noted in Sec. II C (Eq. (20)) for each equivalence class $E_k \in \mathcal{E}^+$ and symbol x with $\mathbf{P}(x|E_k) > 0$, there is a unique equivalence class $\delta(E_k, x)$ to which equivalence class E_k transitions on symbol x, and by Claim 16 in App. E, $\delta(E_k, x) \in \mathcal{E}^+$. It follows immediately from the construction of the HMM $(\mathcal{E}^+, \mathcal{X}, \{T^{(x)}\})$ that this HMM is unifilar and its transition function δ (as defined in Sec. II B) is the same δ. Also, by construction we have $\mathbf{P}(x|E_k) = \mathbf{P}_{E_K}(x)$ for all $x \in \mathcal{X}$, so the statement holds for words of length $|w| = 1$.

To extend to words of length $|w| > 1$, let us write $w = vxu$, where v is the longest proper prefix of w such that $\mathbf{P}(v|E_k) > 0$, x is the next symbol in w after v, and u is the remainder of the word w. u or v may be the null word λ. By definition, we take $\mathbf{P}(\lambda|E_k) = 1$ for all E_k, as in Claim 12 of App. D, so such a v always exists.

Now, let $n = |vx|$, and let w_m be the m^{th} symbol of the word w, starting with w_0. Since $\mathbf{P}(v|E_k) > 0$, we know by Claim 12 that the states E_k^m, $0 \leq m \leq n-1$, defined by $E_k^0 = E_k$ and $E_k^m = \delta(E_k^{m-1}, w_{m-1})$ for $1 \leq m \leq n-1$ are well defined. Thus, there is an allowed state path in the graph of the HMM $(\mathcal{E}^+, \mathcal{X}, \{T^{(x)}\})$ from state E_k following edges labeled with the symbols in the word v, so $\mathbf{P}_{E_k}(v) > 0$.

Therefore, since $\mathbf{P}_{E_k}(v)$ and $\mathbf{P}(v|E_k)$ are both nonzero and the HMM $(\mathcal{E}^+, \mathcal{X}, \{T^{(x)}\})$ is unifilar, we have by Claim 12 and Eq. (15):

$$\mathbf{P}(vx|E_k) = \prod_{m=0}^{n-1} \mathbf{P}(w_m|E_k^m) = \prod_{m=0}^{n-1} \mathbf{P}_{E_k^m}(w_m) = \mathbf{P}_{E_k}(vx) \ ,$$

where implicitly in the second equality we use the fact that both sets of E_k^ms are the same, since the transition function δ between equivalence classes is the same as the transition function δ between states of the HMM.

This proves the statement directly for the case $u = \lambda$. If $u \neq \lambda$, then $\mathbf{P}(vx|E_k) = 0$, so $\|e_k T^{(vx)}\|_1 = \mathbf{P}_{E_k}(vx) = 0$ and $e_k T^{(vx)}$ is the zero vector. And, hence:

$$\mathbf{P}_{E_k}(w) = \mathbf{P}_{E_k}(vxu) = \left\| e_k T^{(vxu)} \right\|_1 = \left\| \left(e_k T^{(vx)} \right) T^{(u)} \right\|_1 = 0 \ .$$

Also, if $u \neq \lambda$, then $\mathbf{P}(w_0 \ldots w_{|w|-2}|E_k) = 0$. So, by Claim 10 in App. D, $\mathbf{P}(w|E_k) = \mathbf{P}(w_0 \ldots w_{|w|-1}|E_k) = 0$. Thus, for $u \neq \lambda$, $\mathbf{P}_{E_k}(w) = \mathbf{P}(w|E_k) = 0$.

□

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

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

$$\begin{aligned} \mathbb{P}(w) &= \sum_{k=1}^N \mathbb{P}(E_{k,w}) \\ &= \sum_{k=1}^N \mathbb{P}(E_k) \cdot \mathbf{P}(w|E_k) \\ &= \sum_{k=1}^N \mu_k \|e_k T^{(w)}\|_1 \\ &= \|\mu T^{(w)}\|_1 \ . \end{aligned}$$

□

Proof. (Theorem 2)

1. *Unifilarity*: As mentioned in the proof of Lemma 4, this is immediate from the history ϵ-machine construction.

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

3. *Strongly Connected Graph*: By Lemma 3, we know the graph 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_i defines a generator ϵ-machine. If two of these components—say, C_i and C_j—were isomorphic via a function $f : C_i$ states $\to C_j$ states, then for states $E_k \in C_i$ and $E_l \in C_j$ with $f(E_k) = E_l$, we would have $\mathbf{P}_{E_k}(w) = \mathbf{P}_{E_l}(w)$ for all $w \in \mathcal{X}^*$. By Lemma 4, however, this implies $\mathbf{P}(w|E_k) = \mathbf{P}(w|E_l)$ for all $w \in \mathcal{X}^*$ as well, which contradicts the fact that E_k and E_l are distinct equivalence classes. Hence, no two of the components $C_i, i = 1, \ldots, n$, can be isomorphic. By Cor. 1, this implies that the stationary processes $\mathcal{P}^i, i = 1, \ldots, n$, generated by each of the generator ϵ-machine components are all distinct. But, by a block diagonalization argument, it follows from Lemma 5 that $\mathcal{P} = \sum_{i=1}^{n} \mu^i \cdot \mathcal{P}^i$, where $\mu^i = \sum_{\{k : E_k \in C_i\}} \mu_k$. That is, for any word $w \in \mathcal{X}^*$, we have:

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

where ρ^i and $T^{i,(w)}$ are, respectively, the stationary state distribution and w-transition matrix for the generator ϵ-machine of component C_i. Since the $\mathcal{P}^i$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 $(\mathcal{E}^+, \mathcal{X}, \{T^{(x)}\})$ is strongly connected there is a unique stationary distribution π over the states satisfying $\pi = \pi T$. But 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.

□

IV. CONCLUSION

We demonstrated the equivalence of finite-state history and generator ϵ-machines. This idea is not new in the development of ϵ-machines, but a rigorous analysis of their equivalence was absent until quite recently. While the results of Ref. [5] also imply equivalence, we feel that the proofs given here, especially for Thm. 1, are more direct and provide novel, constructive intuitions.

For example, the key step in proving the equivalence—or, at least, Thm. 1's new approach—came directly from recent bounds on synchronization rates for finite-state generator ϵ-machines. In addition, as we look forward to generalizing equivalence to larger classes, such as machines with a countably infinite number of states, it seems reasonable that one should attempt to deduce and apply similar synchronization results for countable-state generators. However, synchronization is more subtle for countable-state generators and, more to the point, exponential decay rates as in Lemma 1 no longer always hold. Thus, equivalence in the countable-state case is challenging. Fortunately, Ref. [5] indicates that it holds for countable-state machines if the entropy in the stationary distribution $H[\pi]$ is finite, which it often is.

Acknowledgments

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 that 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 Sec. II.

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 L, $\mathcal{T}_L \equiv \{\overleftarrow{x} : \mathbb{P}(\overleftarrow{x}^L) = 0\}$ is $\mathbb{X}^-$ measurable, since it is $\mathbb{X}^-_L$ measurable, and $\mathbb{P}^-(\mathcal{T}_L) = 0$. Hence, $\mathcal{T} = \bigcup_{L=1}^{\infty} \mathcal{T}_L^-$ 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,*

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

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

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

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

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

Let $\widehat{\mathcal{R}}_w \equiv \{\overleftarrow{x} : \lim_{L\to\infty} Y_{w,L}(\overleftarrow{x}) \text{ exists}\}$. 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 and $\mathbb{P}^-(\mathcal{R}) = 1$. □

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 Sec. II. Recall that, by definition, for any regular past $\overleftarrow{x}$, $\mathbb{P}(\overleftarrow{x}^L) > 0$, for each $L \in \mathbb{N}$. This fact is used implicitly in the proofs of the following claims several times to ensure that various quantities involving $\overleftarrow{x}^L$ 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) $\mathbb{P}(\overleftarrow{x}^L w) > 0$ for each $L \in \mathbb{N}$ and

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

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 $L \in \mathbb{N}$ such that $\mathbb{P}(\overleftarrow{x}^L w) = 0$. Then $\mathbb{P}(\overleftarrow{x}^l w) = 0$ for all $l \geq L$ and, thus, $\mathbb{P}(w|\overleftarrow{x}^l) = \mathbb{P}(\overleftarrow{x}^l w)/\mathbb{P}(\overleftarrow{x}^l) = 0$ for all $l \geq L$ as well. Taking the limit gives $\mathbf{P}(w|\overleftarrow{x}) = \lim_{l\to\infty} \mathbb{P}(w|\overleftarrow{x}^l) = 0$, which is a contradiction. Hence, we must have $\mathbb{P}(\overleftarrow{x}^L w) > 0$ for each L, proving (i). (ii) follows since $\mathbb{P}(w|\overleftarrow{x}^L) = \mathbb{P}(\overleftarrow{x}^L w)/\mathbb{P}(\overleftarrow{x}^L)$ is greater than zero as long as $\mathbb{P}(\overleftarrow{x}^L w) > 0$. □

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, $\mathbb{P}(\overleftarrow{x}^L x)$ and $\mathbb{P}(x|\overleftarrow{x}^L)$ are both nonzero for each $L \in \mathbb{N}$. Thus, the past $\overleftarrow{x}x$ is nontrivial and the conditional probability $\mathbb{P}(w|\overleftarrow{x}^L x)$ is well defined for each $w \in \mathcal{X}^*, L \in \mathbb{N}$ and given by:

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

Moreover, since $\mathbf{P}(x|\overleftarrow{x}) > 0$ the quantity $\mathbf{P}(xw|\overleftarrow{x})/\mathbf{P}(x|\overleftarrow{x})$ is well defined for each $w \in \mathcal{X}^*$ and we have:

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

Step (*) is permissible because the limits in the numerator and the denominator are both known to exist separately (since $\overleftarrow{x}$ is regular), and the limit in the denominator, $\lim_{L\to\infty} \mathbb{P}(x|\overleftarrow{x}^L) = \mathbf{P}(x|\overleftarrow{x})$, is nonzero by assumption. From the last line we see that $\lim_{L\to\infty} \mathbb{P}(w|(\overleftarrow{x}x)^L) = \mathbf{P}(xw|\overleftarrow{x})/\mathbf{P}(x|\overleftarrow{x})$ 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:

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

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

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

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:

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

□

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 Sec. II.

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

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

- Let $\mathcal{A}^+_{w,p,\epsilon,L} \equiv \{\overleftarrow{x} : \mathbb{P}(\overleftarrow{x}^L) > 0,\ \mathbb{P}(w|\overleftarrow{x}^L) \leq p + \epsilon\}$ and $\mathcal{A}^-_{w,p,\epsilon,L} \equiv \{\overleftarrow{x} : \mathbb{P}(\overleftarrow{x}^L) > 0,\ \mathbb{P}(w|\overleftarrow{x}^L) \geq p - \epsilon\}$. $\mathcal{A}^+_{w,p,\epsilon,L}$ and $\mathcal{A}^-_{w,p,\epsilon,L}$ are both $\mathbb{X}^-$ measurable, since they are both $\mathbb{X}^-_L$ measurable.
- Let, $\mathcal{A}^+_{w,p,\epsilon} \equiv \bigcup_{n=1}^\infty \bigcap_{L=n}^\infty \mathcal{A}^+_{w,p,\epsilon,L} = \{\overleftarrow{x} : \mathbb{P}(\overleftarrow{x}^L) > 0$ for all L, there is an $n \in \mathbb{N}$ such that $\mathbb{P}(w|\overleftarrow{x}^L) \leq p + \epsilon$ for $L \geq n\}$. And, $\mathcal{A}^-_{w,p,\epsilon} \equiv \bigcup_{n=1}^\infty \bigcap_{L=n}^\infty \mathcal{A}^-_{w,p,\epsilon,L} = \{\overleftarrow{x} : \mathbb{P}(\overleftarrow{x}^L) > 0$ for all L, there is an $n \in \mathbb{N}$ such that $\mathbb{P}(w|\overleftarrow{x}^L) \geq p - \epsilon$ for $L \geq n\}$. 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 $\mathcal{A}_{w,p,\epsilon} \equiv \mathcal{A}^+_{w,p,\epsilon} \cap \mathcal{A}^-_{w,p,\epsilon} = \left\{\overleftarrow{x} : \mathbb{P}(\overleftarrow{x}^L) > 0 \text{ for all } L, \text{ there is an } n \in \mathbb{N} \text{ such that } \left|\mathbb{P}(w|\overleftarrow{x}^L) - p\right| \leq \epsilon \text{ for } L \geq n\right\}$. Then $\mathcal{A}_{w,p,\epsilon}$ 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 App. E and Claim 12 is used in the proof of Thm. 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 $(\mathcal{X}^-, \mathbb{X}^-, \mathbb{P}^-)$. Other notation is as in Sec. II.

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 $\mathbb{P}(wv|\overleftarrow{x}^L) \leq \mathbb{P}(w|\overleftarrow{x}^L)$ for each L:

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

□

Claim 11. *Let $E_\beta \in \mathcal{E}$, $x \in \mathcal{X}$ with $\mathbf{P}(x|E_\beta) > 0$, and $E_\alpha = \delta(E_\beta, x)$. 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 $\mathbb{P}(\overleftarrow{x}^L x) > 0$ for all L and we have:

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

□

Claim 12. *Let $w = w_0 \ldots w_{n-1} \in \mathcal{X}^*$ be a word of length $n \geq 1$, and let $w^m = w_0 \ldots w_{m-1}$ for $0 \leq m \leq n$. Assume that $\mathbf{P}(w^{n-1}|E_\beta) > 0$ for some $E_\beta \in \mathcal{E}$. Then, the equivalence classes E_β^m, $0 \leq m \leq n-1$, defined by the relations $E_\beta^0 = E_\beta$ and $E_\beta^m = \delta(E_\beta^{m-1}, w_{m-1})$ for $1 \leq m \leq n-1$, are well defined. That is, $\mathbf{P}(w_{m-1}|E_\beta^{m-1}) > 0$ for each $1 \leq m \leq n-1$. And, $\mathbf{P}(w|E_\beta) = \prod_{m=0}^{n-1} \mathbf{P}(w_m|E_\beta^m)$.*

Here, $w^0 = \lambda$ 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 $\mathbf{P}(w_0|E_\beta) > 0$ if $\mathbf{P}(w^{n-1}|E_\beta) > 0$. □

The following theorem from [12, 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 $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \mathcal{F}_3 \ldots$ 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:*

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

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:*

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

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 $\mathbf{P}_{w,L} : \mathcal{X}^\mathbb{Z} \to [0,1]$ defined by:

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

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

$$\begin{aligned} V_w &= \{\overleftrightarrow{x} : \mathbf{P}_{w,L}(\overleftrightarrow{x}) \to \mathbb{E}_w(\overleftrightarrow{x})\} \text{ and} \\ \overline{\mathcal{R}} &= \{\overleftrightarrow{x} : \overleftarrow{x} \text{ is regular}\} \ . \end{aligned}$$

By Claim 3 $\mathbb{P}^-(\mathcal{R}) = 1$, so we know $\mathbb{P}(\overline{\mathcal{R}}) = 1$. And, by the above, $\mathbb{P}(V_w) = 1$. Hence, $\mathbb{P}(W_w) = 1$, where:

$$\begin{aligned} W_w &= V_w \cap \overline{\mathcal{R}} \\ &= \{\overleftrightarrow{x} : \overleftarrow{x} \text{ is regular}, \mathbf{P}(w|\overleftarrow{x}) = \mathbb{E}_w(\overleftrightarrow{x})\} \ . \end{aligned}$$

For each $\overleftrightarrow{x} \in W_w$, however, 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 $\mathbf{P}_{w,L} \xrightarrow{a.s.} \mathbf{P}_w$ and each $\mathbf{P}_{w,L}$ 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 $\overline{E}_\beta = \{\overleftrightarrow{x} : \overleftarrow{x} \in E_\beta\}$. Then $\overline{E}_\beta$ and $A_{w,0}$ are both $\mathbb{X}$ measurable, so their intersection $E_{\beta,w}$ is as well. And, we have:

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

where (a) follows from the fact that $\overline{E}_\beta$ 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 HMM associated with the history ϵ-machine $M_H(\mathcal{P})$ is well defined. 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.

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

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 T between these equivalence classes is $T = \sum_{x \in \mathcal{X}} T^{(x)}$.

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

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

Proof. We have:

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

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

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

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

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

[3] C. R. Shalizi and J. P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *J. Stat. Phys.*, 104:817–879, 2001.

[4] N. Ay and J. P. Crutchfield. Reductions of hidden information sources. *J. Stat. Phys.*, 210(3-4):659–684, 2005.

[5] W. Lohr. *Models of Discrete Time Stochastic Processes and Associated Complexity Measures.* PhD thesis, Max Planck Institute for Mathematics in the Sciences, Leipzig, 2010.

[6] K. Marton and P.C. Shields. How many future measures can there be? *Ergod. Th. and Dynam. Sys.*, 22:257–280, 2002.

[7] W. Krieger and B. Weiss. On g measures in symbolic dynamics. *Israel J. Math*, 176:1–27, 2010.

[8] N. Travers and J. P. Crutchfield. Exact synchronization for finite-state sources. *J. Stat. Phys.*, in press, 2011.

[9] N. Travers and J. P. Crutchfield. Asymptotic synchronization for finite-state sources. *J. Stat. Phys.*, in press, 2011.

[10] D. Levin, Y. Peres, and E.L. Wilmer. *Markov chains and mixing times.* American Mathematical Society, Providence, Rhode Island, 2006.

[11] J. P. Crutchfield. The calculi of emergence: Computation, dynamics, and induction. *Physica D*, 75:11–54, 1994.

[12] R. Durrett. *Probability: Theory and examples.* Wadsworth Publishing Company, Pacific Grove, California, second edition, 1995.