Entropy **2014**, *16*, 1396-1413; doi:10.3390/e16031396

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Infinite Excess Entropy Processes with Countable-State Generators

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

[1] Complexity Sciences Center, University of California at Davis, One Shields Avenue, Davis, CA 95616, USA

[2] Mathematics Department, University of California at Davis, One Shields Avenue, Davis, CA 95616, USA

[3] Physics Department, University of California at Davis, One Shields Avenue, Davis, CA 95616, USA; E-Mails: chaos@ucdavis.edu

[4] Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

[†] Present Address: Mathematics Department, Technion—Israel Institute of Technology, Technion City, Haifa 3200003, Israel; E-Mail: travers@tx.technion.ac.il

* Author to whom correspondence should be addressed; E-Mail: travers@tx.technion.ac.il.

Received: 20 December 2013; in revised form: 24 February 2014 / Accepted: 5 March 2014 / Published: 10 March 2014

Abstract: We present two examples of finite-alphabet, infinite excess entropy processes generated by stationary hidden Markov models (HMMs) with countable state sets. The first, simpler example is not ergodic, but the second is. These are the first explicit constructions of processes of this type.

Keywords: stationary stochastic process; hidden Markov model; epsilon-machine; ergodicity; entropy rate; excess entropy; mutual information

PACS Classification: 02.50.-r 89.70.+c 05.45.Tp 02.50.Ey

---

1. Introduction

For a stationary process (X_t) the *excess entropy* $\mathbf{E}$ is the mutual information between the infinite past $\overleftarrow{X} = \ldots X_{-2}X_{-1}$ and the infinite future $\overrightarrow{X} = X_0X_1\ldots$. It has a long history and is widely employed

as a measure of correlation and complexity in a variety of fields, from ergodic theory and dynamical systems to neuroscience and linguistics [1–6]. For a review the reader is referred to [7].

An important question in classifying a given process is whether the excess entropy is finite or infinite. In the former case the process is said to be *finitary*, and in the latter *infinitary*.

Over a finite alphabet, most of the commonly studied, simple process types are always finitary, including all independent identically distributed (IID) processes, finite-order Markov processes, and processes with finite-state hidden Markov model (HMM) presentations. However, there are also well known examples of finite-alphabet, infinitary processes. For instance, the symbolic dynamics at the onset of chaos in the logistic map and similar dynamical systems [7] and the stationary representation of the binary Fibonacci sequence [8] are both infinitary.

These latter processes, though, only admit stationary HMM presentations with uncountable state sets. Indeed, one can show that any process generated by a stationary, countable-state HMM either has positive entropy rate or consists entirely of periodic sequences, which these do not. Versions of the *Santa Fe Process* introduced in [6] are finite-alphabet, infinitary processes with positive entropy rate. However, they were not constructed directly as hidden Markov processes, and it seems unlikely that they should have any stationary, countable-state presentations either.

Here, we present two examples of stationary, countable-state HMMs that do generate finite-alphabet, infinitary processes. To the best of our knowledge, these are the first explicit constructions of this type in the literature. Although, subsequent to our release of the earlier version of the present work [9], two additional examples were given in [10].

Our first example is nonergodic, and the information conveyed from the past to the future essentially consists of the ergodic component along a given realization. This example is straightforward to construct and, though previously unpublished, others are likely aware of it or similar constructions. The second, ergodic example, though, is more involved, and both its structure and properties are novel.

To put these contributions in perspective, we note that *any* stationary, finite-alphabet process may be trivially presented by a stationary hidden Markov model with an uncountable state set, in which each infinite history $\overleftarrow{x}$ corresponds to a single state. Thus, it is clear that stationary HMMs with uncountable state sets can generate finite-alphabet, infinitary processes. In contrast, for any finite-state HMM $\mathbf{E}$ is always finite—bounded by the logarithm of the number of states. The case of countable-state HMMs lies in-between the finite-state and uncountable-state cases, and it was previously not demonstrated whether it is possible to have countable-state, stationary HMMs that generate infinitary, finite-alphabet processes and, in particular, ergodic ones.

2. Background

2.1. Excess Entropy

We denote by $H[X]$ the Shannon entropy in a random variable X, by $H[X|Y]$ the conditional entropy in X given Y, and by $I[X;Y]$ the mutual information between random variables X and Y. For definitions of these information theoretic quantities, as well as the definitions of stationarity and ergodicity for a stochastic process (X_t), the reader is referred to [11].

Definition 1. *For a stationary, finite-alphabet process* $(X_t)_{t\in\mathbb{Z}}$ *the* excess entropy $\mathbf{E}$ *is the mutual information between the infinite past* $\overleftarrow{X} = \ldots X_{-2}X_{-1}$ *and the infinite future* $\overrightarrow{X} = X_0X_1\ldots$ *:*

$$\mathbf{E} = I[\overleftarrow{X}; \overrightarrow{X}] = \lim_{t\to\infty} I[\overleftarrow{X}^t; \overrightarrow{X}^t] \,, \tag{1}$$

where $\overleftarrow{X}^t = X_{-t}\ldots X_{-1}$ *and* $\overrightarrow{X}^t = X_0\ldots X_{t-1}$ *are the length-*t *past and future, respectively.*

As noted in [7,12] this quantity, $\mathbf{E}$, may also be expressed alternatively as:

$$\mathbf{E} = \lim_{t\to\infty} \left(H[\overrightarrow{X}^t] - h\cdot t \right) \,, \tag{2}$$

where h is the process *entropy rate*:

$$h = \lim_{t\to\infty} \frac{H[\overrightarrow{X}^t]}{t} = \lim_{t\to\infty} H[X_t|\overrightarrow{X}^t] \,. \tag{3}$$

That is, the excess entropy $\mathbf{E}$ is the asymptotic amount of entropy (information) in length-t blocks of random variables beyond that explained by the entropy rate. The excess entropy derives its name from this latter formulation. It is also this formulation that we use to establish that the process of Section 3.1 is infinitary.

Expanding the block entropy $H[\overrightarrow{X}^t]$ in Equation (2) with the chain rule and recombining terms gives another important formulation [7]:

$$\mathbf{E} = \sum_{t=1}^{\infty} \left(h(t) - h \right) \,, \tag{4}$$

where $h(t)$ is the *length-*t *entropy-rate approximation*:

$$h(t) = H[X_{t-1}|\overrightarrow{X}^{t-1}] \,, \tag{5}$$

the conditional entropy in the t-th symbol given the previous $t-1$ symbols. This final formulation will be used to establish that the process of Section 3.2 is infinitary.

2.2. Hidden Markov Models

There are two primary types of hidden Markov models: edge-emitting (or *Mealy*) and state-emitting (or *Moore*). We work with the former edge-emitting type, but the two are equivalent in that any model of one type with a finite output alphabet may be converted to a model of the other type without changing the cardinality of the state set by more than a constant factor—the alphabet size. Thus, for our purposes, Mealy HMMs are sufficiently general. We also consider only stationary HMMs with finite output alphabets and countable state sets.

Definition 2. *A* stationary, edge-emitting, countable-state, finite-alphabet hidden Markov model *(hereafter referred to simply as a* countable-state HMM*) is a 4-tuple* $(\mathcal{S}, \mathcal{X}, \{T^{(x)}\}, \pi)$ *where:*

(1) $\mathcal{S}$ *is a countable set of states.*

(2) $\mathcal{X}$ *is a finite alphabet of output symbols.*

(3) $T^{(x)}, x \in \mathcal{X}$, *are symbol labeled transition matrices whose sum* $T = \sum_{x \in \mathcal{X}} T^{(x)}$ *is stochastic.* $T^{(x)}_{\sigma\sigma'}$ *is the probability that state* σ *transitions to state* σ' *on symbol* x.

(4) π *is a* stationary distribution *for the underlying Markov chain over states with transition matrix* T. *That is,* π *satisfies* $\pi = \pi T$.

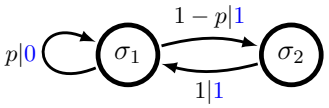
Remarks.

(1) *"Countable" in Property 1 means either finite or countably infinite. If the state set* $\mathcal{S}$ *is finite, we also refer to the HMM as* finite-state.

(2) *We do not assume, in general, that the underlying Markov chain over states with transition matrix* T *is irreducible. Thus, even in the case that* $\mathcal{S}$ *is finite, the stationary distribution* π *is not necessarily uniquely defined by the matrix* T *and is, therefore, specified separately.*

Visually, a hidden Markov model may be depicted as a directed graph with labeled edges. The vertices are the states $\sigma \in \mathcal{S}$ and, for all $\sigma, \sigma' \in \mathcal{S}$ with $T^{(x)}_{\sigma\sigma'} > 0$, there is a directed edge from state σ to state σ' labeled $p|x$ for the symbol x and transition probability $p = T^{(x)}_{\sigma\sigma'}$. These probabilities are normalized so that the sum of probabilities on all outgoing edges from each state is 1. An example is given in Figure 1.

Figure 1. A hidden Markov model (the ϵ-machine) 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. 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 associated Markov chain over states is finite-state and irreducible and, thus, has a unique stationary distribution $\pi = (\pi_1, \pi_2) = (1/(2-p), (1-p)/(2-p))$. The graphical representation of the machine is given on the left, with the corresponding transition matrices on the right. In the graphical representation the symbols labeling the transitions have been colored blue, for visual contrast, while the transition probabilities are black.



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

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

The operation of a HMM may be thought of as a weighted random walk on the associated graph. From the current state σ the next state σ' is determined by following an outgoing edge from σ chosen according to the edge probabilities (or weights). During the transition, the HMM also outputs the symbol x labeling this edge.

We denote the state at time t by S_t and the t-th symbol by X_t, so that symbol X_t is generated upon the transition from state S_t to state S_{t+1}. The state sequence (S_t) is simply a Markov chain with transition matrix T. However, we are interested not simply in this sequence of states, but also in the associated

sequence of output symbols (X_t) that are generated by reading the labels off the edges as they are followed. The interpretation is that an observer of the HMM may directly observe this sequence of output symbols, but not the hidden internal states. Alternatively, one may consider the Markov chain over edges (E_t), of which the observed symbol sequence (X_t) is simply a projection.

In either case, the process (X_t) generated by the HMM $(\mathcal{S}, \mathcal{X}, \{T^{(x)}\}, \pi)$ is defined as the output sequence of edge symbols, which results from running the Markov chain over states according to the stationary law with marginals $\mathbb{P}(S_0) = \mathbb{P}(S_t) = \pi$. It is easy to verify that this process is itself stationary, with word probabilities given by:

$$\mathbb{P}(w) = \|\pi T^{(w)}\|_1 \ , \tag{6}$$

where for a given word $w = w_1 ... w_n \in \mathcal{X}^*$, $T^{(w)}$ is the word transition matrix $T^{(w)} = T^{(w_1)} \cdots T^{(w_n)}$.

Remark. *Even for a* nonstationary HMM $(\mathcal{S}, \mathcal{X}, \{T^{(x)}\}, \rho)$*, where the state distribution* ρ *is not stationary, one may always define a one-sided process* $(X_t)_{t \geq 0}$ *with marginals given by:*

$$\mathbb{P}(\overrightarrow{X}^{|w|} = w) = \|\rho T^{(w)}\|_1 \ . \tag{7}$$

Furthermore, though the state sequence $(S_t)_{t \geq 0}$ *will not be a stationary process if* ρ *is not a stationary distribution for* T*, the output sequence* $(X_t)_{t \geq 0}$ *may still be stationary. In fact, as shown in [12] (Example 2.9), any one-sided process over a finite alphabet* $\mathcal{X}$*, stationary or not, may be represented by a countable-state, nonstationary HMM in which the states correspond to finite-length words in* $\mathcal{X}^*$*, of which there are only countably many. By stationarity, a one-sided stationary process generated by such a nonstationary HMM can be uniquely extended to a two-sided stationary process. So, in a sense, any two-sided stationary process* $(X_t)_{t \in \mathbb{Z}}$ *can be said to be generated by a nonstationary, countable-state HMM. Though, this is a slightly unnatural interpretation of process generation in that the two-sided process* $(X_t)_{t \in \mathbb{Z}}$ *is not directly that obtained by reading symbols off the edges of the HMM as it runs along transitioning between states in bi-infinite time. In either case, the space of stationary, finite-alphabet processes generated by nonstationary, countable-state HMMs is too large: it includes all stationary, finite-alphabet processes. Due to this, we restrict to the case of stationary HMMs where both the state sequence* (S_t) *and output sequence* (X_t) *are stationary processes, and henceforth use the term HMM implicitly to mean stationary HMM. Clearly, if one allows finite-alphabet processes generated by nonstationary, countable-state HMMs there are infinitary examples.*

We consider now an important property known as unifilarity. This property is useful in that many quantities are analytically computable only for unifilar HMMs. In particular, for unifilar HMMs the entropy rate h is often directly computable, unlike in the nonunifilar case. Both of the examples constructed in Section 3 are unifilar, as is the Even Process HMM of Figure 1.

Definition 3. *A HMM* $(\mathcal{S}, \mathcal{X}, \{T^{(x)}\}, \pi)$ *is* unifilar *if for each* $\sigma \in \mathcal{S}$ *and* $x \in \mathcal{X}$ *there is at most one outgoing edge from state* σ *labeled with symbol* x *in the associated graph* G*.*

It is well known that for any finite-state, unifilar HMM the entropy rate in the output process (X_t) is simply the conditional entropy in the next symbol given the current state:

$$h = H[X_0|S_0] = \sum_{\sigma \in \mathcal{S}} \pi_\sigma h_\sigma \ , \tag{8}$$

where π_σ is the stationary probability of state σ and $h_\sigma = H[X_0|S_0 = \sigma]$ is the conditional entropy in the next symbol given that the current state is σ.

We are unaware, though, of any proof that this is generally true for countable-state HMMs. If the entropy in the stationary distribution $H[\pi]$ is finite, then a proof along the lines given in [13] carries through to the countable-state case and Equation (8) still holds. However, countable-state HMMs may sometimes have $H[\pi] = \infty$. Furthermore, it can be shown [12] that the excess entropy $\mathbf{E}$ is always bounded above by $H[\pi]$. So, for the infinitary process of Section 3.2 we need slightly more than unifilarity to establish the value of h. To this end, we consider a property known as *exactness* [14].

Definition 4. *A HMM is said to be* exact *if for a.e. infinite future $\overrightarrow{x} = x_0x_1...$ generated by the HMM an observer synchronizes to the internal state after a finite time. That is, for a.e. $\overrightarrow{x}$ there exists $t \in \mathbb{N}$ such that $H[S_t|\overrightarrow{X}^t = \overrightarrow{x}^t] = 0$, where $\overrightarrow{x}^t = x_0x_1...x_{t-1}$ denotes the the first t symbols of a given $\overrightarrow{x}$.*

In the appendix we prove the following proposition.

Proposition 1. *For any countable-state, exact, unifilar HMM the entropy rate is given by the standard formula of Equation (8).*

The HMM constructed in Section 3.2 is both exact and unifilar, so Proposition 1 applies. Using this explicit formula for h, we will show that $\mathbf{E} = \sum_{t=1}^{\infty} (h(t) - h)$ is infinite.

3. Constructions

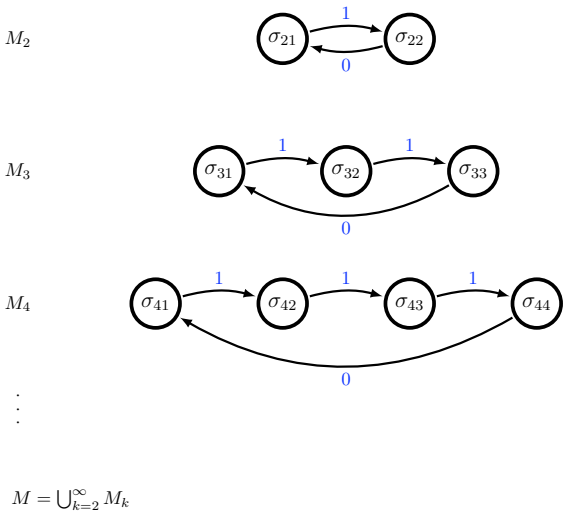
We now present the two constructions of (stationary) countable-state HMMs that generate infinitary processes. In the first example the output process is not ergodic, but in the second it is.

3.1. Heavy-Tailed Periodic Mixture: An infinitary nonergodic process with a countable-state presentation

Figure 2 depicts a countable-state HMM M, for a nonergodic infinitary process $\mathcal{P}$. The machine M consists of a countable collection of disjoint strongly connected subcomponents M_i, $i \geq 2$. For each i, the component M_i generates the periodic process $\mathcal{P}_i$ consisting of $i-1$ 1s followed by a 0. The weighting $(\mu_2, \mu_3, ...,)$ over components is taken as a heavy-tailed distribution with infinite entropy. For this reason, we refer to the process M generates as the *Heavy-Tailed Periodic Mixture* (HPM) process.

Intuitively, the information transmitted from the past to the future for the HPM Process is the ergodic component i along with the phase of the period-i process $\mathcal{P}_i$ in this component. This is more information than simply the ergodic component i, which is itself an infinite amount of information: $H[(\mu_2, \mu_3, ...,)] = \infty$. Hence, $\mathbf{E}$ should be infinite. This intuition can be made precise using the ergodic decomposition theorem of Debowski [15], but we present a more direct proof here.

Figure 2. A countable-state hidden Markov model (HMM) for the Heavy-Tailed Periodic Mixture Process. The machine M is the union of the machines $M_i, i \geq 2$, generating the period-i processes of $i-1$ 1 s followed by a 0. All topologically allowed transitions have probability 1. So, for visual clarity these probabilities are omitted from the edge labels and only the symbols labeling the transitions are given. The stationary distribution π is chosen such that the combined probability μ_i of all states in the the i-th component is $\mu_i = C/(i \log^2 i)$, where $C = 1/\left(\sum_{i=2}^{\infty} 1/(i \log^2 i)\right)$ is a normalizing constant. Formally, the HMM $M = (\mathcal{S}, \mathcal{X}, \{T^{(x)}\}, \pi)$ has alphabet $\mathcal{X} = \{0, 1\}$, state set $\mathcal{S} = \{\sigma_{ij} : i \geq 2, 1 \leq j \leq i\}$, stationary distribution π defined by $\pi_{ij} = C/(i^2 \log^2 i)$, and transition probabilities $T^{(1)}_{ij,i(j+1)} = 1$ for $i \geq 2$ and $1 \leq j < i$, $T^{(0)}_{ii,i1} = 1$ for $i \geq 2$, and all other transitions probabilities 0. Note that all logs here (and throughout) are taken base 2, as is typical when using information-theoretic quantities.



Proposition 2. *The HPM Process has infinite excess entropy.*

Proof. For the HPM Process $\mathcal{P}$ we will show that (i) $\lim_{t\to\infty} H[\overrightarrow{X}^t] = \infty$ and (ii) $h = 0$. The conclusion then follows immediately from Equation (2). To this end, we define sets:

$$W_{i,t} = \{w : |w| = t \text{ and } w \text{ is in the support of process } \mathcal{P}_i\},$$
$$U_t = \bigcup_{2 \leq i \leq t/2} W_{i,t} \text{ , and}$$
$$V_t = \bigcup_{i > t/2} W_{i,t} \text{ .}$$

Note that any word $w \in W_{i,t}$ with $i \leq t/2$ contains at least two 0s. Therefore:

(1) No two distinct states σ_{ij} and $\sigma_{ij'}$ with $i \leq t/2$ generate the same length t word.

(2) The sets $W_{i,t}, i \leq t/2$, are disjoint from both each other and V_t.

It follows that each word $w \in W_{i,t}$, with $i \leq t/2$, can only be generated from a single state σ_{ij} of the HMM and has probability:

$$\begin{aligned}\mathbb{P}(w) &= \mathbb{P}(\overrightarrow{X}^t = w) \\ &= \mathbb{P}(S_0 = \sigma_{ij}) \cdot \mathbb{P}(\overrightarrow{X}^t = w | S_0 = \sigma_{ij}) \\ &= \pi_{ij} \cdot 1 \\ &= C/(i^2 \log^2 i) \; . \end{aligned} \tag{9}$$

Hence, for any fixed t:

$$\begin{aligned}H[\overrightarrow{X}^t] &= \sum_{|w|=t} \mathbb{P}(w) \log \left(\frac{1}{\mathbb{P}(w)} \right) \\ &\geq \sum_{i=2}^{\lfloor t/2 \rfloor} \sum_{w \in W_{i,t}} \frac{C}{i^2 \log^2(i)} \log \left(\frac{i^2 \log^2(i)}{C} \right) \\ &= \sum_{i=2}^{\lfloor t/2 \rfloor} \frac{C}{i \log^2(i)} \log \left(\frac{i^2 \log^2(i)}{C} \right) \; , \end{aligned}$$

so:

$$\lim_{t \to \infty} H[\overrightarrow{X}^t] \geq \sum_{i=2}^{\infty} \frac{C}{i \log^2(i)} \log \left(\frac{i^2 \log^2(i)}{C} \right) = \infty \; , \tag{10}$$

which proves Claim (i). Now, to prove Claim (ii) consider the quantity:

$$\begin{aligned}h(t+1) &= H[X_t | \overrightarrow{X}^t] \\ &= \sum_{w \in U_t} \mathbb{P}(w) \cdot H[X_t | \overrightarrow{X}^t = w] + \sum_{w \in V_t} \mathbb{P}(w) \cdot H[X_t | \overrightarrow{X}^t = w] \; . \end{aligned} \tag{11}$$

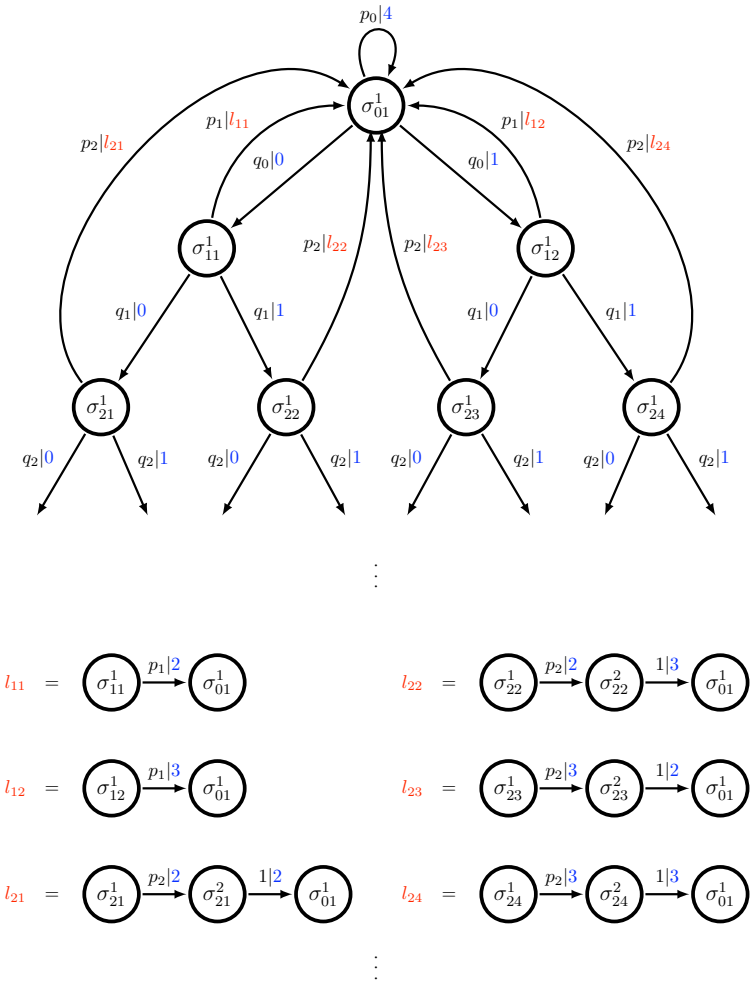
On the one hand, for $w \in U_t$, $H[X_t | \overrightarrow{X}^t = w] = 0$ since the current state and, hence, entire future are completely determined by any word $w \in U_t$. On the other hand, for $w \in V_t$, $H[X_t | \overrightarrow{X}^t = w] \leq 1$ since the alphabet is binary. Moreover, the combined probability of all words in the set V_t is simply the probability of starting in some component M_i with $i > t/2$: $\mathbb{P}(V_t) = \sum_{i > t/2} \mu_i$. Thus, by Equation (11), $h(t+1) \leq \sum_{i > t/2} \mu_i$. Since $\sum_i \mu_i$ converges, it follows that $h(t) \searrow 0$, which verifies Claim (ii). □

3.2. Branching Copy Process: An infinitary ergodic process with a countable-state presentation

Figure 3 depicts a countable-state HMM M for the ergodic, infinitary *Branching Copy Process*. Essentially, the machine M consists of a binary tree with loop backs to the root node. From the root a path is chosen down the tree with each left-right (or 0-1) choice equally likely. But, at each step there is also a chance of turning back towards the root. The path back is a not a single step, however. It has length equal to the number of steps taken down the tree before returning back, and copies the path taken down symbol-wise with 0 s replaced by 2 s and 1 s replaced by 3 s. There is also a high self-loop probability at the root node on symbol 4, so some number of 4 s will normally be generated after returning to the root node before preceding again down the tree. The process generated by this

machine is referred to as the Branching Copy (BC) Process, because the branch taken down the tree is copied on the loop back to the root.

Figure 3. A countable-state HMM for the Branching Copy Process. The machine M is essentially a binary tree with loop-back paths from each node in the tree to the root node and a self-loop on the root. At each node σ_{ij}^1 in the tree there is a probability $2q_i$ of continuing down the tree and a probability $p_i = 1-2q_i$ of turning back towards the root σ_{01}^1 on path $l_{ij} \sim \sigma_{ij}^1 \to \sigma_{ij}^2 \to \sigma_{ij}^3 ... \to \sigma_{ij}^i \to \sigma_{01}^1$. If the choice is made to head back, the next $i-1$ transitions are deterministic. The path of 0s and 1s taken to get from σ_{01}^1 to σ_{ij}^1 is copied on the return with 0 s replaced by 2 s and 1 s replaced by 3 s. Formally, the alphabet is $\mathcal{X} = \{0, 1, 2, 3, 4\}$ and the state set is $\mathcal{S} = \{\sigma_{ij}^k : i \geq 0, 1 \leq j \leq 2^i, 1 \leq k \leq \max\{i, 1\}\}$. The nonzero transition probabilities are as depicted graphically with $p_i = 1 - 2q_i$ for all $i \geq 0$, $q_i = i^2/[2(i+1)^2]$ for all $i \geq 1$, and $q_0 > 0$ taken sufficiently small so that $H[(p_0, q_0, q_0)] \leq 1/300$. The graph is strongly connected so the Markov chain over states is irreducible. Claim 1 shows that the Markov chain is also positive recurrent and, hence, has a unique stationary distribution π. Claim 2 gives the form of π.



By inspection we see that the machine is unifilar with *synchronizing word* $w = 4$, *i.e.*, $H[S_1|X_0 = 4] = 0$. Since the underlying Markov chain over states (S_t) is positive recurrent, the state sequence (S_t)

and symbol sequence (X_t) are both ergodic. Thus, a.e. infinite future $\overrightarrow{x}$ contains a 4, so the machine is exact. Therefore, Proposition 1 may be applied, and we know the entropy rate h is given by the standard formula of Equation (8): $h = \sum_\sigma \pi_\sigma h_\sigma$. Since $\mathbb{P}(S_t = \sigma) = \pi_\sigma$ for any $t \in \mathbb{N}$, we may alternatively represent this entropy rate as:

$$\begin{aligned} h &= \sum_\sigma \left(\sum_{w \in \mathcal{L}_t} \mathbb{P}(w)\phi(w)_\sigma \right) h_\sigma \\ &= \sum_{w \in \mathcal{L}_t} \mathbb{P}(w) \left(\sum_\sigma \phi(w)_\sigma h_\sigma \right) \\ &= \sum_{w \in \mathcal{L}_t} \mathbb{P}(w)\tilde{h}_w \ , \end{aligned} \tag{12}$$

where $\mathcal{L}_t = \{w : |w| = t, \mathbb{P}(w) > 0\}$ is the set of length t words in the process language $\mathcal{L}$, $\phi(w)$ is the conditional state distribution induced by the word w (*i.e.*, $\phi(w)_\sigma = \mathbb{P}(S_t = \sigma | \overrightarrow{X}^t = w)$), and $\tilde{h}_w = \sum_\sigma \phi(w)_\sigma h_\sigma$ is the $\phi(w)$-weighted average entropy in the next symbol given knowledge of the current state σ. Similarly, for any $t \in \mathbb{N}$ the entropy-rate approximation $h(t+1)$ may be expressed as:

$$h(t+1) = H[X_t | \overrightarrow{X}^t] = \sum_{w \in \mathcal{L}_t} \mathbb{P}(w) h_w \ , \tag{13}$$

where $h_w = H[X_t | \overrightarrow{X}^t = w]$ is the entropy in the next symbol after observing the word w. Combining Equations (12) and (13) we have for any $t \in \mathbb{N}$:

$$h(t+1) - h = \sum_{w \in \mathcal{L}_t} \mathbb{P}(w)(h_w - \tilde{h}_w) \ . \tag{14}$$

As we will show in Claim 6, concavity of the entropy function implies the quantity $h_w - \tilde{h}_w$ is always nonnegative. Furthermore, in Claim 5 we will show that $h_w - \tilde{h}_w$ is always bounded below by some fixed positive constant for any word w consisting entirely of 2s and 3s. Also, in Claim 3 we will show that $\mathbb{P}(W_t)$ scales as $1/t$, where W_t is the set of length-t words consisting entirely of 2s and 3s. Combining these results it follows that $h(t+1) - h \gtrsim 1/t$ and, hence, the sum $\mathbf{E} = \sum_{t=1}^\infty (h(t) - h)$ is infinite.

A more detailed analysis with the claims and their proofs is given below. In this we will use the following notation:

- $\mathbb{P}_\sigma(\cdot) = \mathbb{P}(\cdot | S_0 = \sigma)$,
- $V_t = \{w \in \mathcal{L}_t : w \text{ contains only 0s and 1s}\}$ and $W_t = \{w \in \mathcal{L}_t : w \text{ contains only 2s and 3s}\}$,
- $\pi_{ij}^k = \mathbb{P}(\sigma_{ij}^k)$ is the stationary probability of state σ_{ij}^k,
- $R_{ij} = \{\sigma_{ij}^1, \sigma_{ij}^2, ..., \sigma_{ij}^i\}$, and
- $\pi_{ij} = \sum_{k=1}^i \pi_{ij}^k$ and $\pi_i^1 = \sum_{j=1}^{2^i} \pi_{ij}^1$.

Note that:

$$\mathbb{P}_{\sigma_{01}^1}(\overrightarrow{X}^t \in V_t) = \frac{1 - p_0}{t^2} \ , \text{ for all } t \geq 1 \ , \tag{15}$$

and:

$$p_i = \frac{2i+1}{(i+1)^2} \leq \frac{2}{i}, \text{ for all } i \geq 1. \tag{16}$$

These facts will be used in the proof of Claim 1.

Claim 1. *The underlying Markov chain over states for the HMM is positive recurrent.*

Proof. Let $\tau_{\sigma_{01}^1} = \min\{t > 0 : S_t = \sigma_{01}^1\}$ be the first return time to state σ_{01}^1. Then, by continuity:

$$\begin{aligned}
\mathbb{P}_{\sigma_{01}^1}(\tau_{\sigma_{01}^1} = \infty) &= \lim_{t\to\infty} \mathbb{P}_{\sigma_{01}^1}(\tau_{\sigma_{01}^1} > 2t) \\
&= \lim_{t\to\infty} \mathbb{P}_{\sigma_{01}^1}(\overrightarrow{X}^{t+1} \in V_{t+1}) \\
&= \lim_{t\to\infty} \frac{1-p_0}{(t+1)^2} \\
&= 0 \, .
\end{aligned}$$

Hence, the Markov chain is recurrent and we have:

$$\begin{aligned}
\mathbb{E}_{\sigma_{01}^1}(\tau_{\sigma_{01}^1}) &= \sum_{t=1}^{\infty} \mathbb{P}_{\sigma_{01}^1}(\tau_{\sigma_{01}^1} = t) \cdot t \\
&= p_0 \cdot 1 + \sum_{t=1}^{\infty} \mathbb{P}_{\sigma_{01}^1}(\tau_{\sigma_{01}^1} = 2t) \cdot 2t \\
&= p_0 + \sum_{t=1}^{\infty} \mathbb{P}_{\sigma_{01}^1}(\overrightarrow{X}^t \in V_t) \cdot p_t \cdot 2t \\
&\leq p_0 + \sum_{t=1}^{\infty} \frac{1-p_0}{t^2} \cdot \frac{2}{t} \cdot 2t \\
&< \infty \, ,
\end{aligned}$$

from which it follows that the chain is also positive recurrent. Note that the topology of the chain implies the first return time may not be an odd integer greater than 1. □

Claim 2. *The stationary distribution π has:*

$$\pi_{ij}^1 = \frac{C}{i^2 \cdot 2^i} \ , \ i \geq 1, \ 1 \leq j \leq 2^i \, , \tag{17}$$

$$\pi_{ij}^k = \frac{C}{i^2 \cdot 2^i} \cdot \frac{2i+1}{(i+1)^2} \ , \ i \geq 2, \ 1 \leq j \leq 2^i, \ 2 \leq k \leq i \, , \tag{18}$$

where $C = \pi_{01}^1(1-p_0)$.

Proof. Existence of a unique stationary distribution π is guaranteed by Claim 1. Given this, clearly $\pi_1^1 = \pi_{01}^1(1-p_0)$. Similarly, for $i \geq 1$, $\pi_{i+1}^1 = \pi_i^1(1-p_i) = \pi_i^1 \frac{i^2}{(i+1)^2}$, from which it follows by induction that $\pi_i^1 = \pi_{01}^1(1-p_0)/i^2$, for all $i \geq 1$. By symmetry $\pi_{ij}^1 = \pi_i^1/2^i$ for each $i \in \mathbb{N}$ and $1 \leq j \leq 2^i$. Therefore, for each $i \in \mathbb{N}$, $1 \leq j \leq 2^i$ we have $\pi_{ij}^1 = \pi_{01}^1(1-p_0)/(i^2 \cdot 2^i) = C/(i^2 \cdot 2^i)$ as was claimed. Moreover, for $i \geq 2$, $\pi_{ij}^2 = \pi_{ij}^1 \cdot p_i = \pi_{ij}^1 \cdot \frac{2i+1}{(i+1)^2}$. Combining with the expression for π_{ij}^1 gives $\pi_{ij}^2 = \frac{C}{i^2 \cdot 2^i} \cdot \frac{2i+1}{(i+1)^2}$. By induction, $\pi_{ij}^2 = \pi_{ij}^3 = \ldots = \pi_{ij}^i$, so this completes the proof. □

Note that for all $i \geq 1$ and $1 \leq j \leq 2^i$:

$$\pi_{ij} = \frac{C}{2^i \cdot i^2} + (i-1)\frac{C}{2^i \cdot i^2} \cdot \frac{2i+1}{(i+1)^2} \geq \frac{C}{2^i \cdot i^2}\ , \text{ and} \tag{19}$$

$$\pi_{ij} = \frac{C}{2^i \cdot i^2} + (i-1)\frac{C}{2^i \cdot i^2} \cdot \frac{2i+1}{(i+1)^2} \leq \frac{3C}{2^i \cdot i^2}\ . \tag{20}$$

Also note that for any $t \in \mathbb{N}$ and $i \geq 2t$ we have for each $1 \leq j \leq 2^i$:

(1) $\mathbb{P}(\overrightarrow{X}^t \in W_t | S_0 = \sigma_{ij}^k) = 1$, for $2 \leq k \leq \lceil i/2 \rceil + 1$.

(2) $\left(\sum_{k=2}^{i} \pi_{ij}^k\right)/\pi_{ij} \geq 1/3$ and $|\{k : 2 \leq k \leq \lceil i/2 \rceil + 1\}| \geq \frac{1}{2} \cdot |\{k : 2 \leq k \leq i\}|$. Hence, $\left(\sum_{k=2}^{\lceil i/2 \rceil + 1} \pi_{ij}^k\right)/\pi_{ij} \geq 1/6$.

Therefore, for each $t \in \mathbb{N}$:

$$\mathbb{P}(\overrightarrow{X}^t \in W_t | S_0 \in R_{ij}) \geq 1/6\ ,\ \text{for all } i \geq 2t \text{ and } 1 \leq j \leq 2^i\ . \tag{21}$$

Equations (19), (20), and (21) will be used in the proof of Claim 3 below, along with the following simple lemma.

Lemma 1 (Integral Test)**.** *Let $n \in \mathbb{N}$ and let $f : [n, \infty] \to \mathbb{R}$ be a positive, continuous, monotone-decreasing function, then:*

$$\int_n^\infty f(x)dx \leq \sum_{k=n}^{\infty} f(k) \leq f(n) + \int_n^\infty f(x)dx\ .$$

Claim 3. *$\mathbb{P}(W_t)$ decays roughly as $1/t$. More exactly, $C/12t \leq \mathbb{P}(W_t) \leq 6C/t$ for all $t \in \mathbb{N}$.*

Proof. For any state σ_{ij}^k with $i < t$, $\mathbb{P}(\overrightarrow{X}^t \in W_t | S_0 = \sigma_{ij}^k) = 0$. Thus, we have:

$$\begin{aligned}
\mathbb{P}(W_t) &= \mathbb{P}(\overrightarrow{X}^t \in W_t) \\
&= \sum_{i=t}^{\infty} \sum_{j=1}^{2^i} \mathbb{P}(S_0 \in R_{ij}) \cdot \mathbb{P}(\overrightarrow{X}^t \in W_t | S_0 \in R_{ij}) \\
&= \sum_{i=t}^{\infty} 2^i \cdot \mathbb{P}(S_0 \in R_{i1}) \cdot \mathbb{P}(\overrightarrow{X}^t \in W_t | S_0 \in R_{i1})\ ,
\end{aligned} \tag{22}$$

where the final equality follows from symmetry. We prove the bounds from above and below on $\mathbb{P}(W_t)$ separately using Equation (22).

- *Bound from below*:

$$\begin{aligned}\mathbb{P}(W_t) &= \sum_{i=t}^{\infty} 2^i \cdot \mathbb{P}(S_0 \in R_{i1}) \cdot \mathbb{P}(\overrightarrow{X}^t \in W_t | S_0 \in R_{i1}) \\ &\geq \sum_{i=2t}^{\infty} 2^i \cdot \mathbb{P}(S_0 \in R_{i1}) \cdot \mathbb{P}(\overrightarrow{X}^t \in W_t | S_0 \in R_{i1}) \\ &\overset{(a)}{\geq} \sum_{i=2t}^{\infty} 2^i \cdot \frac{C}{2^i \cdot i^2} \cdot \frac{1}{6} \\ &= \frac{C}{6} \sum_{i=2t}^{\infty} \frac{1}{i^2} \\ &\overset{(b)}{\geq} \frac{C}{6} \int_{2t}^{\infty} \frac{1}{x^2} dx \\ &= \frac{C}{12t} \,. \end{aligned} \tag{23}$$

  Here, (a) follows from Equations (19) and (21) and (b) from Lemma 1.

- *Bound from above*:

$$\begin{aligned}\mathbb{P}(W_t) &= \sum_{i=t}^{\infty} 2^i \cdot \mathbb{P}(S_0 \in R_{i1}) \cdot \mathbb{P}(\overrightarrow{X}^t \in W_t | S_0 \in R_{i1}) \\ &\overset{(a)}{\leq} \sum_{i=t}^{\infty} 2^i \cdot \frac{3C}{2^i \cdot i^2} \cdot 1 \\ &= 3C \sum_{i=t}^{\infty} \frac{1}{i^2} \\ &\overset{(b)}{\leq} 3C \left(\frac{1}{t^2} + \int_{t}^{\infty} \frac{1}{x^2} dx \right) \\ &= 3C \cdot \left(\frac{1}{t^2} + \frac{1}{t} \right) \\ &\leq \frac{6C}{t} \,. \end{aligned} \tag{24}$$

  Here, (a) follows from Equation (20) and (b) from Lemma 1. □

Claim 4. $\mathbb{P}(X_t \in \{2,3\} | \overrightarrow{X}^t = w) \geq 1/150$, *for all* $t \in \mathbb{N}$ *and* $w \in W_t$.

Proof. Applying Claim 3 we have for any $t \in \mathbb{N}$:

$$\begin{aligned}\mathbb{P}(X_t \in \{2,3\} | \overrightarrow{X}^t \in W_t) &= \mathbb{P}(\overrightarrow{X}^{t+1} \in W_{t+1} | \overrightarrow{X}^t \in W_t) \\ &= \mathbb{P}(\overrightarrow{X}^{t+1} \in W_{t+1}, \overrightarrow{X}^t \in W_t) / \mathbb{P}(\overrightarrow{X}^t \in W_t) \\ &= \mathbb{P}(\overrightarrow{X}^{t+1} \in W_{t+1}) / \mathbb{P}(\overrightarrow{X}^t \in W_t) \\ &\geq \frac{C/12(t+1)}{6C/t} \\ &= \frac{1}{72} \cdot \frac{t}{t+1} \\ &\geq \frac{1}{150} \,. \end{aligned}$$

By symmetry, $\mathbb{P}(X_t \in \{2,3\}|\overrightarrow{X}^t = w)$ is the same for each $w \in W_t$. Thus, the same bound must also hold for each $w \in W_t$ individually: $\mathbb{P}(X_t \in \{2,3\}|\overrightarrow{X}^t = w) \geq 1/150$ for all $w \in W_t$. □

Claim 5. *For each $t \in \mathbb{N}$ and $w \in W_t$,*
(i) $\tilde{h}_w \leq 1/300$ and
(ii) $h_w \geq 1/150$.
Hence, $h_w - \tilde{h}_w \geq 1/300$.

Proof of (i). $h_{\sigma_{ij}^k} = 0$, for all $i \geq 1$, $1 \leq j \leq 2^i$, and $k \geq 2$. And, for each $w \in W_t$, $\phi(w)_{\sigma_{ij}^1} = 0$, for all $i \geq 1$ and $1 \leq j \leq 2^i$. Hence, for each $w \in W_t$, $\tilde{h}_w = \sum_{\sigma \in \mathcal{S}} \phi(w)_\sigma h_\sigma = \phi(w)_{\sigma_{01}^1} h_{\sigma_{01}^1}$. By construction of the machine $h_{\sigma_{01}^1} \leq 1/300$ and, clearly, $\phi(w)_{\sigma_{01}^1}$ can never exceed 1. Thus, $\tilde{h}_w \leq 1/300$ for all $w \in W_t$. □

Proof of (ii). Let the random variable Z_t be defined by: $Z_t = 1$ if $X_t \in \{2,3\}$ and $Z_t = 0$ if $X_t \notin \{2,3\}$. By Claim 4, $\mathbb{P}(Z_t = 1|\overrightarrow{X}^t = w) \geq 1/150$ for any $w \in W_t$. Also, by symmetry, the probabilities of a 2 or a 3 following any word $w \in W_t$ are equal, so $\mathbb{P}(X_t = 2|\overrightarrow{X}^t = w, Z_t = 1) = \mathbb{P}(X_t = 3|\overrightarrow{X}^t = w, Z_t = 1) = 1/2$. Therefore, for any $w \in W_t$:

$$\begin{aligned} h_w &= H[X_t|\overrightarrow{X}^t = w] \\ &\geq H[X_t|\overrightarrow{X}^t = w, Z_t] \\ &\geq \mathbb{P}(Z_t = 1|\overrightarrow{X}^t = w) \cdot H[X_t|\overrightarrow{X}^t = w, Z_t = 1] \\ &\geq 1/150 \cdot 1 \ . \end{aligned}$$

□

Claim 6. *For each $t \in \mathbb{N}$ and $w \in \mathcal{L}_t$, $h_w - \tilde{h}_w \geq 0$.*

Proof. For $w \in \mathcal{L}_t$, let $P_w = \mathbb{P}(X_t|\overrightarrow{X}^t = w)$ denote the probability distribution over the next output symbol after observing the word w. Also, for $\sigma \in \mathcal{S}$, let $P_\sigma = \mathbb{P}(X_t|S_t = \sigma)$ denote the probability distribution over the next output symbol given that the current state is σ. Then, by concavity of the entropy function $H[\cdot]$, we have that for any $w \in \mathcal{L}_t$:

$$h_w \equiv H[P_w] = H\left[\sum_{\sigma \in \mathcal{S}} \phi(w)_\sigma \cdot P_\sigma\right] \geq \sum_{\sigma \in \mathcal{S}} \phi(w)_\sigma \cdot H[P_\sigma] = \sum_{\sigma \in \mathcal{S}} \phi(w)_\sigma h_\sigma \equiv \tilde{h}_w.$$

□

Claim 7. *The quantity $h(t) - h$ decays at a rate no faster than $1/t$. More exactly, $h(t+1) - h \geq \frac{C}{3600t}$, for all $t \in \mathbb{N}$.*

Proof. As noted above, since the machine satisfies the conditions of Proposition 1, the entropy rate is given by Equation (8) and the difference $h(t+1) - h$ is given by Equation (14). Therefore, applying Claims 3, 5, and 6 we may bound this difference $h(t+1) - h$ as follows:

$$\begin{aligned} h(t+1) - h &= \sum_{w \in \mathcal{L}_t} \mathbb{P}(w)(h_w - \tilde{h}_w) \\ &\geq \sum_{w \in W_t} \mathbb{P}(w)(h_w - \tilde{h}_w) \\ &\geq \mathbb{P}(W_t) \cdot \frac{1}{300} \\ &\geq \frac{C}{3600t} \,. \end{aligned}$$

□

With the above decay on $h(t)$ established we easily see the Branching Copy Process must have infinite excess entropy.

Proposition 3. *The excess entropy* $\mathbf{E}$ *for the BC Process is infinite.*

Proof. $\mathbf{E} = \sum_{t=1}^{\infty} (h(t) - h)$. By Claim 7, this sum must diverge. □

4. Conclusions

Any stationary, finite-alphabet process can be presented by a stationary HMM with an uncountable state set. Thus, there exist stationary HMMs with uncountable state sets capable of generating infinitary, finite-alphabet processes. It is impossible, however, to have a finite-state, stationary HMM that generates an infinitary process. The excess entropy $\mathbf{E}$ is always bounded by the entropy in the stationary distribution $H[\pi]$, which is finite for any finite-state HMM. Countable-state HMMs are intermediate between the finite and uncountable cases, and it was previously not shown whether infinite excess entropy was possible in this case, or not. We have demonstrated that it is indeed possible, by giving two explicit constructions of finite-alphabet, infinitary processes generated by stationary HMMs with countable state sets.

The second example, the Branching Copy Process, is also ergodic—a strong restriction. It is a priori quite plausible that infinite $\mathbf{E}$ might only occur in the countable-state case for nonergodic processes. Moreover, both HMMs we constructed are unifilar, so the ϵ-machines [12,16] of the processes have countable state sets as well. Again, unifilarity is a strong restriction to impose, and it is a priori conceivable that infinite $\mathbf{E}$ might only occur in the countable-state case for nonunifilar HMMs. Our examples have shown, though, that infinite $\mathbf{E}$ is possible for countable-state HMMs, even if one requires both ergodicity and unifilarity.

Following the original release of the above results [9] two additional examples of both ergodic and nonergodic infinitary, finite-alphabet processes with countable-state HMM presentations appeared [10]. For these examples it was shown that the mutual information $\mathbf{E}(t) = I[\overleftarrow{X}^t; \overrightarrow{X}^t]$ between length-t blocks diverges as a power law. Whereas, in our nonergodic example it diverges sublogarithmically and in our ergodic example, presumably, at most logarithmically. The ergodic example given in [10] is also somewhat simpler than ours. However, the HMM presentation for the ergodic process there is not unifilar and, moreover, one does not expect the ϵ-machine for this process to have a countable state set either.

Taking this all into account leaves open the question: Is power law divergence of $\mathbf{E}(t)$ possible for ergodic processes with unifilar, countable-state HMM presentations?

Acknowledgments

The authors thank Lukasz Debowski for helpful discussions. Nicholas F. Travers was partially supported on a National Science Foundation VIGRE fellowship. This material is based upon work supported by, or in part by, the US Army Research Laboratory and the US Army Research Office under grant number W911NF-12-1-0288 and the Defense Advanced Research Projects Agency (DARPA) Physical Intelligence project via subcontract No. 9060-000709. The views, opinions, and findings here are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the DARPA or the Department of Defense.

Author Contribution

Nicholas F. Travers and James P. Crutchfield designed research; Nicholas F. Travers performed research; Nicholas F. Travers and James P. Crutchfield wrote the paper. Both authors read and approved the final manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Del Junco, A.; Rahe, M. Finitary codings and weak Bernoulli partitions. *Proc. AMS* **1979**, *75*, doi:10.2307/2042753 .
2. Crutchfield, J.P.; Packard, N.H. Symbolic dynamics of one-dimensional maps: Entropies, finite precision, and noise. *Int. J. Theor. Phys.* **1982**, *21*, 433.
3. Grassberger, P. Toward a quantitative theory of self-generated complexity. *Int. J. Theor. Phys.* **1986**, *25*, 907–938.
4. Lindgren, K.; Norhdal, M.G. Complexity measures and cellular automata. *Complex Syst.* **1988**, *2*, 409–440.
5. Bialek, W.; Nemenman, I.; Tishby, N. Predictability, complexity, and learning. *Neural Comput.* **2001**, *13*, 2409–2463.
6. Debowski, L. Excess entropy in natural language: Present state and perspectives. *Chaos* **2011**, *21*, 037105.
7. Crutchfield, J.P.; Feldman, D.P. Regularities unseen, randomness observed: Levels of entropy convergence. *Chaos* **2003**, *13*, 25–54.
8. Ebeling, W. Prediction and entropy of nonlinear dynamical systems and symbolic sequences with LRO. *Physica D* **1997**, *109*, 42–52.
9. Travers, N.F.; Crutchfield, J.P. Infinite excess entropy processes with countable-state generators. **2011**, arXiv:1111.3393.

10. Debowski, L. On hidden Markov processes with infinite excess entropy. *J. Theor. Probab.* **2012**, doi: 10.1007/s10959-012-0468-6.
11. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley: New York, NY, USA, 2006.
12. Löhr, W. *Models of Discrete Time Stochastic Processes and Associated Complexity Measures*. Ph.D Thesis, Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany, 2010.
13. Travers, N.F.; Crutchfield, J.P. Asymptotic synchronization for finite-state sources. *J. Stat. Phys.* **2011**, *145*, 1202–1223.
14. Travers, N.F.; Crutchfield, J.P. Exact synchronization for finite-state sources. *J. Stat. Phys.* **2011**, *145*, 1181–1201.
15. Debowski, L. A general definition of conditional information and its application to ergodic decomposition. *Stat. Probab. Lett.* **2009**, *79*, 1260–1268.
16. Crutchfield, J.P.; Young, K. Inferring statistical complexity. *Phys. Rev. Lett.* **1989**, *63*, 105–108.

Appendix

We prove Proposition 1 from Section 2.2, which states that the entropy rate of any countable-state, exact, unifilar HMM is given by the standard formula:

$$h = H[X_0|S_0] = \sum_{\sigma \in \mathcal{S}} \pi_\sigma h_\sigma \ . \tag{25}$$

Proof. Let $\mathcal{L}_t = \{w : |w| = t, \mathbb{P}(w) > 0\}$ be the set of length t words in the process language $\mathcal{L}$, and let $\phi(w)$ be the conditional state distribution induced by a word $w \in \mathcal{L}_t$: *i.e.*, $\phi(w)_\sigma = \mathbb{P}(S_t = \sigma | \overrightarrow{X}^t = w)$. Furthermore, let $\tilde{h}_w = \sum_\sigma \phi(w)_\sigma h_\sigma$ be the $\phi(w)$-weighted average entropy in the next symbol given knowledge of the current state σ. And, let $h_w = H[X_t|\overrightarrow{X}^t = w]$ be the entropy in the next symbol after observing the word w. Note that:

(1) $h(t+1) = H[X_t|\overrightarrow{X}^t] = \sum_{w \in \mathcal{L}_t} \mathbb{P}(w) h_w$, and

(2) $\sum_\sigma \pi_\sigma h_\sigma = \sum_\sigma \left(\sum_{w \in \mathcal{L}_t} \mathbb{P}(w) \phi(w)_\sigma \right) h_\sigma = \sum_{w \in \mathcal{L}_t} \mathbb{P}(w) \left(\sum_\sigma \phi(w)_\sigma h_\sigma \right) = \sum_{w \in \mathcal{L}_t} \mathbb{P}(w) \tilde{h}_w$.

Thus, since we know $h(t)$ limits to h, it suffices to show that:

$$\lim_{t \to \infty} \sum_{w \in \mathcal{L}_t} \mathbb{P}(w)(h_w - \tilde{h}_w) = 0 \ . \tag{26}$$

Now, for any for any $w \in \mathcal{L}_t$, we have $|h_w - \tilde{h}_w| \leq \log |\mathcal{X}|$. However, for a *synchronizing word* $w = w_1 ... w_t$ with $H[S_t|\overrightarrow{X}^t = w] = 0$, $h_w - \tilde{h}_w$ is always 0, since the distribution $\phi(w)$ is concentrated only on a single state. Combining these two facts gives the estimate:

$$\left| \sum_{w \in \mathcal{L}_t} \mathbb{P}(w)(h_w - \tilde{h}_w) \right| \leq \sum_{w \in \mathcal{L}_t} \mathbb{P}(w) \cdot |h_w - \tilde{h}_w| \leq \log |\mathcal{X}| \cdot \mathbb{P}(NS_t) \ , \tag{27}$$

where NS_t is the set of length-t words that are nonsynchronizing and $\mathbb{P}(NS_t)$ is the combined probability of all words in this set. Since the HMM is exact, we know that for a.e. infinite future $\overrightarrow{x}$ an observer

will synchronize exactly at some finite time $t = t(\overrightarrow{x})$. And, since it is unifilar, the observer will remain synchronized for all $t' \geq t$. It follows that $\mathbb{P}(NS_t)$ must be monotonically decreasing and limit to 0:

$$\lim_{t \to \infty} \mathbb{P}(NS_t) = 0 \,. \tag{28}$$

Combining Equation (27) with Equation (28) shows that Equation (26) does in fact hold, which completes the proof. □