Title
Divergent Predictive States: The Statistical Complexity Dimension of Stationary, Ergodic Hidden Markov Processes

Authors
Jurgens, Alexandra M
Crutchfield, James P

Publication Date
2021-02-20

Peer reviewed

Divergent Predictive States: The Statistical Complexity Dimension of Stationary, Ergodic Hidden Markov Processes

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

(Dated: February 23, 2021)

Inferring models from samples of stochastic processes is challenging, even in the most basic setting in which processes are stationary and ergodic. A principle reason for this, discovered by Blackwell in 1957, is that finite-state generators produce realizations with arbitrarily-long dependencies that require an uncountable infinity of probabilistic features be tracked for optimal prediction. This, in turn, means predictive models are generically infinite-state. Specifically, hidden Markov chains, even if finite, generate stochastic processes that are irreducibly complicated. The consequences are dramatic. For one, no finite expression for their Shannon entropy rate exists. Said simply, one cannot make general statements about how random they are and finite models incur an irreducible excess degree of unpredictability. This was the state of affairs until a recently-introduced method showed how to accurately calculate their entropy rate and, constructively, to determine the minimal set of infinite predictive features. Leveraging this, here we address the complementary challenge of determining how structured hidden Markov processes are by calculating the rate of statistical complexity divergence—the information dimension of the minimal set of predictive features.

Keywords: Hidden Markov process, hidden Markov model, iterated function system, optimal prediction, predictive feature, mixed-state presentation, Blackwell measure, spectrum of Lyapunov characteristic exponents, fractal dimension, information dimension

I. INTRODUCTION

A paradox lives at the heart of highly complex systems—the intricate patterns they generate arise through an interplay between determinism and stochasticity. Despite progress identifying and measuring their degrees of randomness and unpredictability, basic questions remain. Specifically, how do we quantify correlation and "structure"? Can we detect a system's emergent patterns and quantify their organization?

Clearly posing these questions and developing the tools to answer them required, over the recent decades, integrating Turing's computation theory [1–3], Shannon's information theory [4], and Kolmogorov's dynamical systems theory [5–9]. Together they highlighted the central role that *information*—its generation, transmission, and storage—plays in investigating complex systems. Drawing from the convergence, *computational mechanics* [10] introduced a definition of the structural organization and memory of stochastic processes—the *statistical complexity* which measures the number and distribution of optimally-predictive features.

Answers to the randomness-structure problem have been carefully outlined and successfully implemented for processes that can be optimally predicted with countably many predictive features [11–13]. However, many complex systems arising in engineering, physical, and biological systems [14–18] require an infinite number of predictive features. Somewhat soberingly, these truly complex systems are implicated in a range of natural phenomena, from the geophysics of earthquakes [19] and physiological measurements of neural avalanches [20] to semantics in natural language [21] and cascading failures in power transmission grids [22].

In point of fact, as first established by Blackwell in the 1950s [14], calculating the entropy rate of processes generated by discrete time, N-state hidden Markov chains (HMCs) requires tracking an uncountably-infinite set of distributions over an HMC's states. The following establishes that optimally predicting these processes must use these sets of distributions as predictive features. These sets, living on the $(N-1)$-simplex, are complex: generically highly ramified and self-similar, they support similarly-complicated measures [23]. This renders prediction of even simply-defined processes very challenging. Probing their structure is even more difficult. Nevertheless, the broad popularity and application of HMCs—not only in the study of complex systems [10], but also in coding theory [24], stochastic processes [25], stochastic thermodynamics [26], speech recognition [27], computational biology [28, 29], epidemiology [30], and finance [31]—gives testimony to the ubiquity of truly complex systems, in both theory and nature.

Our recent work [32] addressed this state of affairs, showing how to generate infinite sets of predictive features and accurately calculate the entropy rate for pro-

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

cesses generated by HMCs. This gave a new and efficient tool for consistently describing the randomness of truly complex systems. However, those results did not address the complementary side of the complex-system paradox—the structural aspect of the interplay between structure and randomness. Troublingly, for processes with uncountably infinite sets of predictive features, the statistical complexity diverges, substantially circumscribing its usefulness as a metric of system organization. As we will show, quantifying the structure of these truly complex systems requires a new approach and new tools and methods.

Historically, the need for such a measure of divergent information storage and Blackwell's discovery were perhaps anticipated by Shannon's definition in the 1940s of *dimension rate* [4]:

$$\lambda = \lim_{\delta \to 0} \lim_{\epsilon \to 0} \lim_{T \to \infty} \frac{N(\epsilon, \delta, T)}{T \log \epsilon} ,$$

where $N(\epsilon, \delta, T)$ is the smallest number of elements that may be chosen such that all elements of a trajectory ensemble generated over time T, apart from a set of measure δ, are within the distance ϵ of at least one chosen trajectory. This is the minimal "number of dimensions" required to specify a member of a trajectory (or message) ensemble. Unfortunately, Shannon devotes barely a paragraph to the concept, leaving it largely unmotivated and uninterpreted. Nonetheless, we take inspiration from this to develop the *statistical complexity dimension*, the asymptotic growth rate of the statistical complexity for an uncountably infinite set of predictive features.

We conjecture that the statistical complexity dimension is the same dimension rate proposed by Shannon. However, the following goes beyond Shannon's brief mention to provide constructive and accurate methods for determining this important system invariant. Technically, statistical complexity dimension is defined as the information dimension of the (self-similar set of) predictive states. Several distinct steps are involved. Determining this information dimension requires establishing ergodicity, calculating the Lyapunov spectrum of an HMC's mixed-state presentation, and applying a suitably modified version of the Lyapunov-information dimension conjecture from dynamical systems that connects the spectrum to the dimension.

To highlight the usefulness of these informational quantities, that otherwise appear rather abstracted from natural systems, it should be noted that the following and its predecessor [32] 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 [33]. The second solved a longstanding problem on exactly determining the thermodynamic functioning of Maxwellian demons, aka information engines [18]. That is, the predecessor and the present development (along with a sequel to be announced in the conclusion) lay out the mathematical and algorithmic tools required to successfully analyze structure and randomness in these applied problems. Taken together, we believe the new approach will find even wider use than in these application areas.

In the following, we introduce a practical and computable measure of structural complexity analogous to Shannon's dimension rate in the form of *statistical complexity dimension* d_μ. Section II recalls the necessary background in stochastic processes, hidden Markov chains, and information theory. Section III recounts mixed states and their dynamic—the mixed-state presentation—as well as the connection to iterated function systems (IFSs) previously demonstrated. The main result follows in Sec. IV, where the Lyapunov-information dimension conjecture is reviewed and updated to our needs and the statistical complexity dimension d_μ is introduced. Finally, in Sec. VI d_μ of a three-state parametrized HMC is calculated across a wide region of parameter space, demonstrating the insights afforded by and computational efficiency of our methods.

II. HIDDEN MARKOV PROCESSES

Our main objects of study are stochastic processes and the mechanisms that generate them—hidden Markov chains, mixed states, and the ϵ-machine. We touch on several of their important properties, including stationarity, ergodicity, randomness, and memory. Readers familiar with the previous work in this series [32] may skip to Section II D.

A. 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 X_t 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. 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 $t \in \mathbb{Z}$, $\ell \in \mathbb{Z}^+$, and for which individual realizations obey all of those statistics. In such cases, we only need to consider a process' length-ℓ *word distributions* $\Pr(X_{0:\ell})$.

A *Markov process* is one that exhibits memory over a single time step: $\Pr(X_t|X_{-\infty:t}) = \Pr(X_t|X_{t-1})$. A *hidden Markov process* is the output of a memoryless channel [34] whose input is a Markov process [25].

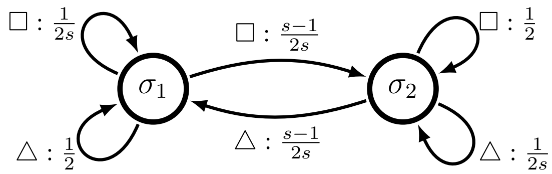


FIG. 1. A hidden Markov chain with two states $\boldsymbol{\mathcal{S}} = \{\sigma_1, \sigma_2\}$ and two symbols $\mathcal{A} = \{\square, \triangle\}$. It is nonunifilar and parametrized with $s \in [1, \infty)$. It becomes unifilar in the limit of $s \to \infty$.

B. Presentations

Directly working with processes—nominally, infinite sets of infinite sequences and their probabilities—is cumbersome. So, we turn to consider finitely-specified mechanistic models that generate them.

Definition 1. *A finite-state edge-labeled* hidden Markov chain *(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 associated state-to-state transitions are described by the row-stochastic matrix $T = \sum_{x \in \mathcal{A}} T^{(x)}$. The *internal-state Markov chain* is given by $\{\boldsymbol{\mathcal{S}}, T\}$. The asymptotic, stationary state distribution is $\pi = \{\Pr(\sigma), \sigma \in \boldsymbol{\mathcal{S}}\}$ and, as a vector, is given by T's left eigenvector normalized in probability: $\pi = \pi T$.

A 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 determining a 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.*

One consequence is that a uHMC's states are *predictive* in the sense that the distribution of words following a state is the same as $\mathcal{P}$'s word distribution conditioned on the words that lead to the state. (This need not hold for the states of nonunifilar HMCs.)

Although there can be many presentations for a process $\mathcal{P}$, there is a canonical presentation that is unique: a process' ϵ*-machine* [10].

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 optimally-predictive, minimal presentation, in the sense that the set $\boldsymbol{\mathcal{S}}$ of predictive states is minimal compared to all its other unifilar presentations. That said, $\boldsymbol{\mathcal{S}}$ may be finite, countably infinite, or uncountably infinite. By capturing a process' structure and not merely being predictive, an ϵ-machine's states are called *causal states*.

C. Process Intrinsic Randomness: HMC Entropy Rate

A process' intrinsic randomness is the information in the present measurement, discounted by having observed the preceding infinitely-long history. It is measured by Shannon's source entropy rate [4].

Definition 4. *A process'* entropy rate h_μ *is the asymptotic average Shannon entropy per symbol [11]:*

$$h_\mu = \lim_{\ell \to \infty} \mathrm{H}[X_{0:\ell}]/\ell \ , \quad (1)$$

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

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

Given a finite-state unifilar presentation M_u of a process $\mathcal{P}$, we may directly calculate the process' entropy rate from its uHMC's transition matrices [4]:

$$h_\mu = - \sum_{\sigma \in \mathcal{S}} \pi_\sigma \sum_{x \in \mathcal{A}} \Pr(x|\sigma) \log_2 \Pr(x|\sigma) \ . \quad (3)$$

In stark contrast, for processes generated by nonunifilar HMCs there is no closed-form expression for the entropy rate [14]. For these processes, the closed-form expression Eq. (3) applied to the HMC states and transition matrices substantially misestimates the generated process' entropy rate.

Addressing this nonunifilar case was the focus of our previous development [32]. We showed that the entropy rate of a general HMC may be determined using its *mixed states*; reviewed shortly in Section III. Tracking an HMC's mixed states allows one to find the entropy rate of the generated process and so the latter's intrinsic randomness.

D. Process Intrinsic Structure

A process' memory is determined using its ϵ-machine (minimal) presentation M. Depending on the specific need, this may be measured either in terms of the number $|\boldsymbol{\mathcal{S}}|$ of M's causal states or the amount of historical Shannon entropy they store—that is, the *statistical complexity* C_μ.

Definition 5. *A process'* statistical complexity *is the Shannon entropy stored in its ϵ-machine's causal states:*

$$C_\mu = \mathrm{H}[\Pr(\mathcal{S})] \\ = - \sum_{\sigma \in \mathcal{S}} \pi_\sigma \log_2 \pi_\sigma \ . \quad (4)$$

From the definitions above, a process' ϵ-machine is its smallest uHMC presentation, in the sense that both $|\boldsymbol{\mathcal{S}}|$ and C_μ are minimized by a process' ϵ-machine, compared to all other unifilar (predictive) presentations. Due to the ϵ-machine's minimality, we can identify the ϵ-machine's C_μ as the process' statistical complexity.

A challenge similar to that encountered with the entropy rate of nonunifilar HMCs arises: there is no closed-form expression for the C_μ of the processes they generate. We now turn to give a constructive answer to this challenge. The preceding presentation types, though, give a useful path to understanding how a process' different presentations help or hinder determining process properties. The strategy in the following turns on yet another presentation type. Here on in, with nothing else said, reference to an HMC means the general case—a nonunfilar HMC.

III. OBSERVER-PROCESS SYNCHRONIZATION

Previously, we introduced mixed-state presentations of HMCs and established their equivalence to random dynamical systems known as *iterated function systems* (IFSs) [32]. We now briefly review this construction. (Readers familiar with the previous results may skip to Section IV.)

Assume that an observer has a finite HMC presentation M for a process $\mathcal{P}$ that it is monitoring. Consider the *observer-process synchronization problem* in which the observer determines at each moment $\mathcal{P}$'s HMC state from observed process data. (An equivalent framing is that we assume M is generating $\mathcal{P}$ and the observer seeks to determine at each moment which state σ M is in.)

Since the process is hidden (and since M is an HMC), the observer cannot directly detect the state. The observer's best guess initially is that the states occur according to M's internal-state stationary distribution π. Using knowledge of M's structure, the observer then refines this guess by monitoring the output data $x_0 \, x_1 \, x_2 \ldots$ that M generates. If and when the observer knows with certainty in which state the process is, they have *synchronized* to the process.

A. Mixed State Presentation

1. Of A Process

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:

$$\eta(w) \equiv \Pr(\mathcal{S}_\ell | X_{0:\ell} = w, \mathcal{S}_0 \sim \pi) \ . \quad (5)$$

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 for components of the mixed state vector η to avoid confusion with temporal indexing.

Synchronization then occurs when a word w is observed such that $\Pr(\mathcal{S}_\ell = \sigma | X_{0:\ell} = w) = 1$, for one state σ.

The belief distributions $\eta(w)$ that an HMC can visit defines its set 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 [14]. Figure 2 shows a case where the HMC generates the Cantor set as its mixed state set.

The probability of transitioning from $\langle \eta(w)|$ to $\langle \eta(wx)|$ on observing symbol x follows from Eq. (5) 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) $\mathcal{U}(\mathcal{P}) = \{\boldsymbol{\mathcal{R}}, \mathcal{W}\}$.

2. Of an HMC

Above, we defined a process' $\mathcal{U}$ abstractly. However, given any HMC M that generates the process, we may explicitly write down the MSP $\mathcal{U}(M) = \{\boldsymbol{\mathcal{R}}, \mathcal{W}\}$. Assume we have an $N+1$-state HMC presentation M with k symbols $x \in \mathcal{A}$. We let the initial mixed state be the invariant probability π over the states of M, so $\langle \eta_0| = \langle \delta_\pi|$. In the context of the mixed-state dynamic, mixed-state subscripts denote time.

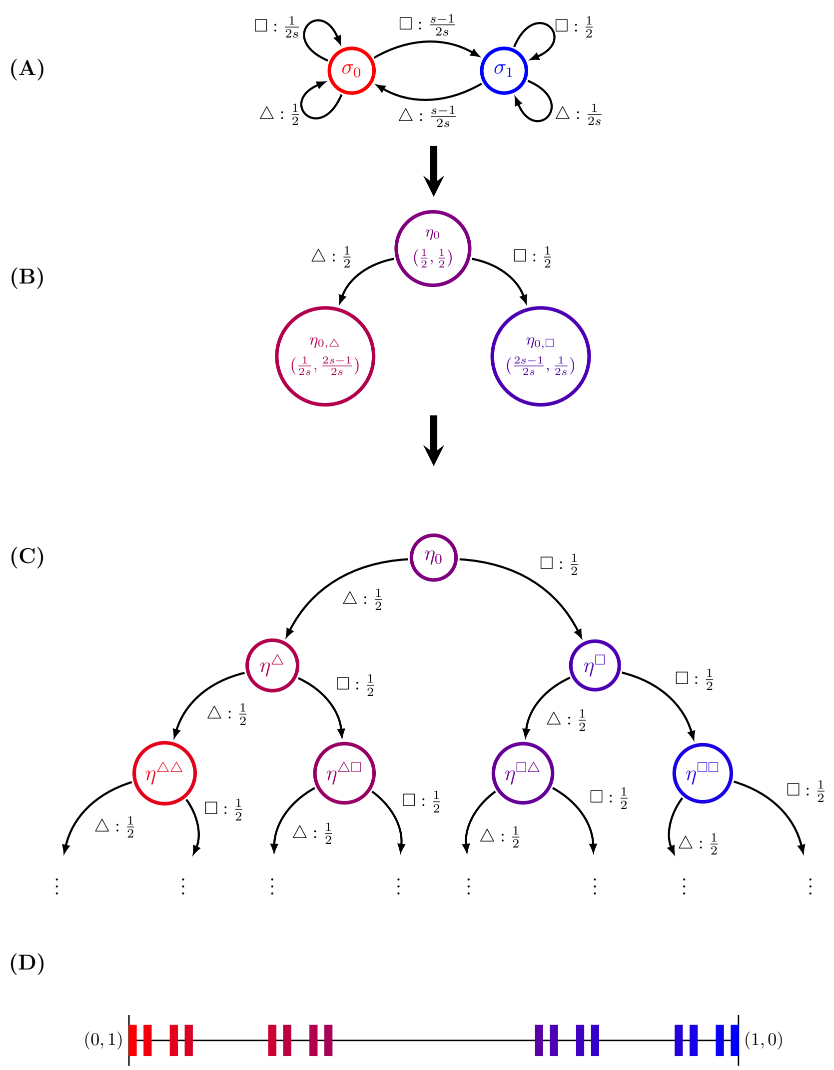


FIG. 2. Constructing the mixed-state presentation of the 2-state s-parametrized nonunifilar HMC shown in (A): The invariant state distribution $\pi = (1/2, 1/2)$. It becomes the first mixed state η_0 used in (B) to calculate the next set of mixed states—those having seen past words of length one. (C) In this example, infinitely many mixed states are generated and $\boldsymbol{\mathcal{R}}$ is the middle-$1/s$ Cantor set. Color indicates the relative closeness of each mixed state to the original states in (A). From η_0, one would need to see a word of infinite $\triangle$s to reach σ_0. In (D) the mixed states for $s = 3$ are pictured on the 1-simplex—the unit interval from $\eta = (0, 1)$ to $\eta = (1, 0)$. In this representation, the relationship to the Cantor set is visually clear.

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

$$\Pr(x|\eta) = \langle \eta | \, T^{(x)} \, |\mathbf{1}\rangle \ , \tag{6}$$

where $T^{(x)}$ is the symbol-labeled transition matrix associated with the symbol x. Now given a mixed state at time t, we may 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:

$$\langle \eta_{t+1,x} | = \frac{\langle \eta_t | \, T^{(x)}}{\langle \eta_t | \, T^{(x)} \, |\mathbf{1}\rangle} \ . \tag{7}$$

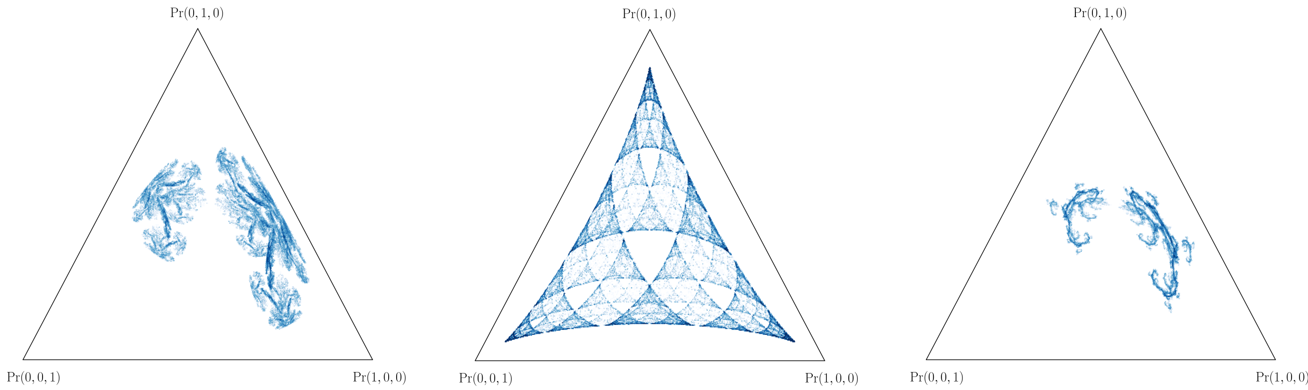
Equation (7) tells us that, by construction, the MSP is unifilar, since each possible output symbol uniquely determines the next (mixed) state. Taken together, Eqs. (6) and (7) define the mixed-state transition dynamic $\mathcal{W}$ as:

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

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

B. Constructing the Mixed-State Presentation

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



(a) Mixed state attractor for 3-state "alpha" machine.

(b) Mixed state attractor for 3-state machine from Eq. (16), with $\alpha = 0.6$ and $x = 0.1$.

(c) Mixed state attractor for 3-state "beta" machine.

FIG. 3. Simple HMCs generate MSPs with a wide variety of structures, many fractal in nature. Each subplot displays 10^4 mixed states of a different, highly-nonunifilar 3-state hidden Markov chain. The HMCs themselves are specified in Appendix A.

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. (6) to calculate $\Pr(x|\eta_t)$ for each $x \in \mathcal{A}$.
5. For $\eta_t \in \boldsymbol{\mathcal{R}}$, use Eq. (7) 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.

This algorithm need not terminate, as shown in Fig. 2, which depicts the MSP construction for the HMC in Fig. 1. However, it can terminate for HMCs described by finite-state ϵ-machines. When dealing with HMCs for which the algorithm does not terminate, one must impose a limit on the number of generated mixed-states, effectively setting a level of approximation for $\boldsymbol{\mathcal{R}}$.

One may ask, given that mixed-state construction returns a unifilar HMC of the underlying process, is the MSP the same as the ϵ-machine? It is not guaranteed to be so, as indeed is the case in Fig. 2. While the so-called Cantor machine generates an uncountably infinite set of mixed states, the ϵ-machine of the underlying process is a single-state fair coin—which has a one-state HMC as its ϵ-machine. We can see this by noting that the symbol-branching probabilities depicted in Fig. 2(C) are identical for every generated mixed state. This may seem like a cause for concern, as we seem to have overestimated the necessary state size for the process in Fig. 2 by an infinite factor. However, this case is both rare and easy to check for, as discussed in Appendix B. By applying a simple check on the uniqueness of mixed states, we can confirm if the MSP is the ϵ-machine of the underlying process. Unless otherwise noted, we take this to be true.

Although the most common purpose of applying mixed-state construction is to *unifilarize* an HMC, we may find the MSP of uHMCs as well. The MSPs of unifilar presentations are interesting and contain additional information beyond the initial unifilar presentation. For example, they typically contain transient causal states and these are employed to calculate many complexity measures that track convergence statistics [35].

Here, we focus on the mixed-state presentations of nonunifilar HMCs, which typically have an infinite mixed-state set $\boldsymbol{\mathcal{R}}$. Even applying mixed-state construction to nominally simple, finite-state, finite-alphabet nonunfilar HMCs results in an explosion of mixed states. Figure 3 gives three examples of MSPs with fractal mixed states sets $\boldsymbol{\mathcal{R}}$, each generated by a three-state nonunifilar HMC.

C. MSP as an IFS

Specifying MSP construction in this way reveals that generating mixed states is a type of random dynamical system known as *place-dependent iterated function system* (IFS) [32]. For finite k and space Δ, place-dependent IFSs are characterized by a set of *mapping functions*:

$$\left\{ f^{(x)} : \Delta \to \Delta \;\middle|\; x \in \{0, 1 \dots k\} \right\} ,$$

and associated *probability functions*:

$$\left\{ p^{(x)} : \Delta \to [0,1] \;\middle|\; x \in \{0, 1 \dots k\} \right\} .$$

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 $\{p^{(x)}(\eta_0) : x = 1, \dots, k\}$ 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 and continue, generating $\eta_0, \eta_1, \eta_2, \ldots$.

An N-state HMC's associated mixed-state construction defines a place-dependent IFS over the $N-1$ simplex, with each symbol-labeled transition matrix $T^{(x)}$ defining a mapping function (Eq. (7)) and associated probability function (Eq. (6)). As our previous results showed [32], an IFS defined by an ergodic HMC has a unique attractor, which is the set of mixed states $\boldsymbol{\mathcal{R}}$. Additionally, this attractor has a unique, attracting, invariant measure known as the *Blackwell measure* $\mu_B(\boldsymbol{\mathcal{R}})$.

IV. STRUCTURE OF INFINITE-STATE PROCESSES

Our prior development showed how to use the MSP to find a process' intrinsic randomness—in the form of Shannon entropy rate h_μ. Our goal here is to complement the measure of randomness with a measure of structure or memory.

Recall that for processes generated by unifilar HMCs, a unique minimal machine known as the ϵ-machine exists, so we can uniquely define the statistical complexity C_μ for a process using Eq. (4). Nonunifilar HMCs have no such canonical minimal presentation. However, as discussed in Section III, we may find the mixed-state presentation $\mathcal{U}(M)$ of uHMCs to unifilarize them. Using the MSP to develop a measure of structure for processes generated by HMCs is a natural choice, since the MSP for a given process is unifilar and unique [32]—indeed, in most cases, the MSP is the ϵ-machine of the underlying process (see Appendix B).

However, the naive approach of simply measuring structure with statistical complexity C_μ introduces a problem: the statistical complexity diverges for an HMC with an uncountably-infinite state set $\boldsymbol{\mathcal{R}}$. In general, MSPs of HMCs are uncountably-infinite state, precluding distinguishing them via C_μ. This being said, it is clear visually from Fig. 3 that HMCs with uncountably-infinite state spaces still have significant and distinct structures. We wish to find a way to measure and distinguish such structure. For this, we take inspiration from Shannon's dimension rate [4] and call on a familiar tool.

Fractal dimension measures the rate at which a chosen size metric of a set diverges with the scale at which the set is observed [36–40]. Fractal dimension is also useful to probe the "size" of objects when cardinality is not informative. For example, the mixed-state presentation, generically, has an uncountable infinity of causal states. That observation is far too coarse, though, to distinguish the clearly distinct mixed-state sets $\boldsymbol{\mathcal{R}}$ in Fig. 3. Each is uncountably infinite, but the $\boldsymbol{\mathcal{R}}$'s geometries differ. Determining their fractal and other dimensions will allow us to distinguish them and allow us to introduce additional insights into the original process' intrinsic information processing.

A. Dimensions

Consider the mixed-state set $\boldsymbol{\mathcal{R}}$ on the simplex for an N-state HMC M that generates a process $\mathcal{P}$. We consider two types of dimension for $\boldsymbol{\mathcal{R}}$: the Minkowski-Bouligand or box-counting dimension, often simply called the fractal dimension, and the information dimension.

To calculate the first, coarse-grain the N-simplex with evenly spaced subsimplex cells of side length ϵ. Let $\mathcal{F}(\epsilon)$ be the set of cells that encompass at least one mixed state. Then $\boldsymbol{\mathcal{R}}$'s *box-counting dimension* is:

$$d_0(\boldsymbol{\mathcal{R}}) = -\lim_{\epsilon \to 0} \frac{\log |\mathcal{F}(\epsilon)|}{\log \epsilon} ~, \quad (8)$$

where $|C|$ is the size of set C.

The information dimension considers how the measure over $\boldsymbol{\mathcal{R}}$ scales. Let each cell in $\mathcal{F}(\epsilon)$ be a state and approximate the dynamic over $\mathcal{U}(M)$ by grouping all transitions to and from states encompassed by the same cell. This results in a Markov chain that generates an approximation of the original process $\mathcal{P}$ that has a stationary distribution $\mu(\mathcal{F}(\epsilon))$. Then $\boldsymbol{\mathcal{R}}$'s *information dimension* is:

$$d_1(\mu(\boldsymbol{\mathcal{R}})) = \lim_{\epsilon \to 0} \frac{H_\mu[\mathcal{F}(\epsilon)]}{\log \epsilon} ~, \quad (9)$$

where $H_\mu[\mathcal{F}(\epsilon)] = -\sum_{C_i \in \mathcal{F}(\epsilon)} \mu(C_i) \log \mu(C_i)$ is the Shannon entropy over the set $\mathcal{F}(\epsilon)$ of cells that cover attractor $\boldsymbol{\mathcal{R}}$ with respect to μ.

B. Dimensions and Scaling of HMCs

These dimensions give two complementary resource-scaling laws for HMC-generated processes. Rearranging Eq. (8), we see that the number of mixed states in our finite-state approximation to $\mathcal{U}(M)$ scales algebraically with $\boldsymbol{\mathcal{R}}$'s box-counting dimension:

$$|\mathcal{F}(\epsilon)| \sim \epsilon^{-d_0(\boldsymbol{\mathcal{R}})} ~. \quad (10)$$

In other words, for an uncountably infinite MSP, the rate of growth of mixed states is $d_0(\boldsymbol{\mathcal{R}})$.

Similarly, the entropy of the mixed-state set scales with the information dimension. Rearranging Eq. (9) shows that the state entropy of the finite-state approximation to $\mathcal{U}(M)$ scales logarithmically with $\boldsymbol{\mathcal{R}}$'s information dimension with respect to the Blackwell measure:

$$H_\mu[\mathcal{F}] \sim d_1(\mu) \cdot \log \epsilon ~. \quad (11)$$

As $\epsilon \to 0$, $|\mathcal{F}|$ and $H_\mu(\mathcal{F})$ diverge and d_0 and d_1 are the divergence rates, respectively. The remainder focuses on d_1 as applied to the ϵ-machine, for which it describes the rate of divergence of statistical complexity C_μ.

V. STATISTICAL COMPLEXITY DIMENSION

We call the information dimension $d_1(\mu)$ of the ϵ-machine the *statistical complexity dimension* d_μ.

Applying d_1 to the Blackwell measure $\mu_B(\boldsymbol{\mathcal{R}})$ gives the rate of divergence of C_μ as one constructs increasingly better finite-state approximations to the infinite-state ϵ-machine. In this way, d_μ describes the divergence of memory resources when attempting to optimally predict a process that requires an uncountably-infinite number of predictive features. This is a unique, minimal description of the process' structural complexity. This solves the challenge posed in the introduction: quantifying structure for truly complex systems.

When a process may be optimally predicted with a countable number of predictive features, the statistical complexity dimension vanishes. In this case, the more relevant complexity measure is the original ϵ-machine statistical complexity C_μ, which is is finite. Statistical complexity dimension for a process that may be minimally generated with N states is less than or equal to $N-1$. This is the associated IFS's *embedding dimension*, since the mixed states lie in a space of dimension $N-1$.

Unfortunately, directly calculating the information dimension using Eq. (9)—and therefore calculating the statistical complexity dimension d_μ—is nontrivial, as it requires estimating a fractal measure. Fortunately, to calculate the d_μ of the mixed state attractor $\boldsymbol{\mathcal{R}}$, we can leverage the associated generating dynamical system (see Section III C).

A. Dimension from Dynamical (In)Stabilities

We can link the information dimension of an MSP's mixed state set $\boldsymbol{\mathcal{R}}$ to the stability properties of the associated IFS. This starts with determining the local time-average stability and instability of orbits within an attractor via the *spectrum of Lyapunov characteristic exponents* $\Gamma = \{\lambda_1, \ldots, \lambda_N : \lambda_i \geq \lambda_{i+1}\}$ [41, 42]. Individual LCEs λ_i measure the average local growth or decay rate of orbit perturbations. The net result is a list of quantities that indicate long-term orbit instability ($\lambda_i > 0$) and orbit stability ($\lambda_i < 0$) in complementary directions.

Usefully, their sum gives the net state-space divergence—volume loss for dissipative systems. The sum of the positive LCEs is the dynamical system's entropy rate [40, 41, 43–45]—the net information generation.

To motivate our present, somewhat indirect, use of the Lyapunov spectrum, it will help to develop a simple intuition for the LCEs as quantities of stretching and contraction. Imagine the attractor of a two-dimensional ($N=2$) map with LCEs $\lambda_1 > 0 > \lambda_2$ and whose state space is covered with equally spaced squares of side length ϵ. After iterating the map q times, for ϵ small enough, the local action of the map is approximately linear. From the LCE definition, this means it takes the initial square cells to rectangles of average length $(e^{\lambda_1 q})\epsilon$ and average width $(e^{\lambda_2 q})\epsilon$. Now, covering the attractor with squares of side length $(e^{\lambda_2 q})\epsilon$ requires roughly $e^{\left(\frac{\lambda_1}{\lambda_2}\right)q}$ squares per rectangle. In this way, the Lyapunov exponents describe scalings analogous to that seen with the box-counting dimension [46].

This suggests defining a *Lyapunov dimension* in terms of the spectrum Γ [47]:

$$d_\Gamma = \begin{cases} k + \dfrac{\sum_i^k \lambda_i}{|\lambda_{k+1}|} \ , & \displaystyle\sum_i^N \lambda_i < 0 \\ N \ , & \displaystyle\sum_i^N \lambda_i \geq 0 \end{cases} , \tag{12}$$

where k is the largest index such that the summation $\sum_i^k \lambda_i$ remains positive. If $\lambda_1 < 0$, $d_\Gamma = 0$. It's name helps distinguish the conditions under which the relationships between the various dimensions actually hold. (There are many conditions and system classes that this summary necessarily leaves out.)

It was conjectured [47] that for "typical dynamical systems", the Lyapunov dimension d_Γ equals the information dimension d_1. What has been shown, however, is that for any ergodic, invariant probability measure μ:

$$d_1(\mu) \leq d_\Gamma \ ,$$

with equality when μ is a Sinai-Bowen-Ruelle (SBR) measure [48]. This remarkable relationship directly relates a system's dynamics to the geometry and natural measure of its attractor. Additionally, and usefully, Eq. (12) gives us a tractable method to find the information dimension of an attractor generated by a dynamical system.

B. Calculating Statistical Complexity Dimension

We now have in hand two important pieces. First, the definition of statistical complexity dimension d_μ as the information dimension of an ϵ-machine. Second, we have a bound on the information dimension of an attractor, given knowledge of the generating system's dynamics. To complete our picture, we now address the final puzzle piece.

As Section V A discussed, there is a direct relationship between the information dimension of a chaotic attractor and the dynamics of the system to which the attractor belongs. Furthermore, as discussed in Section III C, every HMC has an associated random dynamical system—the iterated function system (IFS)—which has the HMC's set of mixed states $\boldsymbol{\mathcal{R}}$ as its unique attractor. Combining these two facts allows us to exactly calculate d_μ in many cases of interest.

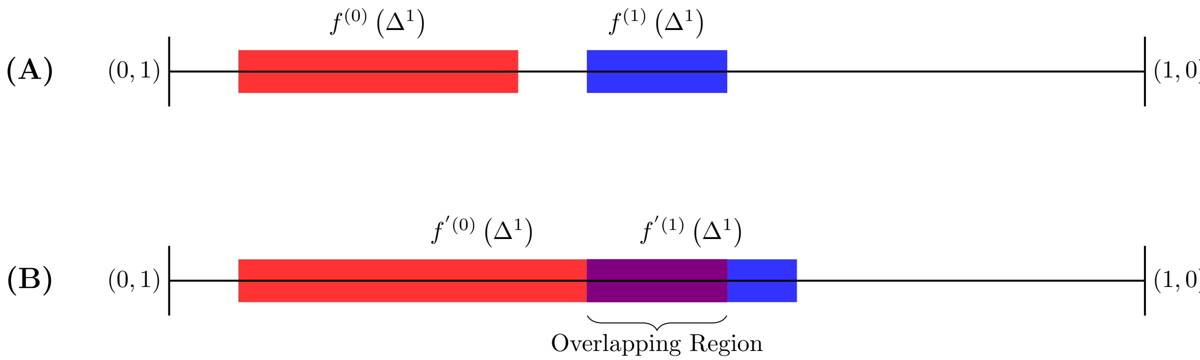


FIG. 4. Overlap problem on the 1-simplex Δ^1: Two distinct IFSs are considered, each with two mapping functions. The images of the mapping functions over the entire simplex are depicted in red and blue. (A) Images of the mapping functions $f^{(0)}$ and $f^{(1)}$ do not overlap—every mixed state $\eta_t \in \boldsymbol{\mathcal{R}}$ has a unique pre-image. (B) Images of the mapping functions overlap (purple Overlapping Region)—there exist $\eta_1, \eta_2 \in \boldsymbol{\mathcal{R}}$ such that $f^{'(0)}(\eta_1) = f^{'(1)}(\eta_2) = \eta_3$. This case is an *overlapping IFS*.

An advantage of working with IFSs defined by HMCs is a clean division exhibited by their Lyapunov spectrum. It has been shown that the entropy rate of the generated process is equivalent to the largest Lyapunov exponent [49, 50]. Calculating this was the main topic of the prequel to the present work [32]. Furthermore, due to the contractivity of the IFS mapping functions, all other Lyapunov exponents will necessarily be negative. For a review of calculating the Lyapunov exponents for IFSs, see Appendix E.

Therefore, the Lyapunov dimension of an IFS is:

$$\widetilde{d_\Gamma} = \begin{cases} k - 1 + \dfrac{h_\mu + \sum_i^{k-1} \lambda_i}{|\lambda_k|}, & h_\mu + \displaystyle\sum_i^N \lambda_i < 0 \\ N \ , & h_\mu + \displaystyle\sum_i^N \lambda_i \geq 0 \end{cases} , \quad (13)$$

where k is now the largest index for which $h_\mu + \sum_i^k \lambda_i > 0$. This is simply Eq. (12), as if h_μ was the largest Lyapunov exponent.

Under specific technical conditions to be discussed shortly, the IFS d_Γ is exactly equal to the information dimension of the IFS's attractor and, therefore, is d_μ [51]. In general, relaxing those conditions, $\widetilde{d_\Gamma}$ upper bounds the statistical complexity dimension:

$$\widetilde{d_\Gamma} \geq d_\mu \ . \quad (14)$$

Assembling these pieces together determines the basic algorithm to calculate (or bound) the statistical complexity dimension:

1. For an N-state HMC M with $|\mathcal{A}| = k$, write down the associated IFS with k symbol-labeled mapping functions and probability functions.
2. Calculate the entropy rate h_μ using the Blackwell limit (see [32]).
3. Calculate the negative Lyapunov exponents $\{\lambda_1, \ldots, \lambda_{N-1}\}$ (see Appendix E).
4. Compute the Lyapunov dimension d_Γ using Eq. (13).

As mentioned, in specific cases, the Lyapunov dimension is exactly equal to the statistical complexity dimension, and our task is complete. However, there are major technical concerns with when we have only the bound in Eq. (14) and with its tightness then.

C. The Overlap Problem

A subtle disadvantage of working with IFSs is a direct result of the stochastic nature of them as random dynamical systems. We must consider the *overlap problem*, which concerns the ranges of the symbol-labeled mapping functions $f^{(i)}$, illustrated in Fig. 4. Specifically, the problem means that we must distinguish between IFSs that meet the open set condition and those that do not.

Definition 6. *An iterated function system with mapping functions* $f^{(\eta)} : \Delta^N \to \Delta^N$ *satisfies the* open set condition *(OSC) if there exists an open set* $U \in \Delta^N$ *such that for all* $\eta, \zeta \in \Delta^N$*:*

$$f^\eta(U) \cap f^\zeta(U) = \emptyset, \quad \eta \neq \zeta \ .$$

IFSs that meet the OSC are nonoverlapping IFSs*.*

When the images of the symbol-labeled mappings overlap the inequality in Eq. (14) is strict. To briefly outline the consequences, for an overlapping IFS the entropy rate h_μ does not accurately capture state-space expansion. And, this causes the IFS d_Γ (Eq. (13)) to overestimate the information dimension. As a rule of thumb, the degree to which the mappings overlap determines the magnitude of the bound's error. The impact of overlaps is significant. It is explored both in Section VI, where we calculate the statistical complexity dimension for HMCs with and without overlap, as well as in the sequel, which diagnoses the problem's origins and outlines a solution.

For now, to give a workable approach, we simply introduce two extra steps to the d_μ algorithm from the previous section:

5. Determine if the Open Set Condition is met using the mapping functions $f^{(i)}$.
6. If the OSC is not met, estimate the degree of overlap to determine the closeness of the bound on d_μ.

D. Statistical Complexity Dimension for Processes Generated by Two-State HMCs

Finally, we analyze two-state HMCs, for which Eq. (13) simplifies significantly and gives exact results for d_μ. (Thus, the concerns just outlined occur only for three or more state HMCs, which are explored in the next section.)

For two-state nonunifilar HMCs, the mixed-state set lives on the 1-simplex Δ^1—the unit interval from $\eta = (0,1)$ to $\eta = (1,0)$. Mixed states $\eta \in \boldsymbol{\mathcal{R}}$ and the dynamic on them exist in a one-dimensional space and, thus, there is a single negative Lyapunov exponent $\lambda_1 < 0$.

In this case, the calculation of the negative Lyapunov exponent is particularly direct, since the maps are all one-dimensional. The negative Lyapunov exponent for a one-dimensional map $\eta_{n+1} = f(\eta_n)$ is:

$$\lambda(\eta_0) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N-1} \log \left| \frac{df(\eta_i)}{d\eta} \right|$$

for an orbit starting at η_0. For an IFS with a set of of mapping functions $\{f^{(x)}\}$, we find λ as the weighted average of the Lyapunov exponents of each map:

$$\lambda_\mu = \int \sum_x p^{(x)}(\eta) \log \left| \frac{df^{(x)}(\eta)}{d\eta} \right| d\mu ,$$

where μ is the IFS's Blackwell measure. We can apply ergodicity to transform this into a summation over time for ease of calculation.

If a two-state HMC has an MSP with an uncountable infinity of mixed states and its corresponding IFS satisfies OSC, there is a simple relationship between the entropy rate, the Lyapunov exponent, and the statistical complexity dimension. This is given by:

$$d_\mu(\mu) = -\frac{h_\mu}{\lambda_\mu} , \tag{15}$$

recalling that $h_\mu > 0$, so that the dimension is always positive. Failing OSC, this ratio is an upper bound on the dimension of the measure μ [52]. For a discussion of the intuition behind this formula, see *Appendix C*.

VI. MULTI-STATE HMC EXAMPLES

Notably, the Lyapunov dimension Eq. (15) for more-than-two-state HMCs is easily shown to be correct when the maps are similitudes and the probability functions are constant. The latter is seen, for example, with the Sierpinski triangle, as discussed in Appendix D.

However, we are generally interested in multi-state HMCs that do not produce the perfectly-self-similar fractals that arise under those conditions. Furthermore, we are often interested in considering physical systems described by parametrized HMCs, such as those that arose in the two prequels on quantum measurement processes and information engine functionality [18, 33]. In such cases, an HMC determined by an application may meet the OSC in some regions of parameter space and fail to do so in others. We will consider an HMC that spans the breadth of these possible behaviors, from zero overlap to complete overlap. This will demonstrate the range of applicability of the statistical complexity dimension algorithm laid about above.

Consider the following HMC with 3 symbols $\{\square, \triangle, \circ\}$ and 3 states:

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

$$T^{\circ} = \begin{pmatrix} \beta y & \beta x & \alpha x \\ \beta x & \beta y & \alpha x \\ \beta x & \beta x & \alpha y \end{pmatrix} , \tag{16}$$

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

Figure 5 shows how the MSP attractors change across the (α, x) parameter space. Each black dot is a generated mixed state, while the colored regions show the range of each symbol-labeled map.

For example, on one hand, in the top left corner with $\alpha = 0.01$ and $x = 0.01$, we find an attractor that extends across the simplex, with moderate amounts of overlap. On the other, $\alpha = 0.79$ and $x = 0.11$ produces an attractor with no overlap, and clearly defined regions.

Moreover, for any α, choosing $x = 1/3$ leads the MSP attractor to collapse to a finite 3-state HMC, since the symbol-labeled mapping functions become constant functions. In this case, there is no overlap, as each symbol-labeled map takes on a different constant value.

However, when $\alpha = \beta = 1/3$, all symbol-labeled mapping functions are identical. Therefore, the attractor is the single fixed-point shared by all three maps—a single-state HMC. This is a case of maximal possible overlap. Along both lines in parameter space the MSP collapses to a finite state HMC, so $d_\mu = 0$, by definition. However, these different mechanisms of state-collapse are relevant in calculation of d_Γ via Eq. (13).

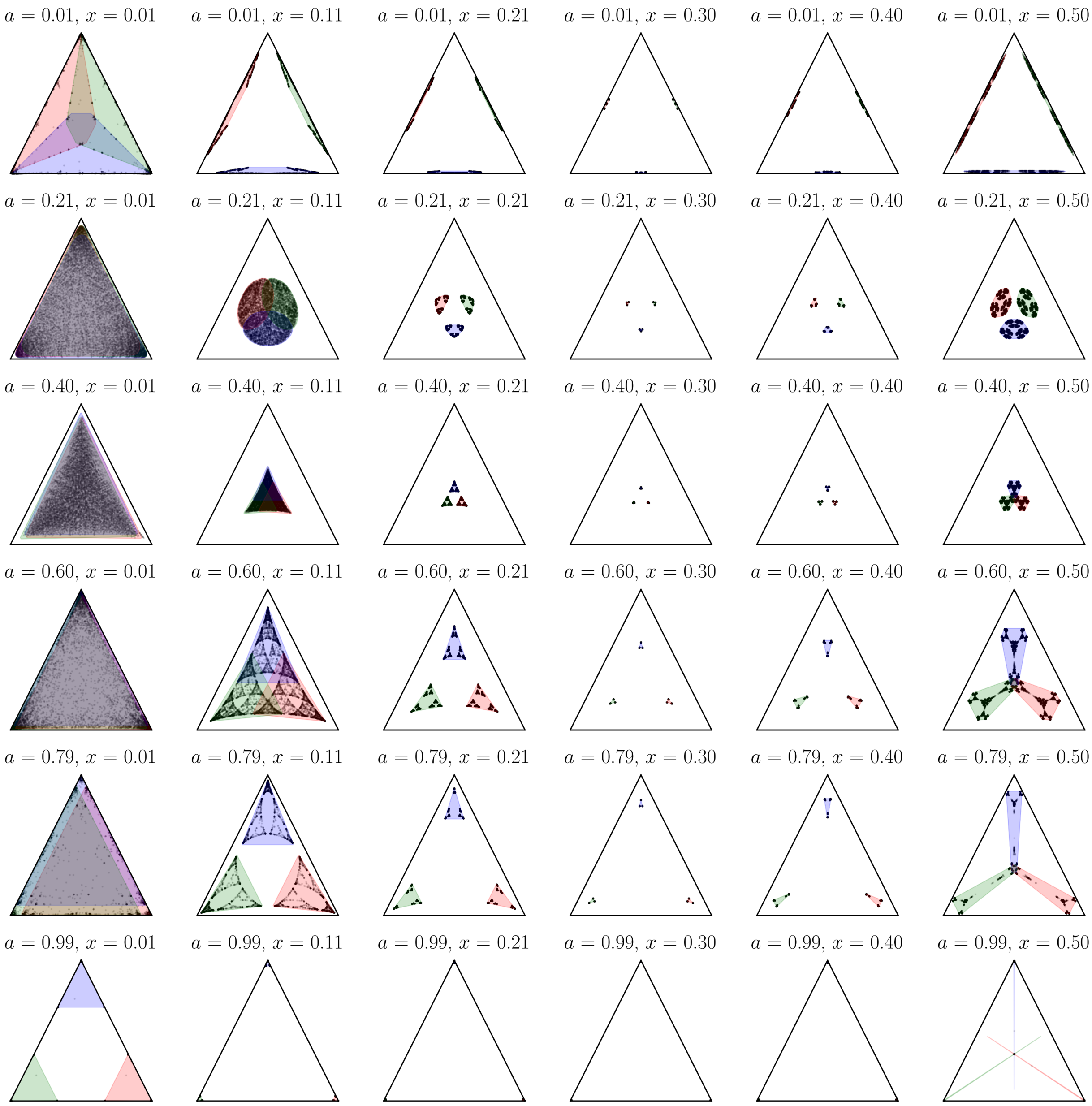
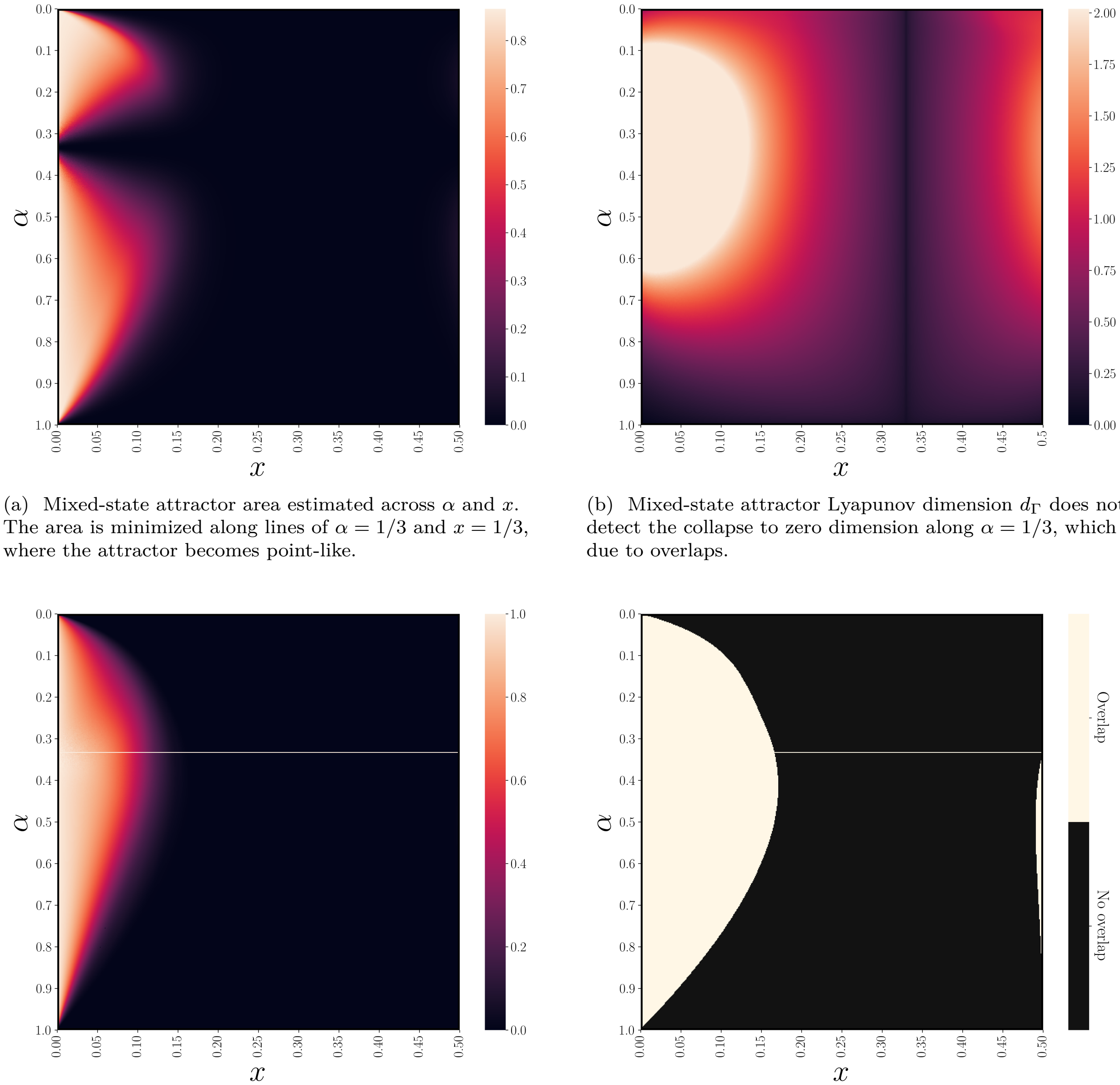


FIG. 5. Mixed-state attractors generated by a 3-state HMC parametrized over $\alpha \in [0, 1]$ and $x \in [0, 0.5]$. The HMC itself is given in Eq. (16). $100,000$ mixed states are plotted for each attractor, with the initial $5,000$ states thrown away as transients. The ranges of the symbol-labeled maps are color shaded, revealing regions of their image overlap on the attractor. Comparing to Eq. (16), the red, blue, and green regions represent the images of the mapping functions defined by $T^{\square}$, $T^{\triangle}$, and T°, respectively.

First, consider Fig. 6a, which illustrates the estimated area on the simplex taken up by the attractor across parameter space. For a discussion of how attractor area was estimated, please see Appendix F. This figure matches the (x, α) grid in Fig. 5: lower values of x produce larger attractors, excepting the region near $\alpha = 1/3$, where the area drops to zero. We know from our analysis of the attractor grid that along this line, the attractor is finite state, and so the statistical complexity dimension d_μ vanishes. However, this is not accurately reflected by d_Γ, as seen in Fig. 6b.

That said, d_Γ clearly—and correctly—vanishes when



(a) Mixed-state attractor area estimated across α and x. The area is minimized along lines of $\alpha = 1/3$ and $x = 1/3$, where the attractor becomes point-like.

(b) Mixed-state attractor Lyapunov dimension d_Γ does not detect the collapse to zero dimension along $\alpha = 1/3$, which is due to overlaps.

(c) Percentage of attractor area in which there is overlap.

(d) Overlap in the attractor area. Comparing with (c), for much of this area, the overlap is very small.

FIG. 6. Attractor area, overlap regions, and Lyapunov dimension d_Γ of the mixed-state attractors shown in Fig. 5, parametrized by $\alpha = [0, 1]$ and $x = [0, 0.5]$. The HMC itself is given in Section VI. See Appendix E and Appendix F for a detailed discussion of the production of these plots.

$x = 1/3$. This is the other line in parameter space where the MSP is finite state and d_μ is known to be zero. This disparity is due to the different mechanisms driving the collapse.

When $x = 1/3$, the symbol-labeled mapping functions become constants, which is reflected in the Lyapunov exponents and consequently in d_Γ. The constant functions have negative Lyapunov exponents of negative infinity, sending Eq. (13) to zero.

In contrast, along the $\alpha = 1/3$ line, the contraction in the state space is a result of the overlapping maps and the symbol-labeled maps are not infinitely contracting, so d_Γ badly overestimates d_μ. This illustrates the importance of the OSC on the bound.

This poses the question, which regions in HMC parameter space exhibit overlap IFS maps? Figure 6d depicts the parameter space as overlap or no overlap. Figure 6c shows this as a percentage of the total attractor area. For a discussion of how overlap was determined, please see Appendix F. Comparing the two, we see that there is a significant region for which overlap does exist for $x < 0.15$ and a smaller region where $x > 0.48$. However, for much of that region the attractor's overlap area is relatively small. As a rule of thumb, the gap between d_Γ and d_μ for the mixed-state set measure $\mu(\boldsymbol{\mathcal{R}})$ is determined by the percentage of the attractor that is affected by overlap. If the overlap region is relatively small, in comparison to the size of the attractor, d_Γ may be very close to d_μ.

However, if the overlap is very large, d_Γ may be a dramatic overestimation of d_μ. This occurs when $\alpha = 1/3$ and $x < 0.15$. The statistical complexity dimension d_μ vanishes, yet the Lyapunov dimension saturates at $d_\Gamma = 2.0$.

We also note that the mechanism driving the collapse of the MSP attractor at $\alpha = 1/3$ is a discontinuity in the parameter space, as compared to $x = 1/3$. This is because state space collapse due to overlap requires the maps to be identical, and even minute differences in the symbol-labeled transition matrices will produce an uncountably-infinite MSP, potentially with $d_\mu = 2.0$. This encourages us to consider not just the statistical complexity dimension, but also the area of the attractor, and the nearby regions in parameter space for a clearer understanding of the underlying HMC. In this, the tools developed here, by greatly facilitating surveys of large regions of parameter space, are particularly useful.

Thus, while there are wide parameter regions in which the quantitative measures developed here are correct and efficiently estimated, this is not the entire story. As the analysis here diagnosed its effects, overlap must be addressed for full generality. The sequel shows how to correct for overestimating statistical complexity dimension, allowing accurate calculation across the entire parameter space [53]. However, the exploratory observations outlined here provide crucial guides to the underlying mechanisms and so to further developments.

VII. CONCLUSION

Our development opened by considering the challenge of quantifying the structure of complex systems. For well over a half a century Shannon entropy rate stood as the standard by which to quantify randomness in time series and in chaotic dynamical systems. Quantifying observable patterns remained a more elusive goal. However, with developments from computational mechanics, it has become possible to answer questions of structure and pattern, at least for stochastic processes generated by finite-state predictive machines, including the symbolic dynamics generated by chaotic dynamical systems.

To handle the processes generated by finite-state nonpredictive (nonunifilar) hidden Markov chains, we developed the mixed-state presentation. This *unifilarized* general HMCs, giving a predictive presentation that itself generates the process. However, adopting a unifilar presentation came at a heavy cost: Generically, they are infinite state and so previous structural measures diverge. Nonetheless, we showed how to work constructively with these infinite 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 to more fully describe the original stochastic process.

Previously, others considered the IFS-HMC connection [54, 55]. Complementing those efforts, we expanded the role of the mixed-state presentation to calculate entropy rate and demonstrated its usefulness in determining the underlying structural properties of the generated process. Indeed, Figs. 3 and 5 show how visually striking—and distinct—mixed state sets generated by HMCs are.

Here, moving in a new direction beyond previous efforts, we established that the information dimension of the mixed-state attractor is exactly the divergence rate of the *statistical complexity* [56]—a measure of a process' structural complexity that tracks memory. Thus, processes in this class effectively increase their use of memory, "creating" mixed or causal states, on the fly. Furthermore, we introduced a method to calculate the information dimension of the mixed-state attractor from the Lyapunov spectrum of the mixed-state IFS. In this way, we demonstrated that coarse-graining the mixed-state simplex—the previous method for studying the structure of infinite-state processes [57]—can be avoided altogether. This greatly improves accuracy and calculational speed.

During the development, we noted several obstacles. Most importantly, the presence of *overlap* and failure to meet the conditions of the OSC causes the Lyapunov dimension to be a strict upper bound, and some times quite a poor one, on the statistical complexity dimension. The final work of our trilogy [53] introduces a measure of how badly the entropy rate overestimates the expansion of the mixed-state set. Combining this measure with the Lyapunov-information dimension conjecture finally yields a correct Eq. (13) to apply to HMCs with overlap; that is, for processes generated by all general HMCs.

To close, we note that the structural tools introduced here and the entropy-rate method introduced previously [32] have been put to practical use in two previous works. One diagnosed the origin of randomness and structural complexity in quantum measurement [33]. The other exactly determined the thermodynamic functioning of Maxwellian information engines [18], when there had been no previous method for this kind of detailed and accurate analysis. At this point, however, we 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 [53].

ACKNOWLEDGMENTS

The authors thank Sam Loomis, Sarah Marzen, Ariadna Venegas-Li and Ryan James 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 grants FQXi-RFP-IPW-1902 and FQXI-RFP-CPW-2007 and U.S. Army Research Laboratory and the U.S. Army Research Office grants W911NF-18-1-0028 and W911NF-21-1-0048.

---

[1] A. M. Turing. On computable numbers, with an application to the entsheidungsproblem. *Proc. Lond. Math. Soc. Ser. 2*, 42:230, 1936. 1

[2] 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.

[3] M. Minsky. *Computation: Finite and Infinite Machines.* Prentice-Hall, Englewood Cliffs, New Jersey, 1967. 1

[4] C. E. Shannon. A mathematical theory of communication. *Bell Sys. Tech. J.*, 27:379–423, 623–656, 1948. 1, 2, 3, 7

[5] A. N. Kolmogorov. *Foundations of the Theory of Probability.* Chelsea Publishing Company, New York, second edition, 1956. 1

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

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

[8] 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.

[9] Ja. G. Sinai. On the notion of entropy of a dynamical system. *Dokl. Akad. Nauk. SSSR*, 124:768, 1959. 1

[10] J. P. Crutchfield. Between order and chaos. *Nature Physics*, 8(January):17–24, 2012. 1, 3

[11] J. P. Crutchfield and D. P. Feldman. Regularities unseen, randomness observed: Levels of entropy convergence. *CHAOS*, 13(1):25–54, 2003. 1, 3

[12] S. Marzen, M. R. DeWeese, and J. P. Crutchfield. Time resolution dependence of information measures for spiking neurons: Scaling and universality. *Front. Comput. Neurosci.*, 9:109, 2015.

[13] C. R. Shalizi and J. P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *J. Stat. Phys.*, 104:817–879, 2001. 1

[14] 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. 1, 3, 4

[15] J. P. Crutchfield and K. Young. Computation at the onset of chaos. In W. Zurek, editor, *Entropy, Complexity, and the Physics of Information*, volume VIII of *SFI Studies in the Sciences of Complexity*, pages 223 – 269, Reading, Massachusetts, 1990. Addison-Wesley.

[16] N. Travers and J. P. Crutchfield. Infinite excess entropy processes with countable-state generators. *Entropy*, 16:1396–1413, 2014.

[17] L. Debowski. On hidden Markov processes with infinite excess entropy. *J. Theo. Prob.*, 27(2):539–551, 2012.

[18] 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. 1, 2, 10, 13

[19] D. L. Turcotte. *Fractals and Chaos in Geology and Geophysics.* Cambridge University Press, Cambridge, United Kingdom, second edition, 1997. 1

[20] J. M. Beggs and D. Plenz. Neuronal avalanches in neocortixal circuits. *J. Neurosci.*, 23(35):11167–11177, 2003. 1

[21] L. Debowski. Excess entropy in natural language: Present state and perspectives. *Chaos*, 21(3):037105, 2011. 1

[22] I. Dobson, B. A. Carreras, V. E. Lynch, and D. E. Newman. Complex systems analysis of series of blackouts: Cascading failure, critical points, and self-organization. *CHAOS*, 17(2):26103, 2007. 1

[23] J. P. Crutchfield. The calculi of emergence: Computation, dynamics, and induction. *Physica D*, 75:11–54, 1994. 1

[24] B. Marcus, K. Petersen, and T. Weissman, editors. *Entropy of Hidden Markov Process and Connections to Dynamical Systems*, volume 385 of *Lecture Notes Series*. London Mathematical Society, 2011. 1

[25] Y. Ephraim and N. Merhav. Hidden Markov processes. *IEEE Trans. Info. Th.*, 48(6):1518–1569, 2002. 1, 2

[26] J. Bechhoefer. Hidden Markov models for stochastic thermodynamics. *New. J. Phys.*, 17:075003, 2015. 1

[27] L. R. Rabiner and B. H. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, January:4–16, 1986. 1

[28] E. Birney. Hidden Markov models in biological sequence analysis. *IBM J. Res. Dev.,*, 45(3.4):449–454, 2001. 1

[29] S. Eddy. What is a hidden Markov model? *Nature Biotech.*, 22:1315–1316, Oct 2004. 1

[30] 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. 1

[31] 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. 1

[32] A. Jurgens and J. P. Crutchfield. Shannon entropy rate of hidden Markov processes. *arXiv.org:2008.12886*, 2020. 1, 2, 3, 4, 6, 7, 9, 13, 5

[33] A. Venegas-Li, A. Jurgens, and J. P. Crutchfield. Measurement-induced randomness and structure in controlled qubit processes. *Phys. Rev. E*, 102(4):040102(R), 2020. 2, 10, 13

[34] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, second edition, 2006. 2

[35] 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. 6

[36] A. Renyi. On the dimension and entropy of probability distributions. *Acta Math. Hung.*, 10:193, 1959. 7

[37] B. B. Mandelbrot. *The Fractal Geometry of Nature*. W. H. Freeman and Company, San Francisco, California, 1982.

[38] G. A. Edgar. *Measure, Topology, and Fractal Geometry*. Springer-Verlag, New York, 1990.

[39] K. Falconer. *Fractal geometry: mathematical foundations and applications*. John Wiley, Chichester, 1990.

[40] Ya. B. Pesin. *Dimension Theory in Dynamical Systems: Contemporary Views and Applications*. Chicago Lectures in Mathematics. University of Chicago Press, Chicago, Illinois, 1997. 7, 8

[41] I. Shimada and T. Nagashima. A numerical approach to ergodic problem of dissipative dynamical systems. *Prog. Theo. Phys.*, 61:1605, 1979. 8

[42] G. Benettin, L. Galgani, A. Giorgilli, and J.-M. Strelcyn. Lyapunov characteristic exponents for smooth dynamical systems and for hamiltonian systems; a method for computing all of them. *Meccanica*, 15:9, 1980. 8

[43] Ja. B. Pesin. Ljapunov characteristic exponents and ergodic properties of smooth dynamical systems with an invariant measure. *Doklady Akademii Nauk*, 226:774–777, 1976. 8

[44] D. Ruelle. Sensitive dependence on initial conditions and turbulent behavior of dynamical systems. *Annals of the New York Academy of Sciences*, 316(1):408–416, 1977.

[45] G. Benettin, L. Galgani, and J.-M. Strelcyn. Kolmogorov entropy and numerical experiments. *Phys. Rev. A*, 14(6):2338–2345, 1976. 8

[46] J. D. Farmer, E. Ott, and J. A. Yorke. The dimension of chaotic attractors. *Physica*, 7D:153, 1983. 8

[47] J. Kaplan and J. Yorke. Chaotic behavior of multidimensional difference equations. In *Functional Differential Equations and Approximation of Fixed Points*, volume 730 of *Lecture Notes in Mathematics*, pages 204–227. Springer, 1979. 8

[48] F. Ledrappier and L. S. Young. The metric entropy of diffeomorphisms: Part II: Relations between entropy, exponents and dimension. *Ann. Mathematics*, 122:540–574, 1985. 8

[49] P. Jacquet, G. Seroussi, and W. Szpankowski. On the entropy of a hidden Markov process. *Theo. Comp. Sci.*, 395:203–219, 2008. 9

[50] T. Holliday, A. Goldsmith, and P. Glynn. Capacity of finite state channels based on Lyapunov exponents of random matrices. *IEEE Trans. Info. Th.*, 52:3509–3532, 2006. 9

[51] B. Barany. On the Ledrappier-Young formula for self-affine measures. *Math. Proc. Cambridge Phil. Soc.*, 159(3):405–432, 2015. 9

[52] J. Jaroszewska and M. Rams. On the Hausdorff dimension of invariant measures of weakly contracting on average measurable IFS. *J. Stat. Physics*, 2008. 10

[53] A. Jurgens and J. P. Crutchfield. Backwards entropy rate of hidden Markov processes. *in preparation*, 2021. 13

[54] M. Rezaeian. Hidden Markov process: A new representation, entropy rate and estimation entropy. *arXiv:0606114*, 2006. 13

[55] W. Słomczyński, J. Kwapień, and K. Życzkowski. Entropy computing via integration over fractal measures. *Chaos: Interdisc. J. Nonlin. Sci.*, 10(1):180–188, Mar 2000. 13

[56] J. P. Crutchfield and K. Young. Inferring statistical complexity. *Phys. Rev. Let.*, 63:105–108, 1989. 13

[57] S. E. Marzen and J. P. Crutchfield. Nearly maximally predictive features and their dimensions. *Phys. Rev. E*, 95(5):051301(R), 2017. 13

[58] A. Jurgens and J. P. Crutchfield. Minimal embedding dimension of minimally infinite hidden Markov processes. *in preparation*, 2021. 2

[59] V. I. Oseledets. A multiplicative ergodic theorem. Lyapunov characteristic numbers for dynamical systems. *Trans. Moscow Math. Soc.*, 19:197–231, 1968. 4

[60] J. H. Elton. An ergodic theorem for iterated maps. *Ergod. Th. Dynam. Sys.*, 7:481–488, 1987. 5

[61] G. Froyland and K. Aihara. Rigorous numerical estimation of Lyapunov exponents and invariant measures of iterated function systems and random matrix products. *Intl. J. Bifn. Chaos*, 10(1):103–122, 2000. 5

[62] A. Boyarsky and Y. S. Lou. A matrix method for approximating fractal measures. *Intl. J. Bifn. Chaos*, 02:167–175, 1992. 5

[63] Shapely: Geometric objects, predicates, and operations. version 1.7.1. 5, 6

Supplementary Materials

Divergent Predictive States: The Statistical Complexity Dimension of Stationary, Ergodic Hidden Markov Processes

Alexandra Jurgens and James P. Crutchfield

arXiv:2102.XXXXX

The Supplementary Materials give the HMCs for the example processes considered, review MSP minimality, draw a correspondence with the Baker's Map on the unit square, layout an HMC whose mixed-state set $\boldsymbol{\mathcal{R}}$ is the well-known Sierpinski Triangle fractal, and review Lyapunov characteristic exponents and calculating their spectrum for IFSs.

Appendix A: Nonunifilar HMC Examples

We reproduce here the HMCs used to create Fig. 3. First, the "alpha" HMC, from Fig. 3a is given by:

$$T^{\square} = \begin{pmatrix} 2.734\times10^{-2} & 0.392 & 1.924\times10^{-2} \\ 0.475 & 2.176\times10^{-2} & 2.766\times10^{-4} \\ 0.224 & 2.711\times10^{-3} & 0.236 \end{pmatrix}, \tag{S1}$$

$$T^{\triangle} = \begin{pmatrix} 1.845\times10^{-3} & 0.133 & 0.259 \\ 3.913\times10^{-2} & 0.315 & 2.789\times10^{-2} \\ 0.467 & 1.015\times10^{-2} & 4.699\times10^{-3} \end{pmatrix}, \text{ and}$$

$$T^{\circ} = \begin{pmatrix} 9.782\times10^{-2} & 3.374\times10^{-2} & 3.644\times10^{-2} \\ 5.422\times10^{-2} & 6.503\times10^{-2} & 2.090\times10^{-3} \\ 5.328\times10^{-2} & 1.278\times10^{-3} & 8.778\times10^{-4} \end{pmatrix}.$$

Figure 3b is given by Eq. (16), at $\alpha = 0.6$ and $x = 0.1$. The "beta" HMC, in Fig. 3c, is given by:

$$T^{\square} = \begin{pmatrix} 5.001\times10^{-2} & 0.388 & 4.251\times10^{-2} \\ 0.464 & 4.484\times10^{-2} & 2.495\times10^{-2} \\ 0.232 & 2.720\times10^{-2} & 0.243 \end{pmatrix}, \tag{S2}$$

$$T^{\triangle} = \begin{pmatrix} 1.708\times10^{-3} & 0.123 & 0.240 \\ 3.623\times10^{-2} & 0.292 & 2.583\times10^{-2} \\ 0.432 & 9.397\times10^{-3} & 4.351\times10^{-3} \end{pmatrix}, \text{ and}$$

$$T^{\circ} = \begin{pmatrix} 9.0576\times10^{-2} & 3.124\times10^{-2} & 3.374\times10^{-2} \\ 5.020\times10^{-2} & 6.021\times10^{-2} & 1.935\times10^{-3} \\ 4.933\times10^{-2} & 1.183\times10^{-3} & 8.127\times10^{-4} \end{pmatrix}.$$

Due to finite numerical accuracy, reproduction of the attractors using these specifications may differ slightly from Fig. 3.

Appendix B: Mixed-State Presentation Minimality

Given an HMC M, minimality of infinite-state mixed-state presentations $\mathcal{U}(M)$ is an open problem. MSPs 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, as seen with the so-called Cantor state process in Figs. 1 and 2. Note that while this HMC generates an uncountable number of mixed states, each one has the same emitted-symbol probability distribution, indicating that all states can be merged into a single state with no loss of predictability. Indeed, the ϵ-machine for the HMC depicted in Fig. 1 is simply the single-state Fair Coin HMC.

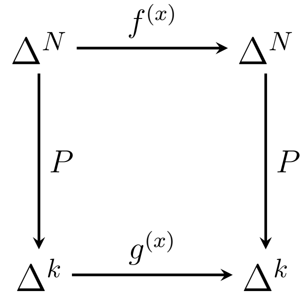


FIG. S1. Commuting diagram for probability functions $P = \{p^{(x)}\}$, mixed-state mapping functions $f^{(x)}$, and proposed symbol-distribution mapping functions $g^{(x)}$.

A proposed solution to this needless presentation verbosity is a short and simple check on *mergeablility* of mixed states. This 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 $L \in \mathbb{N}^+$.

A 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 check this by considering, for an $N+1$ state HMC 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. (S3) 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 $f^{(x)}$. When every mixed state has a unique outgoing probability distribution, these states are also the causal states, and the MSP is the process' ϵ-machine. Our companion work [58] elaborates on this and the implications for identifying the embedding dimension of minimal generators.

Appendix C: Correspondence with Baker's Map

The simple dimension formula in Eq. (15) may not seem easily motivated. Especially, considering that, in general, both positive and negative Lyapunov exponents are required to have a nontrivial attractor. However, for iterated function systems, all Lyapunov exponents are negative and the expansive role played by positive Lyapunov exponents is instead played by an IFS's stochastic map selection, as measured by the entropy rate h_μ.

This is more intuitively appreciated by comparing the two-state IFS with the Baker's map. Consider the Baker's map:

$$x_{n+1} = \begin{cases} \dfrac{x_n}{s_0} \ , & y < p \\ \dfrac{x_n + s_1 - 1}{s_1} \ , & y \geq p \end{cases} \quad \text{and}$$

$$y_{n+1} = \begin{cases} \dfrac{y_n}{p} \ , & y < p \\ \dfrac{y_n - p}{p - 1} \ , & y \geq p \end{cases}$$

It has LCE spectrum $\Lambda = \{\lambda_1, \lambda_2\}$, where:

$$\lambda_1 = p \log(p) + (1-p)\log(1-p)$$
$$\lambda_2 = p \log(1/s_0) + (1-p)\log(1/s_1) \ .$$

Note that $\lambda_1 > 0$ and $\lambda_2 < 0$. Then, the Lyapunov dimension is:

$$d_\Gamma = 1 - \frac{\lambda_1}{\lambda_2} \ .$$

To compare this to an IFS, take:

$$\{f(x)\} = \left\{ \frac{x_n}{s_0}, \frac{x_n + s_1 - 1}{s_1} \right\} \text{ and}$$
$$\{p(x)\} = \{p, 1-p\} \ .$$

Thus, we identify the y coordinate as controlling the stochastic map choice. The dynamic over position in the y direction exactly determines the IFS entropy rate. Since the Baker's map is volume preserving in y, the extra dimension always contributes a plus one in the dimension formula. In other words, the dimension along a slice of constant y equals the IFS dimension.

Appendix D: Sierpinski's Triangle

The Sierpinski triangle is a canonical a Cantor set in two dimensions. An HMC that generates a MSP attractor that is the Sierpinski triangle is:

$$T^{(0)} = \begin{pmatrix} a & 0 & a(s-1) \\ 0 & a & a(s-1) \\ 0 & 0 & as \end{pmatrix}, \ T^{(1)} = \begin{pmatrix} \frac{1-as}{2} & 0 & 0 \\ \frac{(1-as)(s-1)}{2s} & \frac{1-as}{2s} & 0 \\ \frac{(1-as)(s-1)}{2s} & 0 & \frac{1-as}{2s} \end{pmatrix}, \text{ and } T^{(2)} = \begin{pmatrix} \frac{1-as}{2s} & \frac{(1-as)(s-1)}{2s} & 0 \\ 0 & \frac{1-as}{2} & 0 \\ 0 & \frac{(1-as)(s-1)}{2s} & \frac{(1-as)(s-1)}{2s} \end{pmatrix}, \quad \text{(S1)}$$

where s controls the contraction coefficient and a controls the probability of selecting the maps. This HMC produces constant probability functions:

$$p^{(0)} = as \ , \ p^{(1)} = \frac{1-as}{2} \ , \text{ and } p^{(2)} = \frac{1-as}{2} \ .$$

and, therefore, linear mappings, since $f^{(0)} = \langle \eta | T^{(i)} / p^{(i)}(\eta)$. The constant probability functions make the entropy rate trivial to calculate. And, the linearity of the mappings does the same for the Lyapunov exponents.

Setting $s = 2$ and $a = 1/6$, results in equal probability for all maps and gives the standard Sierpinski triangle shown in Fig. S2. In this case, the entropy rate is $h_\mu = \log_2(3)$ and the Lyapunov exponents are both $-\log_2(2)$. Plugging this into Eq. (13) returns the well-known fractal dimension of the Sierpinski triangle, $\log_2 3 / \log_2 2 \approx 1.585$.

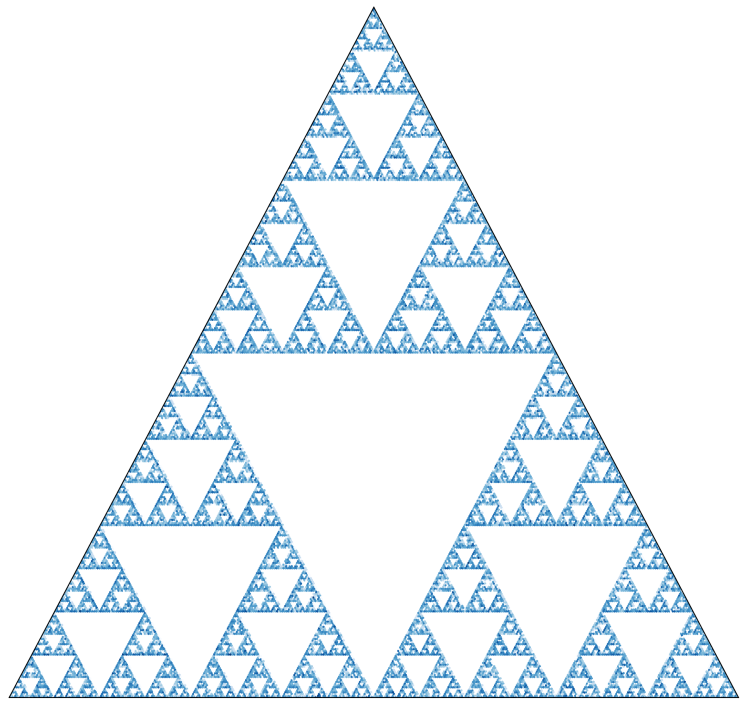


FIG. S2. Attractor of the 3-state, 3-symbol machine specified in Eq. (S1), with $s = 2$ and $a = 1/6$.

Appendix E: Lyapunov Exponents

A positive *Lyapunov characteristic exponent* for a dynamical system measures the exponential rate of separation of trajectories that begin infinitesimally close. Since, typically, the separation rate depends on the direction of the initial separation, we use a spectrum of Lyapunov exponents, with one exponent for each state-space dimension. In a chaotic dynamical system, at least one Lyapunov exponent is positive. In general, the Lyapunov exponent spectrum for an N-dimensional dynamical system with mapping $x_{n+1} = F(x_n)$ depends on the initial condition x_0. However, here we consider ergodic systems, for which the spectrum does not.

Consider the map's *Jacobian* matrix:

$$J = \frac{\partial F}{\partial x}$$

and the evolution of vectors in the tangent space, controlled by:

$$\dot{Y} = YJ ~,$$

where $Y(0) = \mathbb{I}_N$ and $Y(t)$ describes how an infinitesimal change in $x(0)$ has propagated to $x(t)$. Let $\{y_1, \ldots, y_N\}$ be the eigenvalues of the matrix $Y(t)Y(t)^\intercal$. Then, the Lyapunov exponents are:

$$\lambda_i = \lim_{t \to \infty} \frac{1}{2t} \log y_i ~.$$

The *Lyapunov numbers* were introduced and proven to exist by Oseledets [59]. The Lyapunov exponents are merely the logarithms of Lyapunov numbers.

The most common way of calculating an IFS's Lyapunov spectrum, employed to produce the results in Fig. 6, is the *pull-back* method. The basic idea is that for an IFS in the $N-1$ simplex, defined by an N-state HMC, there will be $N-1$ independent directions of contraction. These directions are represented by a coordinate frame of $N-1$ vectors that are kept orthogonal and normalized. This coordinate frame is carried along a long orbit of the IFS. At each time step, the Jacobian is used to evolve the frame. We track the contraction rate of each vector, and this becomes the estimation of the Lyapunov exponents.

In contrast to applying this approach to deterministic dynamical systems, as is more familiar, an IFS's stochastic nature introduces additional error. Since the Jacobian varies not just across the simplex, but also for the selected maps, the orbit must be long enough to sample from the IFS attractor's distribution accurately for each possible mapping function. This being said, there it has been established that the pull-back method works for IFS spectra, given a sufficiently long orbit [60].

The prequel, on estimating the entropy rate of HMC processes, made use of error-bounding techniques from Markov chain Monte Carlo (MCMC) [32]. Since here we are estimating Lyapunov exponents by sampling from the Blackwell distribution, similar error-bounding techniques apply. In this analysis, there are two fundamental sources of estimation error. First, 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.

Bounding these error sources requires estimating the autocorrelation function, which can be done from long sequences of samples. If we have the nonunifilar HMC in hand, it is a simple matter of sweeping through increasingly long sequences of generated samples until we observe convergence of the autocorrelation function. An alternative method of approximating the infinite-state HMC with a finite-state approximation is discussed in detail in our previous work [32]. The upshot is that the method here generally efficiently leads to accurate estimates of the LCE spectrum.

For completeness, we note that there are alternative methods to calculate Lyapunov exponents; see, e.g., Refs. [61, 62]. These methods may be more appropriate in specific applications. That said, the accuracy and applicability of Lyapunov exponent estimation is not the focus here.

Appendix F: Overlap Estimation

To estimate the size of a mixed-state attractor and overlap of mapping functions in Fig. 6, a combination of techniques were used. Let's briefly summarize the method here.

First, $250,000$ different HMCs were generated using a 500×500 parameter grid over $\alpha = [0, 1]$ and $x = [0, 0.5]$. Each HMC was defined by plugging the appropriate parameter values into the symbol-labeled transition matrices in Eq. (16). From this HMC, the mapping and probability functions were defined (see Eq. (7) and Eq. (6)), producing a place-dependent IFS.

For each IFS and associated HMC, $10,000$ mixed states were generated from an initial randomized state, throwing away the first $5,000$ as transients. Using a spatial algorithm from the SciPy Python package, a convex hull was drawn around this set of points, with a small buffer. This convex hull (the attractor "outline") was converted into a polygon. This polygon was then evolved independently by each symbol-labeled mapping function, producing three polygons, each associated with a symbol. This may be visualized by referencing Fig. 5, where the evolved polygons are depicted on top of the mixed-state attractor, each with a distinct color. We can see that the combination of these polygons must necessarily cover the attractor.

These three symbol-labeled polygons were then combined into a single polygon or multipolygon (a polygon with "holes" that are themselves polygons) using the geometry-processing Python module Shapely [63]. This produces a more accurate outline of the attractor than the convex hull. This process may then be repeated with the new outline for as many iterations as desired, until a polygon or multipolygon that covers the mixed state attractor with the desired level of accuracy is produced. The same result could be achieved by beginning with the entire simplex as the initial outline, without any production of mixed-states. However, the step of estimating the convex hull sharply reduces the number of required iterations and, more importantly, makes the required number more equal across parameter space. To see this, consider that attractors which take up less of the simplex require several iterations to converge to the small size of the attractor. By initializing with the convex hull, the process of converging to the basic shape of the attractor is skipped, and the iterations are merely refinements.

In rendering Fig. 6, we found that we could produce good outlines across parameter space by evolving the convex hull three times. To produce Fig. 6a, the area of the resultant polygon or multipolygon was found using Shapely

[63]. To produce Figs. 6c and 6d, the outline was evolved one more time by each map, and the resultant polygons or multipolygons were checked for intersection. For the binary overlap/no-overlap plot in Fig. 6d, only the existence of overlap somewhere on the attractor was considered. For the percentage overlap in Fig. 6c, the area of the total outline that was comprised of overlapping polygons—whether only two or all three—was compared to the total area. The subtlety of whether a region of overlap includes two or three maps was largely ignored here, but will be considered in future discussions of the overlap problem.