Shannon Entropy Rate of Hidden Markov Processes

Alexandra M. Jurgens* and James P. Crutchfield†
Complexity Sciences Center, Physics Department
University of California at Davis
Davis, California 95616

(Dated: October 6, 2020)

Hidden Markov chains are widely applied statistical models of stochastic processes, from fundamental physics and chemistry to finance, health, and artificial intelligence. The hidden Markov processes they generate are notoriously complicated, however, even if the chain is finite state: no finite expression for their Shannon entropy rate exists, as the set of their predictive features is generically infinite. As such, to date one cannot make general statements about how random they are nor how structured. Here, we address the first part of this challenge by showing how to efficiently and accurately calculate their entropy rates. We also show how this method gives the minimal set of infinite predictive features. A sequel addresses the challenge's second part on structure.

Keywords: Markov process, Shannon entropy, iterated function system, mixed state, predictive feature, optimal prediction, Blackwell measure

I. INTRODUCTION

Randomness is as necessary to physics as determinism. Indeed, since Henri Poincaré's failed attempt to establish the orderliness of planetary motion, it has been understood that both determinism and randomness are essential and unavoidable in the study of physical systems [1–4]. In the 1960s and 1970s, the rise of dynamical systems theory and the exploration of statistical physics of critical phenomena offered up new perspectives on this duality. The lesson was that intricate structures in a system's state space amplify uncertainty, guiding it and eventually installing it—paradoxically—in complex spatiotemporal patterns. Accepting this state of affairs prompts basic, but as-yet unanswered questions. How is this emergence monitored? How do we measure a system's randomness or quantify its patterns and their organization?

The tools needed to address these questions arose over recent decades during the integration of Turing's computation theory [5–7], Shannon's information theory [8], and Kolmogorov's dynamical systems theory [9–13]. This established the vital role that information plays in physical theories of complex systems. In particular, the application of hidden Markov chains to model and analyze the randomness and structure of physical systems has seen considerable success, not only in complex systems [14], but also in coding theory [15], stochastic processes [16], stochastic thermodynamics [17], speech recognition [18], computational biology [19, 20], epidemiology [21], and finance [22], to offer a nonexhaustive list of examples.

A highly useful property of certain hidden Markov chains (HMCs) is *unifilarity* [23], a structural constraint on their state transitions. Shannon showed that given a process generated by a finite-state unifilar HMC, one may directly and accurately calculate a process' irreducible randomness [8]—now called the *Shannon entropy rate*. Furthermore, for such a process, there is a unique minimal finite-state unifilar HMC that generates the process [24], known as the *ϵ-machine*. The ϵ-machine states—the process' *causal states*—are the minimal set of maximally predictive features. One consequence of the ϵ-machine's uniqueness and minimality is that its mathematical description gives a constructive definition of a process' structural complexity as the amount of memory required to generate the process.

Loosening the unifilar constraint to consider a wider class of generated processes, however, leads to major roadblocks. Predicting a process generated by a *finite-state* nonunifilar HMC requires an *infinite set* of causal states [25]. That is, though "finitely" generated, the process cannot be predicted by any finite unifilar HMC. Practically, this precludes directly determining the process' entropy rate using Shannon's result. Although the problem of the entropy rate of HMCs is well studied [15, 26–31], fully quantifying the randomness and structure of HMCs in general is an open problem.

That said, the causal states of an HMC are (in general, see Appendix B) equivalent to the uncountable set of *mixed states*, or predictive features, formally introduced by Blackwell over a half century ago [27]. To date, working with infinite mixed-states required coarse-graining to produce a finite set of predictive features. Fortunately, the tradeoffs between resource constraints and predictive power induced by such coarse graining can be systematically laid out [32–34].

The following introduces an alternative and more direct approach to working with mixed states. It casts generating mixed states as a chaotic dynamical system—specifically, a (place dependent) *iterated function system* (IFS). This obviates analyzing the underlying HMC via coarse graining. Rather, the complex dynamics of the

* amjurgens@ucdavis.edu
† chaos@ucdavis.edu

new system directly captures the information-theoretic properties of the original process. Specifically, this allows exactly calculating the entropy rate of the process generated by the original nonunifilar finite-state HMC. Additionally, the IFS interpretation of the nonunifilar HMC provides new insight into the structure and complexity of infinite-state processes. This has direct application to the study of randomness and structure in a wide range of physical systems.

In point of fact, the following and its sequel [35] were proceeded by two companions that applied the theoretical results here to two, rather different, physical domains. The first analyzed the origin of randomness and structural complexity engendered by quantum measurement [36]. The second solved a longstanding problem on exactly determining the thermodynamic functioning of Maxwellian demons, aka information engines [37]. That is, the following and its sequel lay out the mathematical and algorithmic tools required to successfully analyze these applied problems. We believe the new approach is destined to find even wider applications.

Section II recalls the necessary background in stochastic processes, hidden Markov chains, and information theory. Section III reviews the needed results on iterated function systems; while Sec. IV develops mixed states and their dynamic—the mixed-state presentation. The main result connecting these then follows in Sec. V, showing that the mixed-state presentation is an IFS and that it produces an ergodic process. Section VI recalls Blackwell's theory, updating it for our present purpose of determining the entropy rate of any HMC. The Supplementary Materials provide background on the asymptotic equipartition property and minimality of the mixed states. They also constructively work through the results for several example nonunifilar HMCs. They close with the statistical error analysis underlying entropy-rate estimation.

II. HIDDEN MARKOV PROCESSES

A *stochastic process* $\mathcal{P}$ is a probability measure over a bi-infinite chain $\ldots X_{t-2}\, X_{t-1}\, X_t\, X_{t+1}\, X_{t+2} \ldots$ of random variables, each denoted by a capital letter. A particular *realization* $\ldots x_{t-2}\, x_{t-1}\, x_t\, x_{t+1}\, x_{t+2} \ldots$ is denoted via lowercase letters. We assume values x_t belong to a discrete alphabet $\mathcal{A}$. We work with blocks $X_{t:t'}$, where the first index is inclusive and the second exclusive: $X_{t:t'} = X_t \ldots X_{t'-1}$. $\mathcal{P}$'s measure is defined via the collection of distributions over blocks: $\{\Pr(X_{t:t'}) : t < t', t, t' \in \mathbb{Z}\}$.

To simplify the development, we restrict to stationary, ergodic processes: those for which $\Pr(X_{t:t+\ell}) = \Pr(X_{0:\ell})$ for all $t \in \mathbb{Z}$, $\ell \in \mathbb{Z}^+$. In such cases, we only need to consider a process's length-ℓ *word distributions* $\Pr(X_{0:\ell})$.

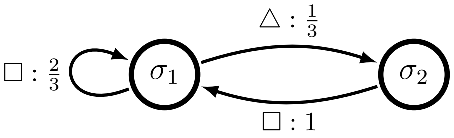


FIG. 1. A hidden Markov chain (HMC) with two states, $\{\sigma_1, \sigma_2\}$ and two symbols $\{\square, \triangle\}$. This machine is unifilar.

A *Markov process* is one for which $\Pr(X_t|X_{-\infty:t}) = \Pr(X_t|X_{t-1})$. A *hidden Markov process* is the output of a memoryless channel [38] whose input is a Markov process [16]. Working with processes directly is cumbersome, so we turn to consider finitely-specified mechanistic models that generate them.

Definition 1. *A finite-state edge-labeled* hidden MC *(HMC) consists of:*

1. *a finite set of states* $\boldsymbol{\mathcal{S}} = \{\sigma_1, ..., \sigma_N\}$,
2. *a finite alphabet* $\mathcal{A}$ *of* k *symbols* $x \in \mathcal{A}$, *and*
3. *a set of* N *by* N *symbol-labeled transition matrices* $T^{(x)}$, $x \in \mathcal{A}$: $T^{(x)}_{ij} = \Pr(\sigma_j, x|\sigma_i)$. *The corresponding overall state-to-state transitions are described by the row-stochastic matrix* $T = \sum_{x\in\mathcal{A}} T^{(x)}$.

Any given stochastic process can be generated by any number of HMCs. These are called a process' *presentations*.

We now introduce a structural property of HMCs that has important consequences in characterizing process randomness and structure.

Definition 2. *A* unifilar HMC *(uHMC) is an HMC such that for each state* $\sigma_i \in \boldsymbol{\mathcal{S}}$ *and each symbol* $x \in \mathcal{A}$ *there is at most one outgoing edge from state* σ_i *labeled with symbol* x.

Although there are many presentations for a process $\mathcal{P}$, there is a canonical presentation that is unique: a process' ϵ*-machine*.

Definition 3. *An* ϵ-machine *is a uHMC with* probabilistically distinct states*: For each pair of distinct states* $\sigma_i, \sigma_j \in \boldsymbol{\mathcal{S}}$ *there exists a finite word* $w = x_{0:\ell-1}$ *such that:*

$$\Pr(X_{0:\ell} = w|\mathcal{S}_0 = \sigma_k) \neq \Pr(X_{0:\ell} = w|\mathcal{S}_0 = \sigma_j) \ .$$

A process' ϵ-machine is its optimal, minimal presentation, in the sense that the set $\boldsymbol{\mathcal{S}}$ of predictive (or *causal*) states is minimal compared to all its other unifilar presentations [39].

A. Entropy Rate of HMCs

A process' intrinsic randomness is the information in the present measurement, discounted by having observed the information in an infinitely long history. It is measured by Shannon's source entropy rate [8].

Definition 4. *A process'* entropy rate h_μ *is the asymptotic average entropy per symbol [40]:*

$$h_\mu = \lim_{\ell \to \infty} \frac{\mathrm{H}[X_{0:\ell}]}{\ell} \ , \tag{1}$$

where $H[X_{0:\ell}]$ is the Shannon entropy of block $X_{0:\ell}$:

$$H[X_{0:\ell}] = - \sum_{x_{0:\ell} \in \mathcal{A}^\ell} \Pr(x_{0:\ell}) \log_2 \Pr(x_{0:\ell}) \ . \tag{2}$$

Given a finite-state unifilar presentation M_u of a process $\mathcal{P}$, we may directly calculate the entropy rate from the transition matrices of the uHMC [8]:

$$\begin{aligned} h_\mu(\mathcal{P}) &= h_\mu(M_u) \\ &= - \sum_{\sigma \in \boldsymbol{\mathcal{S}}} \Pr(\sigma) \sum_{x \in \mathcal{A}} T^{(x)}_{\sigma\sigma'} \log_2 T^{(x)}_{\sigma\sigma'} \ . \end{aligned} \tag{3}$$

Blackwell showed, though, that in general for processes generated by HMCs there is no closed-form expression for the entropy rate [27]. For a process generated by an nonunifilar HMC M, applying Eq. (3) to M typically overestimates the true entropy rate of the process $h_\mu(\mathcal{P})$:

$$h_\mu(M) \geq h_\mu(\mathcal{P}) \ .$$

Overcoming this limitation is one of our central results. We now embark on introducing the necessary tools for this.

III. ITERATED FUNCTION SYSTEMS

To get there, we must take a short detour to review iterated function systems (IFSs) [41], as they play a critical role in analyzing HMCs. Speaking simply, we show that HMCs are stochastic dynamical systems—namely, IFSs.

Let (Δ^N, d) be a compact metric space with $d(\cdot,\cdot)$ a distance. This notation anticipates our later application, in which Δ^N is N-simplex of discrete-event probability distributions (see Section IV A). However, the results here are general.

Let $f^{(x)} : \Delta^N \to \Delta^N$ for $x = 1, \ldots, k$ be a set of Lipschitz functions with:

$$d\left(f^{(x)}(\eta), f^{(x)}(\zeta)\right) \leq \tau^{(x)} d(\eta, \zeta) \ ,$$

for all $\eta, \zeta \in \Delta^N$ and where $\tau^{(x)}$ is a constant. This notation is chosen to draw an explicit parallel to the stochastic processes discussed in Section II and to avoid confusion with the lowercase Latin characters used for realizations of stochastic processes. In particular, note that the superscript (x) here and elsewhere parallels that of the HMC symbol-labeled transition matrices $T^{(x)}$. The reasons for this will soon become clear.

The Lipschitz constant $\tau^{(x)}$ is the *contractivity* of map $f^{(x)}$. Let $p^{(x)} : \Delta^N \to [0,1]$ be continuous, with $p^{(x)}(\eta) \geq 0$ and $\sum_{x=1}^k p^{(x)}(\eta) = 1$ for all η in M. The triplet $\{\Delta^N, \{p^{(x)}\}, \{f^{(x)}\} : x \in \mathcal{A}\}$ defines a *place-dependent* IFS.

A place-dependent IFS generates a stochastic process over $\eta \in \Delta^N$ as follows. Given an initial position $\eta_0 \in \Delta^N$, the probability distribution $\{p^{(x)}(\eta_0) : x = 1, \ldots, k\}$ is sampled. According to the sample x, apply $f^{(x)}$ to map η_0 to the next position $\eta_1 = f^{(x)}(\eta_0)$. Resample x from the distribution $p^{(x)}(\eta_1)$ and continue, generating $\eta_0, \eta_1, \eta_2, \ldots$.

If each map $f^{(x)}$ is a contraction—i.e., $\tau^{(x)} < 1$ for all $\eta, \zeta \in \Delta^N$—it is well known that there exists a unique nonempty compact set $\Lambda \subset \Delta^N$ that is invariant under the IFS's action:

$$\Lambda = \bigcap_{x=1}^{k} f^{(x)}(\Lambda) \ .$$

Λ is the IFS's *attractor.*

Consider the operator $V : M(\Delta^N) \to M(\Delta^N)$ on the space of Borel measures on the N-simplex:

$$V\mu(B) = \sum_{x=1}^{k} \int_{\left(f^{(x)}\right)^{-1}(B)} p^{(x)}(\eta) d\mu(\eta) \ . \tag{4}$$

A Borel probability measure μ is said to be *invariant* or *stationary* if $V\mu = \mu$. It is *attractive* if for any probability measure ν in $M(\Delta^N)$:

$$\int g d(V^n \nu) \to \int g \mu \ ,$$

for all g in the space of bounded continuous functions on Δ^N.

Let's recall here a key result concerning the existence of attractive, invariant measures for place-dependent IFSs.

Theorem 1. [42, Thm. 2.1] Suppose there exists $r < 1$ and $q > 0$ such that:

$$\sum_{x \in \mathcal{A}} p^{(x)}(\eta) d^q \left(f^{(x)}(\eta), f^{(x)}(\zeta)\right) \leq r^q d^q(\eta, \zeta) \ ,$$

for all $\eta, \zeta \in \Delta^N$. Assume that the modulus of uniform continuity of each $p^{(x)}$ satisfies Dini's condition and that there exists a $\delta > 0$ such that:

$$\sum_{x : d(f^{(x)}(\eta), f^{(x)}(\zeta)) \leq r d(\eta, \zeta)} p^{(x)}(\eta) p^{(x)}(\zeta) \leq \delta^2 \ , \tag{5}$$

for all $\eta, \zeta \in \Delta^N$. Then there is an attractive, unique, invariant probability measure for the Markov process generated by the place-dependent IFS.

In addition, under these same conditions Ref. [43] established an ergodic theorem for IFS *orbits*. That is, for any $\eta \in \Delta^N$ and $g : \Delta^N \to \Delta^N$:

$$\frac{1}{n+1} \sum_{k=0}^{n} g(w_{x_k} \circ \cdots \circ w_{x_1} \eta) \to \int g d\mu \ . \tag{6}$$

IV. MIXED-STATE PRESENTATION

We now return to stochastic processes and their HMC presentations. When calculating entropy rates from various presentations, we noted that HMC presentations led to difficulties: (i) the internal Markov-chain entropy-rate overestimates the process' entropy rate and (ii) there is no closed-form entropy-rate expression. To develop the tools needed to resolve these problems, we introduce HMC *mixed states* and their dynamic.

Assume that an observer has a finite HMC presentation M for a process $\mathcal{P}$. Since the process is hidden, the observer does not directly measure M's internal states. Absent output data, the best guess for M's hidden states is that they occur according to the state stationary distribution π. The observer can improve on this guess by monitoring the output data $x_0 \, x_1 \, x_2 \ldots$ that M generates. Given knowledge of M, determining the internal state from observed data is the problem of *observer-process synchronization*.

A. Mixed States

For a length-ℓ word w generated by M let $\eta(w) = \Pr(\boldsymbol{\mathcal{S}}|w)$ be the observer's *belief distribution* as to the process' current state after observing w:

$$\eta(w) \equiv \Pr(\mathcal{S}_\ell | X_{0:\ell} = w, \mathcal{S}_0 \sim \pi) \ . \tag{7}$$

When observing a N-state machine, the vector $\langle \eta(w)|$ lives in the *(N-1)-simplex* Δ^{N-1}, the set such that:

$$\{\eta \in \mathbb{R}^N : \langle \eta | \mathbf{1} \rangle = 1, \langle \eta | \delta_i \rangle \geq 0, i = 1, \ldots, N\} \ ,$$

where $\langle \delta_i | = \begin{pmatrix} 0 & 0 & \ldots & 1 & \ldots & 0 \end{pmatrix}$ and $|\mathbf{1}\rangle = \begin{pmatrix} 1 & 1 & \ldots & 1 \end{pmatrix}$. We use this notation to indicate components of the belief distribution vector η in order to avoid confusion with temporal indexing.

The 0-simplex Δ^0 is the single point $|\eta\rangle = (1)$, the 1-simplex Δ^1 is the line segment $[0,1]$ from $|\eta\rangle = (0,1)$ to $|\eta\rangle = (1,0)$, and so on.

The set of belief distributions $\eta(w)$ that an HMC can visit defines its set $\boldsymbol{\mathcal{R}}$ of *mixed states*:

$$\boldsymbol{\mathcal{R}} = \{\eta(w) : w \in \mathcal{A}^+, \Pr(w) > 0\} \ .$$

Generically, the mixed-state set $\boldsymbol{\mathcal{R}}$ for an N-state HMC is infinite, even for finite N [27].

Note that when a mixed state appears in probability expressions, the notation refers to the random variable η, not the row vector $|\eta\rangle$, and we drop the bra-ket notation. Bra-ket notation is used in vector-matrix expressions.

B. Mixed-State Dynamic

The probability of transitioning from $\langle \eta(w)|$ to $\langle \eta(wx)|$ on observing symbol x follows from Eq. (7) immediately; we have:

$$\Pr(\eta(wx)|\eta(w)) = \Pr(x|\mathcal{S}_\ell \sim \eta(w)) \ .$$

This defines the mixed-state transition dynamic $\mathcal{W}$. Together the mixed states and their dynamic define an HMC that is unifilar by construction. This is a process' *mixed-state presentation* (MSP) $\mathcal{U}(\mathcal{P}) = \{\boldsymbol{\mathcal{R}}, \mathcal{W}\}$.

We defined a process' $\mathcal{U}$ abstractly. The $\mathcal{U}$ typically has an uncountably infinite set of mixed states, making it challenging to work with in the form laid out in Section IV A. Usefully, however, given any HMC M that generates the process, we may explicitly write down the dynamic $\mathcal{W}$. Assume we have an $N+1$-state HMC presentation M with k symbols $x \in \mathcal{A}$. The initial condition is the invariant probability π over the states of M, so that $\langle \eta_0 | = \langle \delta_\pi |$. In the context of the mixed-state dynamic, mixed-state subscripts denote time.

The probability of generating symbol x when in mixed state η is:

$$\Pr(x|\eta) = \langle \eta | T^{(x)} | \mathbf{1} \rangle \ , \tag{8}$$

where $T^{(x)}$ is M's symbol-labeled transition matrix associated with the symbol x.

From η_0, we calculate the probability of seeing each $x \in \mathcal{A}$. Upon seeing symbol x, the current mixed state $\langle \eta_t |$ is updated according to:

$$\langle \eta_{t+1,x} | = \frac{\langle \eta_t | T^{(x)}}{\langle \eta_t | T^{(x)} | \mathbf{1} \rangle} \ . \tag{9}$$

Thus, given an HMC presentation we can restate Eq. (7) as:

$$\begin{aligned} \langle \eta(w) | &= \frac{\langle \eta_0 | T^{(w)}}{\langle \eta_0 | T^{(w)} | \mathbf{1} \rangle} \\ &= \frac{\langle \pi | T^{(w)}}{\langle \pi | T^{(w)} | \mathbf{1} \rangle} \ . \end{aligned}$$

Equation (9) tells us that, by construction, the MSP is unifilar, since each possible output symbol uniquely determines the next (mixed) state. Taken together, Eqs. (8)

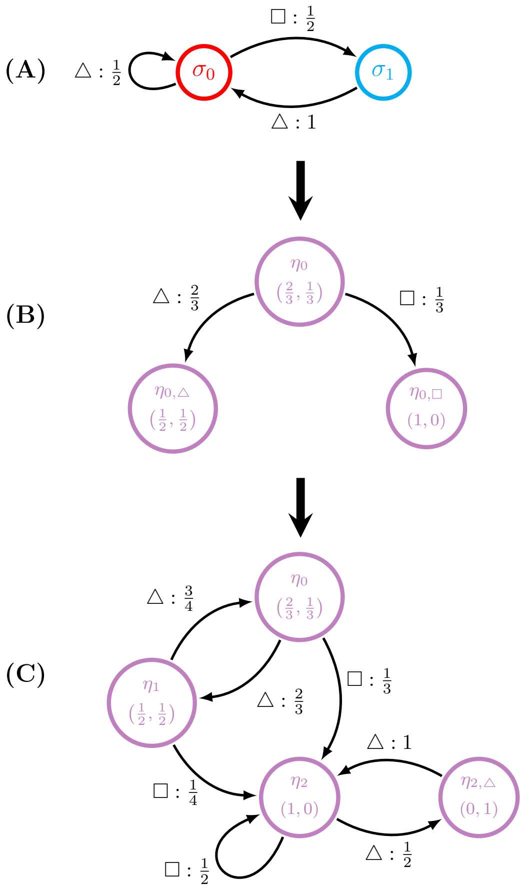


FIG. 2. Determining the mixed-state presentation (MSP) of the 2-state unifilar HMC shown in (A): The invariant state distribution $\pi = (2/3, 1/3)$. It becomes the first mixed state η_0 used in (B) to calculate the next set of mixed states. (C) The full set of mixed states seen from all allowed words. In this case, we recover the unifilar HMC shown in (A) as the MSP's recurrent states.

and (9) define the mixed-state transition dynamic $\mathcal{W}$ as:

$$\begin{aligned} \Pr(\eta_{t+1}, x|\eta_t) &= \Pr(x|\eta_t) \\ &= \langle \eta_t | \, T^{(x)} \, |\mathbf{1}\rangle \ , \end{aligned}$$

for all $\eta \in \boldsymbol{\mathcal{R}}$, $x \in \mathcal{A}$.

To find the MSP $\mathcal{U} = \{\boldsymbol{\mathcal{R}}, \mathcal{W}\}$ for a given HMC M we apply *mixed-state construction*:

1. Set $\mathcal{U} = \{\boldsymbol{\mathcal{R}} = \emptyset, \mathcal{W} = \emptyset\}$.
2. Calculate M's invariant state distribution: $\pi = \pi T$.
3. Take η_0 to be $\langle \delta_\pi |$ and add it to $\boldsymbol{\mathcal{R}}$.
4. For each current mixed state $\eta_t \in \boldsymbol{\mathcal{R}}$, use Eq. (8) to calculate $\Pr(x|\eta_t)$ for each $x \in \mathcal{A}$.
5. For $\eta_t \in \boldsymbol{\mathcal{R}}$, use Eq. (9) to find the updated mixed state $\eta_{t+1,x}$ for each $x \in \mathcal{A}$.
6. Add η_t's transitions to $\mathcal{W}$ and each $\eta_{t+1,x}$ to $\boldsymbol{\mathcal{R}}$, merging duplicate states.
7. For each new η_{t+1}, repeat steps 4-6 until no new mixed states are produced.

With the MSP $\mathcal{U}(M)$ in hand, the next issue is determining it's (equivalent) ϵ-machine. There are several cases.

Beginning with a finite, unifilar HMC M generating a process $\mathcal{P}$, the MSP $\mathcal{U}(M)$ is a finite, optimally-predictive rival presentation to $\mathcal{P}$'s ϵ-machine, as seen in Fig. 2. In this case, the starting HMC depicted in Fig. 2 (A) is an ϵ-machine, and reducing the MSP in Fig. 2 (C) by trimming the transient states returns the process' recurrent-state ϵ-machine. When starting with the ϵ-machine, trimming the resultant $\mathcal{U}(\epsilon\text{-machine})$ in this way always returns the ϵ-machine.

In general, if $\mathcal{U}(M)$ is finite, we find the ϵ-machine by minimizing $\mathcal{U}(M)$ via merging duplicate states: repeat mixed-state construction on $\mathcal{U}(M)$ and trim transient states once more. Minimizing countably-infinite and uncountably-infinite $\mathcal{U}(M)$ is discussed further in Appendix B.

The MSPs of unifilar presentations are interesting and contain additional information beyond the unifilar presentations. For example, containing transient causal states, they are employed in calculating many complexity measures that track convergence statistics [44].

However, here we focus on the mixed-state presentations of nonunifilar HMCs, which typically have an infinite mixed-state set $\boldsymbol{\mathcal{R}}$. Figure 3 illustrates applying mixed-state construction to a finite, nonunifilar HMC. This produces an infinite sequence of mixed states on $\Delta^1 = [0, 1]$, as plotted in Fig. 3(B). In this particular example, the MSP is highly structured and $\boldsymbol{\mathcal{R}}$ is countably infinite, allowing us to better understand the underlying process $\mathcal{P}$; compared, say, to the 2-state nonunifilar HMC in Fig. 3(A). MSPs of nonunifilar HMCs typically have an uncountably-infinite mixed-state set $\boldsymbol{\mathcal{R}}$.

V. MSP AS AN IFS

With this setup, our intentions in reviewing iterated function systems (IFSs) become explicit. The mixed-state presentation (MSP) exactly defines a place-dependent IFS, where the mapping functions are the set of symbol-labeled mixed-state update functions as given in Eq. (9) and the set of place-dependent probability functions are given by Eq. (8). We then have a mapping function and associated probability function for each symbol $x \in \mathcal{A}$ that can be derived from the symbol-labeled transition matrix $T^{(x)}$.

If these probability and mapping functions meet the conditions of Theorem 1, we identify the attractor Λ as the set of mixed states $\boldsymbol{\mathcal{R}}$ and the invariant measure μ as the invariant distribution π of the potentially infinite-state $\mathcal{U}$. This is the original HMC's *Blackwell measure.* Since all Lipschitz continuous functions are Dini continuous, the probability functions meet the conditions by

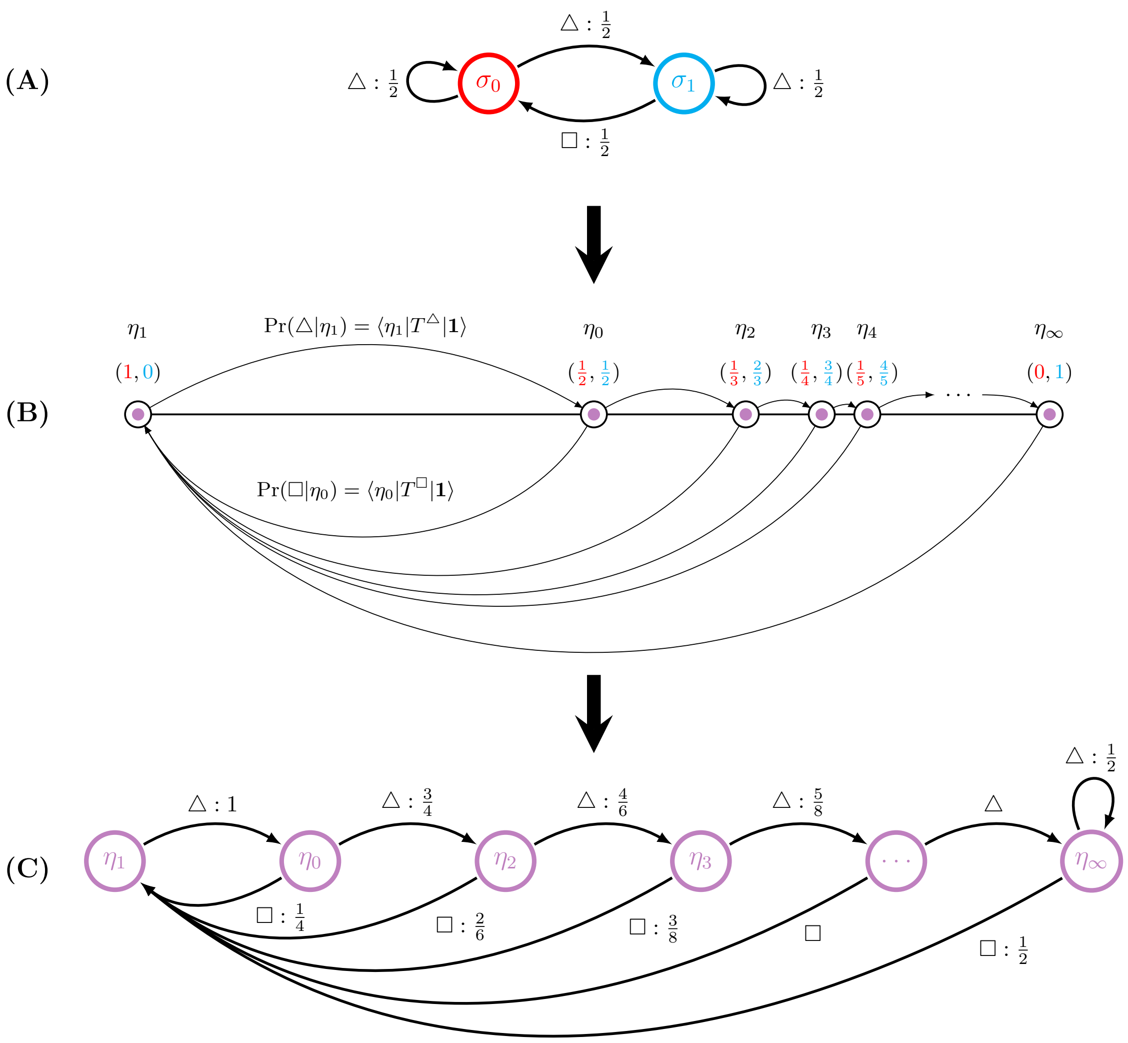

(A)
$\triangle : \frac{1}{2}$
σ_0
σ_1
$\triangle : \frac{1}{2}$
$\triangle : \frac{1}{2}$
$\square : \frac{1}{2}$
(B)
η_1
$\Pr(\triangle|\eta_1) = \langle \eta_1 | T^{\triangle} | \mathbf{1} \rangle$
η_0
η_2
η_3
η_4
η_∞
$(1, 0)$
$(\frac{1}{2}, \frac{1}{2})$
$(\frac{1}{3}, \frac{2}{3})$
$(\frac{1}{4}, \frac{3}{4})$
$(\frac{1}{5}, \frac{4}{5})$
$\cdots$
$(0, 1)$
$\Pr(\square|\eta_0) = \langle \eta_0 | T^{\square} | \mathbf{1} \rangle$
(C)
$\triangle : \frac{1}{2}$
$\triangle : 1$
$\triangle : \frac{3}{4}$
$\triangle : \frac{4}{6}$
$\triangle : \frac{5}{8}$
$\triangle$
η_1
η_0
η_2
η_3
$\cdots$
η_∞
$\square : \frac{1}{4}$
$\square : \frac{2}{6}$
$\square : \frac{3}{8}$
$\square$
$\square : \frac{1}{2}$

FIG. 3. Determining the mixed-state presentation of the 2-state *nonunifilar* HMC shown in (A). The invariant distribution $\pi = (1/2, 1/2)$. It is the first mixed state η_0 used in (B) to calculate the next set of mixed states. (B) plots the mixed states along the 1-simplex $\Delta^1 = [0, 1]$. In (C), we translated the points on the simplex to the states of an infinite-state, unifilar HMC.

inspection. We now establish that the maps are contractions, by appealing to Birkhoff's 1957 proof that a positive linear map preserving a convex cone is a contraction under the Hilbert projection metric [45].

Given an integer $N \geq 2$, let C^N be the nonnegative cone in $\mathbb{R}^N$, so that C^N consists of all vectors $z = (z_1, z_2, \ldots, z_N)$ satisfying $z \neq 0$ and $z_i \geq 0$ for all i. The *projective distance* $d : C^N \times C^N \to [0, \infty)$ is defined:

$$d(z, y) := \max \left\{ \left| \log \left(\frac{z_r}{z_s} \frac{y_s}{y_r} \right) \right| : r, s = 1, \ldots, N; r \neq s \right\} \qquad (10)$$

for $z, y \in C^N$, where $d(z, z) = 0$. If one of the points is on the cone boundary, the distance is taken to be $+\infty$. Note that the projective distance, by construction, defines $d(\alpha z, \beta y) = d(z, y)$, where $\alpha, \beta \in \mathbb{R}^+$. In other words, for two mixed states $\eta, \zeta \in \Delta^N$, $d\left(f^{(x)}(\eta), f^{(x)}(\zeta)\right) = d(\eta T^{(x)}, \zeta T^{(x)})$.

If $T^{(x)}$ is an $N \times N$ positive matrix, we have $d(zT^{(x)}, yT^{(x)}) < d_N(z, y)$ for every $z, y \in C^N$ such that $d(y, z) > 0$. We define the *projective contractivity* $\tau^{(x)}$ associated with $T^{(x)}$ as:

$$\tau_x := \sup_{\{z, y \in C_N : d(z,y) > 0\}} \frac{d(zT^{(x)}, yT^{(x)})}{d(z, y)} ,$$

so that $\tau^{(x)}$ satisfies $\tau^{(x)} \leq 1$. As the theorem below indicates, this inequality is strict.

Theorem 2. ([46, Thm. 1].) Let the integers $m, n \geq 2$ be arbitrary. For each matrix $T^{(x)} = \left[t_{ij}^{(x)} \right]$ of order $m \times n$

with positive components, $\tau^{(x)}$ is given by the following Birkhoff formula:

$$\tau^{(x)} = \frac{1 - \left(\phi^{(x)}\right)^{1/2}}{1 + \left(\phi^{(x)}\right)^{1/2}} ~,$$

where:

$$\phi(H) := \min_{r,s,j,k} \frac{t^{(x)}_{rj} t^{(x)}_{sk}}{t^{(x)}_{sj} t^{(x)}_{rk}} ~.$$

By inspection we see that $\phi(H) > 0$ and $\tau < 1$. As Ref. [47] notes, not only does the projective metric turn all positive linear transformations into contraction mappings, it is the only metric that does so.

Positivity of the transition matrix guarantees that any boundary points are mapped inside Δ^N. This is not generally true for our transition matrices—they are restricted merely to be nonnegative. However, the above result extends to any nonnegative matrix $T^{(x)}$ for which there exists an $N \in \mathbb{N}^+$ such that $\left(T^{(x)}\right)^N$ is a positive matrix. Then there will be a $\tau^{(x)} < 1$ such that $d\left(\eta \left(T^{(x)}\right)^N, \zeta \left(T^{(x)}\right)^N\right) < d(\eta, \zeta)$. This is equivalent to a requirement that $T^{(x)}$ be aperiodic and irreducible.

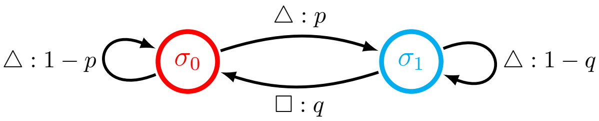


FIG. 4. Simple Nonunifilar Source (SNS): The symbol-labeled transition matrices given in Eq. (11) are both reducible, but the place-dependent IFS still has an attractor with an invariant probability distribution. By setting $p = q = 1/2$, we return the nonunifilar HMC from Fig. 3.

Still, we are not guaranteed irreducibility and aperiodicity for our symbol-labeled transition matrices. Indeed, the *Simple Nonunifilar Source*, depicted in Fig. 4, has the symbol-labeled transition matrices:

$$T^{(\triangle)} = \begin{pmatrix} 1-p & p \\ 0 & 1-q \end{pmatrix} \text{ and } T^{(\square)} = \begin{pmatrix} 0 & 0 \\ q & 0 \end{pmatrix} ~. \qquad (11)$$

Both $T^{(\triangle)}$ and $T^{(\square)}$ are reducible. A quick check is to examine Fig. 4 and ask if there is a length-n sequence consisting of only a single symbol that reaches every state from every other state. Nonetheless, the HMC has a countable set of mixed states $\boldsymbol{\mathcal{R}}$ and an invariant measure μ.

We can determine this from the mapping functions:

$$\begin{aligned} f^{(\triangle)}(\eta) = \Bigg[& \frac{\langle \eta \,|\delta_1\rangle\,(1-p)}{1-(1-\langle \eta \,|\delta_1\rangle)q}, \\ & \frac{\langle \eta \,|\delta_1\rangle\, p + (1-\langle \eta \,|\delta_1\rangle)(1-q)}{1+(1-\langle \eta \,|\delta_1\rangle)q} \Bigg] \text{ and } \quad (12) \\ f^{(\square)}(\eta) = [1,0] & ~. \qquad (13) \end{aligned}$$

Recall here that $\langle \eta \,|\delta_1\rangle$ is simply the first component of η. From any initial state η_0, other than $\eta_0 = \sigma_0 = [1,0]$, the probability of seeing a $\square$ is positive. Once a $\square$ is emitted, the mixed state is guaranteed to be $\eta = \sigma_0 = [1,0]$. In this case, when the mapping function is constant and the contractivity is $-\infty$, we call the symbol a *synchronizing* symbol. From σ_0, the set of mixed states is generated by repeated emissions of $\triangle$s, so that $\boldsymbol{\mathcal{R}} = \left\{ \left(f^{(\triangle)}\right)^n (\sigma_0) : n = 0, \ldots, \infty \right\}$. This is visually depicted in Fig. 3 for the specific case of $p = q = 1/2$. For all p and q, the measure can be determined analytically; see Ref. [48]. Note that this is due to the HMC's highly structured topology. In general, the set of mixed states is uncountable—either a fractal or continuous set—and the measure cannot be analytically expressed.

These arguments, including the uniqueness of the IFS attractor Λ, establish which HMC class generates ergodic processes: those whose total transition matrix $T = \sum_x T^{(x)}$ is nonnegative, irreducible, and aperiodic. Consider an HMC is in this class. Define for any word $w = x_1 \ldots x_\ell \in \mathcal{A}^+$ the associated mapping function $T^{(w)} = T^{(x_1)} \circ \cdots \circ T^{(x_\ell)}$. Consider word w in a process' typical set of realizations (see Appendix A), which set approaches measure one as $|w| \to \infty$. Due to ergodicity, it must be the case that $f^{(w)}$ is either (i) a constant mapping—and, therefore, infinitely contracting—or (ii) $T^{(w)}$ is irreducible.

As an example of the former case, we see that any composition of the SNS functions Eq. (13) is always a constant function, so long as there is at least one $\square$ in the word, the probability of which approaches one as the word grows in length.

As an example of the later case, imagine adding to the SNS in Fig. 4 a transition on $\square$ from σ_0 to σ_1. Then, both symbol-labeled transition matrices are still reducible, but the composite transition matrices for any word including both symbols is now irreducible. Therefore, the map is contracting. While this is not the case for words composed of all $\square$s and all $\triangle$s, these sequences are measure zero as $N \to \infty$. Appendix A discusses this further.

VI. ENTROPY OF GENERAL HMCS

Blackwell analyzed the entropy of *functions of finite-state Markov chains* [27]. With a shift in notation, functions of Markov chains can be identified as general hidden

Markov chains. This is to say, both presentation classes generate the same class of stochastic processes. As we have discussed, the entropy rate problem for unifilar hidden Markov chains is solved, with Shannon's entropy rate expression, Eq. (3). However, according to Blackwell, there is no analogous closed-form expression for the entropy rate of a nonunifilar HMC.

A. Blackwell Entropy Rate

That said, Blackwell gave an expression for the entropy rate of general HMCs, by introducing mixed states over stationary, ergodic, finite-state chains. (Although he does not refer to them as such.) His main result, retaining his notation, is transcribed here and adapted by us to constructively solve the HMC entropy-rate problem.

Theorem 3. ([27, Thm. 1].) Let $\{x_n, -\infty < n < \infty\}$ be a stationary ergodic Markov process with states $i = 1, \ldots, I$ and transition matrix $M = \|m(i,j)\|$. Let Φ be a function defined on $1, \ldots, I$ with values $a = 1, \ldots, A$ and let $y_n = \Phi(x_n)$. The entropy of the $\{y_n\}$ process is given by:

$$H = -\int \sum_a r_a(w) \log r_a(w) dQ(w) \ , \tag{14}$$

where Q is a probability distribution on the Borel sets of the set W of vectors $w = (w_1, \ldots, w_I)$ with $w_i \geq 0$, $\sum_i w_i = 1$, and $r_a(w) = \sum_{i=1}^{I} \sum_{j \ni \Phi(j)=a} w_i m(i,j)$. The distribution Q is concentrated on the sets $W_1, \ldots, W_A$, where W_a consists of all $w \in W$ with $w_i = 0$ for $\Phi(i) \neq a$ and satisfies:

$$Q(E) = \sum_a \int_{f_a^{-1}E} r_a(w) dQ(w) \ , \tag{15}$$

where f_a maps W into W_a, with the jth coordinate of $f_a(w)$ given by $\sum_i w_i m(i,j)/r_a(w)$ for $\Phi(j) = a$.

We can identify the w vectors in Theorem 3 as exactly the mixed states of Section IV. Furthermore, it is clear by inspection that $r_a(w)$ and $f_a(w)$ are the probability and mapping functions of Eqs. (8) and (9), respectively, with a playing the role of our observed symbol x.

Therefore, Blackwell's expression Eq. (14) for the HMC entropy rate, in effect, replaces the average over a finite set $\boldsymbol{\mathcal{S}}$ of unifilar states in Shannon's entropy rate formula Eq. (3) with (i) the mixed states $\boldsymbol{\mathcal{R}}$ and (ii) an integral over the Blackwell measure μ. In our notation, we write Blackwell's entropy formula as:

$$h_\mu^B = -\int_{\boldsymbol{\mathcal{R}}} d\mu(\eta) \sum_{x \in \mathcal{A}} p^{(x)}(\eta) \log_2 p^{(x)}(\eta) \ . \tag{16}$$

Thus, as with Shannon's original expression, this too uses unifilar states—now, though, states from the mixed-state presentation $\mathcal{U}$. This, in turn, maintains the finite-to-one internal (mixed-) state sequence to observed-sequence mapping. Therefore, one can identify the mixed-state entropy rate itself as the process' entropy rate.

B. Calculating the Blackwell HMC Entropy

Appealing to Ref. [43], we have that contractivity of our substochastic transition matrix mappings guarantees ergodicity over the words generated by the mixed-state presentation. And so, we can replace Eq. (16)'s integral over $\boldsymbol{\mathcal{R}}$ with a time average over a mixed-state trajectory $\eta_0, \eta_1, \ldots$ determined by a long allowed word, using Eqs. (8) and (9). This gives a new limit expression for the HMC entropy rate:

$$\widehat{h_\mu}^B = -\lim_{\ell \to \infty} \frac{1}{\ell} \sum_{t=0}^{\ell} \sum_{x \in \mathcal{A}} \Pr(x|\eta_\ell) \log_2 \Pr(x|\eta_\ell) \ , \tag{17}$$

where $\eta_\ell = \eta(w_{0:\ell})$ and $w_{0:\ell}$ is the first ℓ symbols of an arbitrarily long sequence $w_{0:\infty}$ generated by the process.

Note that $w_{0:\ell}$ will be a typical trajectory, if ℓ is sufficiently long. To remove convergence-slowing contributions from transient mixed states, one can ignore some number of the initial mixed states. The exact number of transient states that should be ignored is unknown in general. That said, it depends on the initial mixed state η_0, which is generally taken to be $\langle \delta_\pi |$, and the diameter of the attractor.

This completes our development of the HMC entropy rate. We now turn to practical issues of the resources needed for accurate estimation.

C. Data Requirements

Although we developed our HMC entropy-rate expression in terms of IFSs, determining a process' entropy rate can be recast as Markov chain Monte Carlo (MCMC) estimation. In MCMC, the mean of a function $f(x)$ of interest over a desired probability distribution $\pi(x)$ is estimated by designing a Markov chain with a stationary distribution π. For HMCs the desired distribution is the Blackwell measure μ, which is the stationary distribution μ over the MSP states $\boldsymbol{\mathcal{R}}$. Then, the Markov chain is simply the transition dynamic $\mathcal{W}$ over $\boldsymbol{\mathcal{R}}$.

With this setting, we estimate the entropy rate $\widehat{h_\mu}^B$ as the mean of the stochastic process defined by taking the entropy $H[X_\eta]$ over symbols emitted from state η for a sequence of mixed states generated by $\mathcal{W}$. In effect, we estimate the entropy rate as the mean of this stochastic process:

$$\widehat{h_\mu}^B = \mu_H \\ = \langle H[X_\eta] \rangle_\mu \ . \quad (18)$$

Mathematically, little has changed. The advantage, though, of this alternative description is that it invokes the extensive body of results on MCMC estimation. In this, it is well known that there are two fundamental sources of error in the estimation. First, there is that due to *initialization bias* or undesired statistical trends introduced by the initial transient data produced by the Markov chain before it reaches the desired stationary distribution. Second, there are errors induced by *autocorrelation in equilibrium*. That is, the samples produced by the Markov chain are correlated. And, the consequence is that statistical error cannot be estimated by $1/\sqrt{N}$, as done for N independent samples.

To address these two sources of error, we follow common MCMC practice, considering two "time scales" that arise during estimation. Consider the autocorrelation of the stationary stochastic process:

$$C_f(t) = \langle f_s f_{s+t} \rangle - \mu_f^2 \ ,$$

where μ_f is f's mean. Also, consider the *normalized* autocorrelation, defined:

$$\rho_f(t) = \frac{C_f(t)}{C_f(0)} \ .$$

If the autocorrelation decays exponentially with time, we define the *exponential autocorrelation time*:

$$\tau_{exp,f} = \lim_{t\to\infty} \sup \frac{1}{-\log|\rho_f(t)|}$$

and

$$\tau_{exp} = \sup_f \tau_{exp,f} \ .$$

So, τ_{exp} upper bounds the rate of convergence from an initial nonequilibrium distribution to the equilibrium distribution.

For a given observable, we also define the *integrated autocorrelation time* $\tau_{init,f}$ as:

$$\tau_{int,f} = \frac{1}{2} \sum_{-\infty}^{\infty} \rho_f(t) \ . \quad (19)$$

This relates the correlated samples selected by the chain to the variance of independent samples for the particular function f of interest. The variance of $f(x)$'s sample mean in MCMC is higher by a factor of $2\tau_{int,f}$. In other words, the errors for a sample of length N are of order $\sqrt{\tau_{int,f}/N}$. Thus, targeting 1% accuracy requires $\approx 10^4 \tau_{int,f}$ samples.

In practice, it is difficult to find τ_{exp} and τ_{int} for a generic Markov chain. There are two options. The first is to use numerical approximations that estimate the autocorrelation function, and therefore τ, from data. If we have the nonunifilar model in hand, it is a simple matter of sweeping through increasingly long strings of generated data until we observe convergence of the autocorrelation function.

Alternatively, taking inspiration from previous treatments of nonunifilar models, we make a finite-state approximation to the MSP by coarse-graining the simplex into boxes of length ϵ and employ a suitable method, such as Ulam's, to approximate the transition operator. Using methods previously discussed in Ref. [49], this allows calculating the autocorrelation function directly. Appendix D shows that that the approximation error vanishes as $\epsilon \to 0$.

The net result is that, being cognizant of the data requirements, entropy rate estimation is well behaved, convergent, and accurate.

VII. CONCLUSION

We opened this development considering the role that determinism and randomness play in the behavior of complex physical systems. A central challenge in this has been quantifying randomness, patterns, and structure and doing so in a mathematically-consistent but calculable manner. For well over a half a century Shannon entropy rate has stood as the standard by which to quantify randomness in a time series. Until now, however, calculating it for processes generated by nonunifilar HMCs has been difficult, at best.

We began our analysis of this problem by recalling that, in general, hidden Markov chains that are not unifilar have no closed-form expression for the Shannon entropy rate of the processes they generate. Despite this, these HMCs can be *unifilarized* by calculating the mixed states. The resulting mixed-state presentations are themselves HMCs that generate the process. However, adopting a unifilar presentation comes at a heavy cost: Generically, they are infinite state and so Shannon's expression cannot be used. Nonetheless, we showed how to work constructively with these mixed-state presentations. In particular, we showed that they fall into a common class of dynamical system. The mixed-state presentation is an iterated function system. Due to this, a number of results from dynamical systems theory can be applied.

Specifically, analyzing the IFS dynamics associated with a finite-state nonunfilar HMC allows one to extract useful properties of the original process. For instance, we can easily find the entropy rate of the generated process from long orbits of the IFS. That is, one may select any arbitrary starting point in the mixed-state simplex and calculate the entropy over the IFS's place-dependent probability distribution. We evolve the mixed state according to the IFS and sequentially sample the entropy of the place-dependent probability distribution at each step. Using an arbitrarily long word and taking the mean of these entropies, the method converges on the process' entropy rate.

Although others consider the IFS-HMC connection [50, 51], our development expanded previous work to include consideration of the role of the mixed-state presentation's in calculating the entropy rate and its connection to existing approaches to randomness and structure in complex systems. In particular, while our results focused on quantifying and calculating a process' randomness, we left open questions of pattern and structure. However, the path to achieving the results introduced here strongly suggests that the mixed-state presentation offers insight into answering these questions. For instance, Fig. 3 demonstrated how the highly structured nature of the Simple Nonunifilar Source is made topologically explicit through calculating its mixed-state presentation—which is also its ϵ-machine.

Though space will not let us develop it further here, this connection is not spurious. Indeed, many information-theoretic properties of the underlying process may be directly extracted from its mixed-state presentation. This follows from our showing how the attractor of the IFS defined by an HMC is exactly the set of mixed states $\boldsymbol{\mathcal{R}}$ of that HMC. These sets are often fractal in nature and quite visually striking. See Fig. S2 for several examples.

The sequel [35] to this development establishes that the information dimension of the mixed-state attractor is exactly the divergence rate of the *statistical complexity* [24]—a measure of a process' structural complexity that tracks memory. Furthermore, the sequel introduces a method to calculate the information dimension of the mixed-state attractor from the Lyapunov spectrum of the mixed-state IFS. In this way, it demonstrates that coarse-graining the simplex—the previous approach to study the structure of infinite-state processes—may be avoided altogether.

To close, we note that these structural tools and the entropy-rate method introduced here have already been put to practical use in two previous works. One diagnosed the origin of randomness and structural complexity in quantum measurement [36]. The other exactly determined the thermodynamic functioning of Maxwellian information engines [37], when there had been no previous method for this. At this point, however, we must leave the full explication of these techniques and further analysis on how mixed states reveal the underlying structure of processes generated by hidden Markov chains to the sequel [35].

ACKNOWLEDGMENTS

The authors thank David Gier, Ryan James, Sam Loomis, Sarah Marzen, Ariadna Venegas-Li, and Greg Wimsatt for helpful discussions and the Telluride Science Research Center for hospitality during visits and the participants of the Information Engines Workshops there. JPC acknowledges the kind hospitality of the Santa Fe Institute, Institute for Advanced Study at the University of Amsterdam, and California Institute of Technology for their hospitality during visits. This material is based upon work supported by, or in part by, FQXi Grant number FQXi-RFP-IPW-1902, and U.S. Army Research Laboratory and the U.S. Army Research Office under contract W911NF-13-1-0390 and grant W911NF-18-1-0028.

[1] D. Goroff, editor. *H. Poincaré, New Methods Of Celestial Mechanics, 1: Periodic And Asymptotic Solutions.* American Institute of Physics, New York, 1991.

[2] D. Goroff, editor. *H. Poincaré, New Methods Of Celestial Mechanics, 2: Approximations by Series.* American Institute of Physics, New York, 1993.

[3] D. Goroff, editor. *H. Poincaré, New Methods Of Celestial Mechanics, 3: Integral Invariants and Asymptotic Properties of Certain Solutions.* American Institute of Physics, New York, 1993.

[4] J. P. Crutchfield, N. H. Packard, J. D. Farmer, and R. S. Shaw. Chaos. *Sci. Am.*, 255:46 – 57, 1986.

[5] A. M. Turing. On computable numbers, with an application to the entsheidungsproblem. *Proc. Lond. Math. Soc. Ser. 2*, 42:230, 1936.

[6] C. E. Shannon. A universal Turing machine with two internal states. In C. E. Shannon and J. McCarthy, editors, *Automata Studies*, number 34 in Annals of Mathematical Studies, pages 157–165. Princeton University Press, Princeton, New Jersey, 1956.

[7] M. Minsky. *Computation: Finite and Infinite Machines.* Prentice-Hall, Englewood Cliffs, New Jersey, 1967.

[8] C. E. Shannon. A mathematical theory of communication. *Bell Sys. Tech. J.*, 27:379–423, 623–656, 1948.

[9] A. N. Kolmogorov. *Foundations of the Theory of Probability.* Chelsea Publishing Company, New York, second edition, 1956.

[10] A. N. Kolmogorov. Three approaches to the concept of the amount of information. *Prob. Info. Trans.*, 1:1, 1965.

[11] A. N. Kolmogorov. Combinatorial foundations of information theory and the calculus of probabilities. *Russ. Math. Surveys*, 38:29–40, 1983.

[12] A. N. Kolmogorov. Entropy per unit time as a metric invariant of automorphisms. *Dokl. Akad. Nauk. SSSR*, 124:754, 1959. (Russian) Math. Rev. vol. 21, no. 2035b.

[13] Ja. G. Sinai. On the notion of entropy of a dynamical system. *Dokl. Akad. Nauk. SSSR*, 124:768, 1959.

[14] J. P. Crutchfield. Between order and chaos. *Nature Physics*, 8(January):17–24, 2012.

[15] B. Marcus, K. Petersen, and T. Weissman, editors. *Entropy of Hidden Markov Process and Connections to Dy-*

namical Systems, volume 385 of *Lecture Notes Series*. London Mathematical Society, 2011.

[16] Y. Ephraim and N. Merhav. Hidden Markov processes. *IEEE Trans. Info. Th.*, 48(6):1518–1569, 2002.

[17] J. Bechhoefer. Hidden Markov models for stochastic thermodynamics. *New. J. Phys.*, 17:075003, 2015.

[18] L. R. Rabiner and B. H. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, January:4–16, 1986.

[19] E. Birney. Hidden Markov models in biological sequence analysis. *IBM J. Res. Dev.*,, 45(3.4):449–454, 2001.

[20] S. Eddy. What is a hidden Markov model? *Nature Biotech.*, 22:1315–1316, Oct 2004.

[21] C. Bretó, D. He, E. L. Ionides, and A. A. King. Time series analysis via mechanistic models. *Ann. App. Statistics*, 3(1):319–348, Mar 2009.

[22] T. Rydén, T. Teräsvirta, and S. Åsbrink. Stylized facts of daily return series and the hidden Markov model. *J. App. Econometrics*, 13:217–244, 1998.

[23] R. B. Ash. *Information Theory*. John Wiley and Sons, New York, 1965.

[24] J. P. Crutchfield and K. Young. Inferring statistical complexity. *Phys. Rev. Let.*, 63:105–108, 1989.

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

[26] J. J. Birch. Approximations for the entropy for functions of Markov chains. *Ann. Math. Statist.*, 33(2):930–938, 1962.

[27] D. Blackwell. The entropy of functions of finite-state Markov chains. In *Transactions of the first Prague conference on information theory, Statistical decision functions, Random processes*, volume 28, pages 13–20, Prague, Czechoslovakia, 1957. Publishing House of the Czechoslovak Academy of Sciences.

[28] S. Egner, V. B. Balakirsky, L. Tolhuizen, S. Baggen, and H. Hollmann. On the entropy rate of a hidden markov model. In *International Symposium on Information Theory, 2004. ISIT 2004. Proceedings.*, pages 12–, 2004.

[29] G. Han and B. Marcus. Analyticity of entropy rate of hidden Markov chains. *IEEE Trans. Info. Th.*, 52(12):5251–5266, 2006.

[30] E. Ordentlich and T. Weissman. New bounds on the entropy rate of hidden Markov processes. In C. N. Georghiades, S. Verdu, R. Calderbank, and A. Orlitsky, editors, *2004 IEEE Information Theory Workshop Proceedings*, pages 117–122, Piscataway, New Jersey, 24-29 October 2004. Institute of Electrical and Electronics Engineers.

[31] P. Jacquet, G. Seroussi, and W. Szpankowskic. On the entropy of a hidden Markov process. *Theo. Comp. Sci.*, 395:203–2019, 2008.

[32] F. Creutzig, A. Globerson, and N. Tishby. Past-future information bottleneck in dynamical systems. *Phys. Rev. E*, 79(4):041925, 2009.

[33] S. Still, J. P. Crutchfield, and C. J. Ellison. Optimal causal inference: Estimating stored information and approximating causal architecture. *CHAOS*, 20(3):037111, 2010.

[34] S. Marzen and J. P. Crutchfield. Predictive rate-distortion for infinite-order Markov processes. *J. Stat. Phys.*, 163(6):1312–1338, 2014.

[35] A. Jurgens and J. P. Crutchfield. Infinite complexity of finite state hidden Markov processes. *in preparation*, 2020.

[36] A. Venegas-Li, A. Jurgens, and J. P. Crutchfield. Measurement-induced randomness and structure in quantum dynamics. *Physical Review E*, in press, 2020. arXiv:1908.09053.

[37] A. Jurgens and J. P. Crutchfield. Functional thermodynamics of Maxwellian ratchets: Constructing and deconstructing patterns, randomizing and derandomizing behaviors. *Phys. Rev. Research*, 2(3):033334, 2020.

[38] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, second edition, 2006.

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

[40] J. P. Crutchfield and D. P. Feldman. Regularities unseen, randomness observed: Levels of entropy convergence. *CHAOS*, 13(1):25–54, 2003.

[41] M. Barnsley. *Fractals Everywhere*. Academic Press, New York, 1988.

[42] M. F. Barnsley, S. G. Demko, J. H. Elton, and J. S. Geronimo. Invariant measures arising from iterated function systems with place dependent probabilities. *Ann. Inst. H. Poincare*, 24:367–394, 1988.

[43] J. H. Elton. An ergodic theorem for iterated maps. *Ergod. Th. Dynam. Sys.*, 7:481–488, 1987.

[44] J. P. Crutchfield, P. Riechers, and C. J. Ellison. Exact complexity: Spectral decomposition of intrinsic computation. *Phys. Lett. A*, 380(9-10):998–1002, 2016.

[45] G. Birkhoff. Extensions of Jentzsch's theorem. *Trans. Am. Math. Soc.*, 85(1):219–227, 1957.

[46] R. Cavazos-Cadena. An alternative derivation of Birkhoff's formula for the contraction coefficient of a positive matrix. *Linear Algebra Apps.*, 375:291–297, 2003.

[47] E. Kohlberg and J. W. Pratt. The contraction mapping approach to the Perron-Frobenius theory: Why Hilbert's metric? *Math. Oper. Res.*, 7(2), 1982.

[48] S. E. Marzen and J. P. Crutchfield. Nearly maximally predictive features and their dimensions. *Phys. Rev. E*, 95(5):051301(R), 2017.

[49] P. Riechers and J. P. Crutchfield. Spectral simplicity of apparent complexity, Part II: Exact complexities and complexity spectra. *Chaos*, 28:033116, 2018.

[50] M. Rezaeian. Hidden Markov process: A new representation, entropy rate and estimation entropy. *arXiv:0606114*.

[51] W. Słomczyński, J. Kwapień, and K. Życzkowski. Entropy computing via integration over fractal measures. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 10(1):180–188, Mar 2000.

[52] A. Jurgens and J. P. Crutchfield. Minimal embedding dimension of minimally infinite hidden Markov processes. *in preparation*, 2020.

Supplementary Materials

The Shannon Entropy Rate of Hidden Markov Processes

Alexandra Jurgens and James P. Crutchfield

arXiv:2002.XXXXX

The Supplementary Materials to follow review the notion of typical sets of realizations in a stochastic process, discuss minimality of infinite-state mixed-state presentations, determine the entropy rates of a suite of example hidden Markov chains with infinite mixed-state presentations, and give details of errors that arise when estimating autocorrelation.

Appendix A: Asymptotic Equipartition and Typical Set Contraction

The *asymptotic equipartition property* (AEP) states that for a discrete-time, ergodic, stationary process X:

$$-\frac{1}{n}\log_2 \Pr(X_1, X_2, \ldots, X_n) \to h_\mu(X) \ , \tag{S1}$$

as $n \to \infty$ [38]. This effectively divides the set of sequences into two sets: the *typical set*—sequences for which the AEP holds—and the atypical set, for which it does not. As a consequence of the AEP, it must be the case that the typical set is measure one in the space of all allowed realizations and all sequences in the atypical set approach measure zero as $n \to \infty$.

We argue that while our IFS class includes reducible maps, any composition of maps corresponding to a word in the typical set will be irreducible. This can be seen intuitively by considering the SNS, shown in Fig. 4, and adding an additional transition on a $\square$ from σ_0 to σ_1. This produces an HMC with two reducible symbol-labeled transition matrices, but an irreducible total transition matrix. However, as $|w| \to \infty$, the only words such that $T^{(w)}$ remains reducible are $\square^N$ and $\triangle^N$. We can see that these words cannot possibly be in the typical set, since $-\frac{1}{n}\log_2 \Pr(\square^n) = -\log_2 \Pr(\square) \neq h_\mu(X)$. The entropy rate h_μ is by definition the branching entropy averaged over the mixed states. And so, any word that visits only a restricted subset of the mixed states—i.e., a word with a reducible transition matrix—cannot approach h_μ, regardless of length. Therefore, only words with an irreducible mapping will be in the typical set, implying that there exists an integer word length $|w| > 0$ for which words without a contractive mapping are measure zero.

Appendix B: Minimality of $\mathcal{U}(M)$

The minimality of infinite-state mixed-state presentations $\mathcal{U}(M)$ is an open question. Mixed-state presentations are not guaranteed to be minimal. In fact, it is possible to construct MSPs with an uncountably-infinite number of states for a process that requires only one state to optimally predict [52].

A proposed solution to this problem is a short and simple check on *mergeablility* of mixed states, which here refers to any two distinct mixed states that have the same conditional probability distribution over future strings; i.e., any two mixed states η_0 and ζ_0 for which:

$$\Pr(X_{0:L}|\eta_0) = \Pr(X_{0:L}|\zeta_0) \ , \tag{S1}$$

for all $L \in \mathbb{N}^+$.

Although minimality does not impact the entropy-rate calculation, one benefit of the IFS formalization of the MSP is the ability to directly check for duplicated states and therefore determine if the MSP is nonminimal. We

$$\begin{tikzcd}
\Delta^N \arrow[r, "f^{(x)}"] \arrow[d, "P"] & \Delta^N \arrow[d, "P"] \\
\Delta^k \arrow[r, "g^{(x)}"] & \Delta^k
\end{tikzcd}$$

FIG. S1. Commuting diagram for probability functions $P = \{p^{(x)}\}$, mixed-state mapping functions $f^{(x)}$, and proposed symbol-distribution mapping functions $g^{(x)}$.

check this by considering, for an $N+1$ state machine M with alphabet $\mathcal{A} = \{0, 1, \ldots, k\}$, the dynamic not only over mixed states, but probability distributions over symbols. Let:

$$P(\eta) = \left(p^{(0)}(\eta), \ldots, p^{(k-1)}(\eta)\right) \tag{S2}$$

and consider Fig. S1. For each mixed state $\eta \in \Delta^N$, Eq. (S2) gives the corresponding probability distribution $\rho(\eta) \in \Delta^k$ over the symbols $x \in \mathcal{A}$. Let M emit symbol x, then the dynamic from one such probability distribution $\rho \in \Delta^k$ to the next is given by:

$$\begin{aligned} g^{(x)}(\rho_t) &= P \circ f^{(x)} \circ P^{-1}(\rho) \\ &= \rho_{t+1,x} \ . \end{aligned} \tag{S3}$$

From this, we see that if Eq. (S2) is invertible, $g^{(x)} : \Delta^k \to \Delta^k$ is well defined and has the same functional properties as $f^{(x)}$. In other words, in this case, it is not possible to have two distinct mixed states $\eta, \zeta \in \Delta^N$ with the same probability distribution over symbols. And, the probability distributions can only converge under the action of $g^{(x)}$ if the mixed states also converge under the action of $f^{(x)}$. This, and the implications for identifying the embedding dimension of minimal generators, is discussed further in our companion work [52].

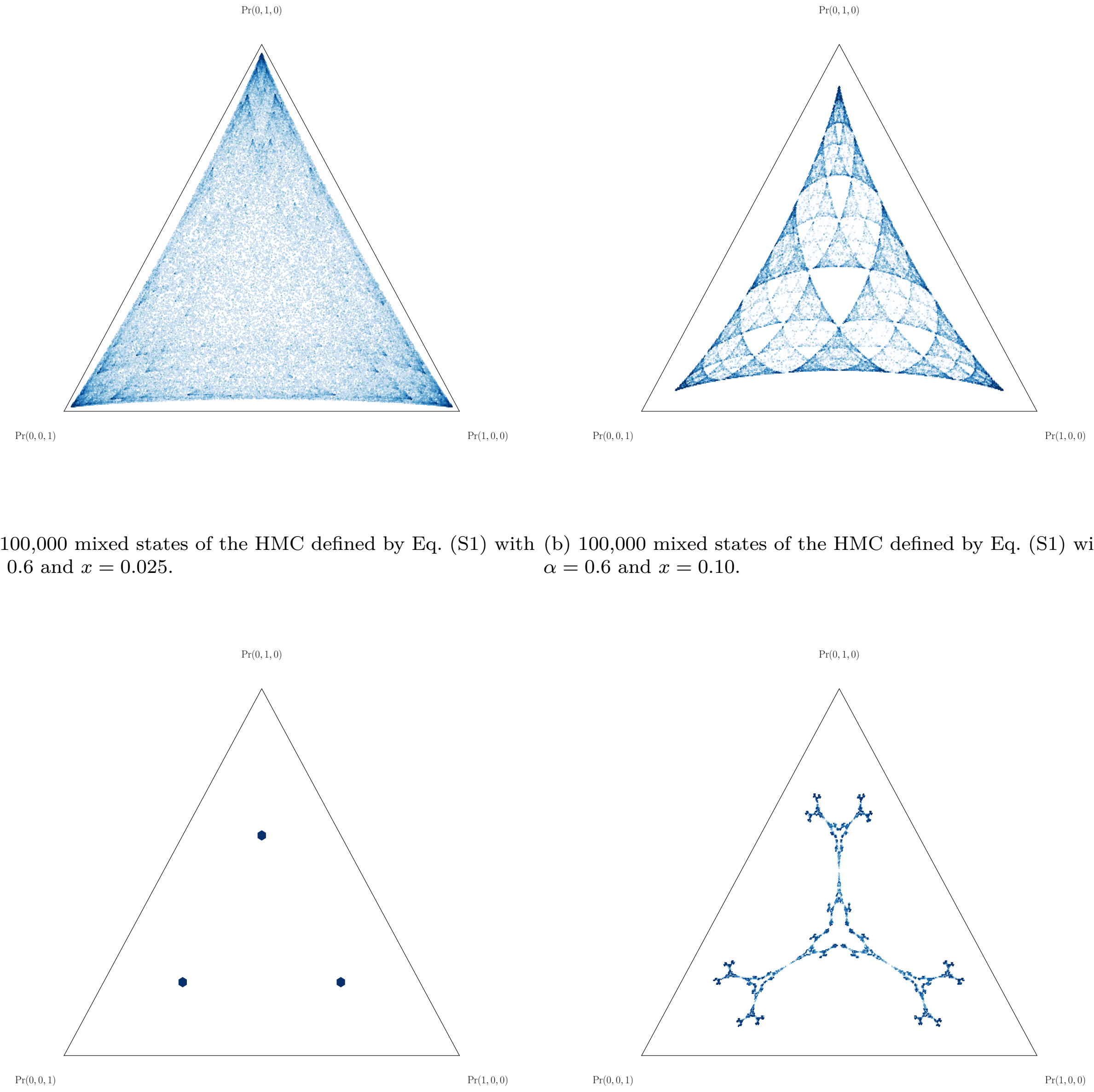
Appendix C: Parametrized HMCs and Their MSPs

As an example of the broad applicability of this entropy calculation method, consider an HMC with 3 symbols and 3 states [48]:

$$T^{\square} = \begin{pmatrix} \alpha y & \beta x & \beta x \\ \alpha x & \beta y & \beta x \\ \alpha x & \beta x & \beta y \end{pmatrix}, \quad T^{\triangle} = \begin{pmatrix} \beta y & \alpha x & \beta x \\ \beta x & \alpha y & \beta x \\ \beta x & \alpha x & \beta y \end{pmatrix}, \text{ and } \quad T^{\circ} = \begin{pmatrix} \beta y & \beta x & \alpha x \\ \beta x & \beta y & \alpha x \\ \beta x & \beta x & \alpha y \end{pmatrix} \ , \tag{S1}$$

with $\beta = (1-\alpha)/2$ and $y = 1 - 2x$. From inspection, we see that α can take on any value from 0 to 1 and x may range from 0 to 1/2.

Choosing $\alpha = 0.6$ and sweeping $x \in [0, 0.5]$ gives us an MSP that first fills nearly the entire simplex, with probability mass concentrated at the corners, then shrinks to a finite machine with 3 states at $x = 1/3$, and finally grows once again into a fractal measure, as Fig. S2 illustrates. To demonstrate the ease and efficiency calculating their entropy rates Fig. S3 plots h_μ as function of $(x, \alpha) \in [0, 0.5] \times [0, 1]$. It is an interesting side note that despite the wildly different structures on display in Fig. S2, we see a smoothly varying entropy rate that does not appear to be strongly affected by the underlying structure. This case and the expected impact of structure on the entropy rate more broadly will be discussed in further detail in the sequel.



(a) 100,000 mixed states of the HMC defined by Eq. (S1) with $\alpha = 0.6$ and $x = 0.025$.

(b) 100,000 mixed states of the HMC defined by Eq. (S1) with $\alpha = 0.6$ and $x = 0.10$.

(c) 100,000 mixed states of the HMC defined by Eq. (S1) with $\alpha = 0.6$ and $x = 0.33$.

(d) 100,000 mixed states of the HMC defined by Eq. (S1) with $\alpha = 0.6$ and $x = 0.49$.

FIG. S2. Parametrized 3-state HMC defined in Eq. (S1) that generates MSPs in a variety of structures, depending on x and α. However, due to the rotational symmetry in the transition matrices, the attractor is radially symmetric around the simplex center.

Appendix D: Estimation Errors for Finite-State Autocorrelation

Coarse-graining the mixed-state simplex into a set $\mathcal{C}$ of boxes of width ϵ, we may construct a finite-state approximation of the infinite-state MSP. It has been shown that given such an approximation, for any given box c, the bound on the difference in the entropy rate over the symbol distribution between the coarse-grained approximation and a

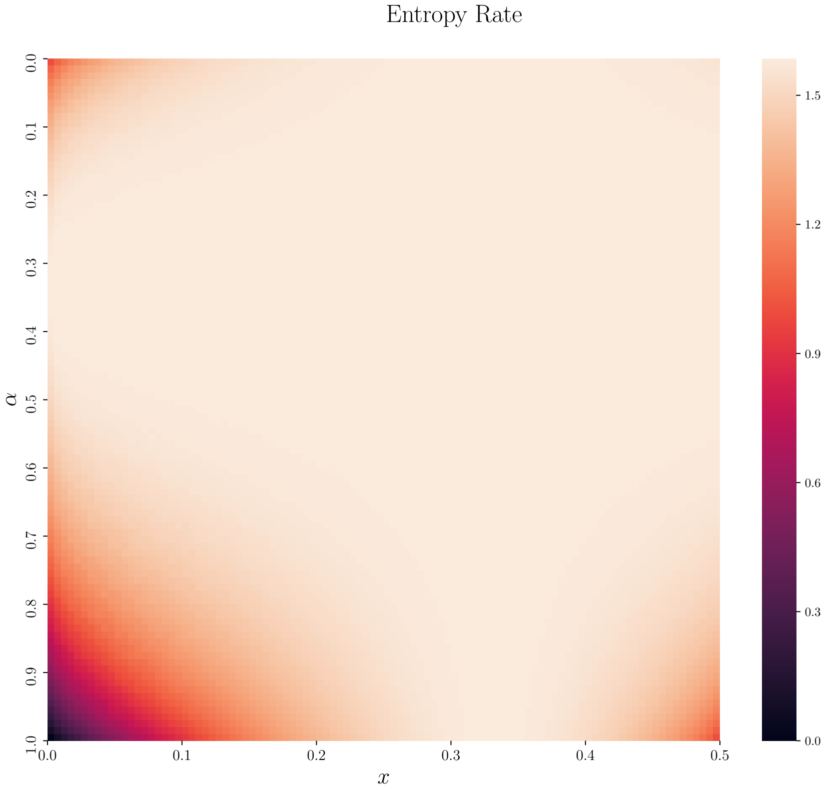


FIG. S3. Entropy rates of the parametrized HMC defined in Eq. (S1) over $x \in [0.0, 0.5]$ and $\alpha \in [0.0, 1.0]$.

mixed state within that box is bounded by [48]:

$$|H[X_0|\mathcal{C} = c] - H[X_0|\eta \in f]| \leq H_b\left(\frac{\sqrt{G}\epsilon}{2}\right) , \tag{S1}$$

where $H_b(\cdot)$ is the binary entropy function. Our task here is to consider the error in the autocorrelation in the sequence of mixed states since, if we can show that this is bounded, the error in the autocorrelation of the branching entropy must also be bounded.

At time zero, the autocorrelation is equal to $A(L = 0) = \langle X_0 \overline{X_0} \rangle$, so for the finite-state approximation, we have:

$$A_{\mathcal{C}}(L = 0) = \sum_i \pi_{\mathcal{C}}(i) c_i \overline{c_i} \ ,$$

where $\pi_{\mathcal{C}}$ is the stationary distribution over the coarse-grained mixed states, $\pi_{\mathcal{C}}(i)$ is the stationary probability of cell i, and c_i is the center of cell i. For the true process, we have:

$$\begin{aligned} A(L = 0) &= \int_{\boldsymbol{\mathcal{R}}} d\mu(\eta) \eta \overline{\eta} \\ &= \sum_i \pi_{\mathcal{C}}(i) \int_{\eta \in \mathcal{C}_i} d\mu(\eta|i) \eta \overline{\eta} \ , \end{aligned}$$

where $d\mu(\eta|i)$ is the distribution over mixed states within cell i. The maximum distance between any two mixed states in a cell i is bounded by:

$$\|\eta - \zeta\|_1 \leq \sqrt{G}\epsilon \ ,$$

the length of the longest diagonal in a hypercube of dimension $|G|$, by construction. Since the gradient of the L_2 norm is simply $\nabla\|\mathbf{x}\|_2 = \mathbf{x}/\|\mathbf{x}\|_2$, we have a bound on the difference in the autocorrelation at time zero:

$$|A_{\mathcal{C}}(L=0) - A(L=0)| \leq N\sqrt{|G|}\epsilon \ .$$

With increasing length we have:

$$A_{\mathcal{C}}(L) = \sum_i \pi_{\mathcal{C}}(i) c_i \sum_{w \in \mathcal{A}^L} \overline{f^{(w)}(c_i)} p^{(w)}(c_i)$$

and:

$$\begin{aligned} A(L) &= \int_{\boldsymbol{\mathcal{R}}} d\mu(\eta)\eta \sum_{w \in \mathcal{A}^L} \overline{f^{(w)}(\eta)} p^{(w)}(\eta) \\ &= \sum_i \pi_{\mathcal{C}}(i) \int_{\eta \in \mathcal{C}_i} d\mu(\eta|i)\eta \sum_{w \in \mathcal{A}^L} \overline{f^{(w)}(\eta)} p^{(w)}(\eta) \ . \end{aligned}$$

Let $\eta = c_i + \delta$ for some mixed state in cell i. Then we can write:

$$|A_{\mathcal{C}}(L) - A(L)| \leq \ \sum_i \pi_{\mathcal{C}}(i) \left[c_i \sum_{w \in \mathcal{A}^L} \overline{f^{(w)}(c_i)} p^{(w)}(c_i) - (c_i + \delta) \sum_{w \in \mathcal{A}^L} \overline{f^{(w)}(c_i + \delta)} p^{(w)}(c_i + \delta) \right] \ .$$

Now, note that:

$$p^{(w)}(c_i + \delta) \approx p^{(w)}(c_i) + \nabla p^{(w)}(c_i) \cdot \delta$$

and:

$$f^{(w)}(c_i + \delta) \approx f^{(w)}(c_i) + e^{\lambda^x} \delta$$

where λ^w is the leading Lyapunov exponent of the mapping function. Substituting this and eliminating terms of order δ^2 gives us:

$$|A_{\mathcal{C}}(L) - A(L)| \leq \ \sum_i \pi_{\mathcal{C}}(i) \left[c_i \sum_{w \in \mathcal{A}^L} \left(\overline{f^{(w)}(c_i)} \nabla p^{(w)} \cdot \delta + e^{\lambda^w} \overline{\delta} p^{(w)}(c_i) \right) + \delta \sum_{w \in \mathcal{A}^L} \overline{f^{(w)}(c_i)} p^{(w)}(c_i) \right] \ .$$

These terms identify three sources of approximation error: (i) that due to a difference in the probability distribution over symbols, (ii) that in the mapping functions, and (iii) that from approximating the points at the center of their cells.

For the first, we note that total variation in the probability distribution over symbols is bounded by the distance between the mixed states at which the distributions are computed. So, for any two mixed states in the same cell, $\| \Pr(X = x|\eta) - \Pr(X = x|\zeta)\|_{TV} \leq \sqrt{G}\epsilon$. Then, the first term is the error due to the difference in the expectation value of the next state, given that we have calculated the probability distribution at $c_i + \delta$, rather than c_i. Using Hölder's inequality, for two distributions over $P(X)$ and $Q(X)$, we may say:

$$\begin{aligned} E[f]_Q - E[f]_P &= \sum_x f(x)(P(x) - Q(x)) \\ \sum_x f(x)(P(x) - Q(x)) &\leq \sum_x f(x)|P(x) - Q(x)| \\ \sum_x f(x)|P(x) - Q(x)| &\leq \|f\|_p \|P - Q\|_q \ , \end{aligned}$$

where $1/p + 1/q = 1$. Setting $q = 1$:

$$E[f]_Q - E[f]_P \leq \|f\|_\infty \|P - Q\|_{TV} \ .$$

So, after taking the product with the cell centers c_i, we have that the first error is bounded by $N\sqrt{G}\epsilon$ at all lengths.

For the second, we note that since the maps are contractions, $\lambda < 0$, and the distance between $f^x(\eta)$ and $f^x(\zeta)$, where η and ζ are in the same cell i, is bounded by $\sqrt{G}\epsilon$. As the length of a word w grows, $\lambda^w \to -\infty$ and the distance $f^w(\eta) - f^w(\zeta) \to 0$. At large L, this term vanishes, at a rate equal to the average maximal Lyapunov exponent of the IFS.

The final error is that in the autocorrelation in the cell approximation which is, likewise, bounded by the cell size—this is the same error from $A(0)$, viz. $N\sqrt{G}\epsilon$.

And so, in combination with the bound on the entropy, we may say, loosely speaking, that the error in the autocorrelation vanishes as $\epsilon \to 0$. Therefore, to find τ and estimate the error in Eq. (18) as a function of sample size, we take finer coarse-grained approximations until convergence in the autocorrelation curve is observed, and then calculate τ directly.