Ambiguity 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: November 17, 2021)

The ϵ-machine is a stochastic process' optimal model—maximally predictive and minimal in size. It often happens that to optimally predict even simply-defined processes, probabilistic models—including the ϵ-machine—must employ an uncountably-infinite set of features. To constructively work with these infinite sets we map the ϵ-machine to a place-dependent iterated function system (IFS)—a stochastic dynamical system. We then introduce the ambiguity rate that, in conjunction with a process' Shannon entropy rate, determines the rate at which this set of predictive features must grow to maintain maximal predictive power over increasing horizons. We demonstrate, as an ancillary technical result that stands on its own, that the ambiguity rate is the (until now missing) correction to the Lyapunov dimension of an IFS's attracting invariant set. For a broad class of complex processes and for the first time, this then allows calculating their statistical complexity dimension—the information dimension of the minimal set of predictive features.

Keywords: Hidden Markov process, minimal stochastic state machines, entropy rate, ambiguity rate, predictive feature, optimal prediction, state growth

I. INTRODUCTION

The bedrock of scientific inquiry is model building. The act of *modeling* a natural system serves many purposes, among them prediction of future behavior, generating surrogate data, pattern recognition, and pattern discovery. These goals are not independent, nonetheless they are sufficiently distinct to merit substantial and separate attention. This is particularly the case when confronted with modeling *complex systems*—those that create intricate and delicate patterns through their internal interplay of stochasticity and determinism. These systems are often identified by the presence of collectively-interacting subsystems, long-range correlations, and visually-striking emergent structures. In this regime, where direct observation is difficult at best and uninformative at worst, the importance of building models is amplified and we require a robust theoretical framework to guide their construction, analysis, and interpretation.

Computational mechanics [1] introduces a suite of tools to analyze complex systems in terms of their informational architecture by integrating Turing's computation theory [2–4], Shannon's information theory [5], and Kolmogorov's dynamical systems theory [6–10]. Its most basic statistic is the *ϵ-machine*—a system's maximally predictive, minimal, and unique model. The ϵ-machine's *causal states* are the minimal set of maximally-predictive features—the unique possible futures conditioned the system's infinite past. Qualitatively, the information stored in the causal states is a process' *statistical complexity* C_μ, a measure of the memory resources a system employs to generate its behavior and support its organization. Due to its uniqueness and minimality, via an Occam's razor argument, we may identify the ϵ-machine as a system's canonical description.

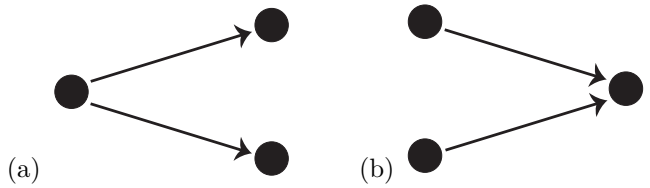


FIG. 1. (a) Equivocation: Same input sequence to a communication channel leads to different outputs. (b) Ambiguity: Two different inputs lead to same output. The strategy underlying Shannon's proof of his second coding theorem is to find channel inputs that are least ambiguous given the channel's distortion properties.

We may choose to view the ϵ-machine either as distinct from the system (a predictor of the true system's behavior) or as representing the system itself (a generator of data statistically indistinguishable from observations). There is a third option, though. Consider the ϵ-machine as a memoryful communication channel, mapping pasts to futures. In this view, we imagine an adaptive channel interacting with a system via a time-series of discrete observations. At each time, it updates its internal configuration to represent the current optimal prediction of the system. The channel's set of possible configurations are the causal states and these require C_μ bits of memory to operate.

The student of information theory will recall that Shannon, in his analysis of information transmission through channels, introduced two mechanisms: *equivocation*, in which the same input may lead to distinct outputs, and *ambiguity*, in which two different inputs may

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

lead to the same output; see Fig. 1. When our channel is taken to be the ϵ-machine, the equivocation rate of the channel is the *entropy rate* h_μ of the underlying system—the rate at which the system generates future information. This is guaranteed by the ϵ-machine's predictive optimality—the only noise in the channel arises from the system's intrinsic randomness h_μ.

The following proposes a parallel quantity—the *ambiguity rate* h_a—as a new intrinsic complexity measure. The ambiguity rate tracks the rate at which an optimal predictor of a system discards information by introducing uncertainty over the infinite past. The difference $h_\mu - h_a$ between this rate and the entropy rate describes the *growth rate* of the information stored in a process' optimally predictive features. When $h_\mu = h_a$, this rate vanishes and the associated ϵ-machine's internal causal-state process is stationary. Explicitly, for the ϵ-machine to consist of a finite set of predictive features, it must forget information at the same rate at which the system generates it, so as to not grow the size of the model over time. However, when $h_\mu > h_a$, any optimal predictor must accumulate new information over time to sustain accurate predictions.

This introduces a challenge, since $h_\mu = h_a$ is atypical for a broad class of stochastic processes—as our prior works demonstrated [11, 12]—in particular, those used not only in the study of complex systems [1], but also in coding theory [13], stochastic processes [14], stochastic thermodynamics [15], speech recognition [16], computational biology [17, 18], epidemiology [19], and finance [20]. In point of fact, for many complex systems, the predictive-feature set is uncountably infinite and the structural complexity C_μ diverges, requiring new tools to characterize these systems' complexity. The recent work introduced a suite of tools to address this state of affairs.

The key realization was identifying a process' ϵ-machine as the attractor of a hidden Markov-*Driven Iterated Function System* (DIFS) [11]. First, we showed that this gave efficient and accurate calculation of a process' Shannon entropy rate h_μ. Second, we introduced a new measure of structural complexity—the *statistical complexity dimension* d_μ—that tracks C_μ's divergence and gives the information dimension of the distribution of predictive features [12]. Previously, accurate calculation of d_μ was contingent on the DIFS meeting restrictive technical conditions. Introducing ambiguity rate h_a reframes these constraints information-theoretically, effectively lifting them. The result is a new method to accurately calculate d_μ for a broad class of complex processes.

Introducing h_a is not merely a means to an end—allowing accurately calculating d_μ—but also a first step towards a full dynamical decomposition of information in physical systems. Several novel informational measures, such as "transfer entropy" [21] and "causation entropy" [22], were introduced to solve this puzzle. Unfortunately, they met with mixed success and were criticized for failing to truly capture "information flow" [4, 21, 23]. We argue that $h_\mu - h_a$ points instead at an *information flux* that measures the difference in rates of information flow into and out of a system. This framing is a natural extension of previously successful information decomposition methods [24] that identified how information is embedded in time series. Although the full tool set necessary for the decomposition is not developed here, we believe that the new perspective introduced is key to finally developing a *dynamics of information* for complex systems and model inference.

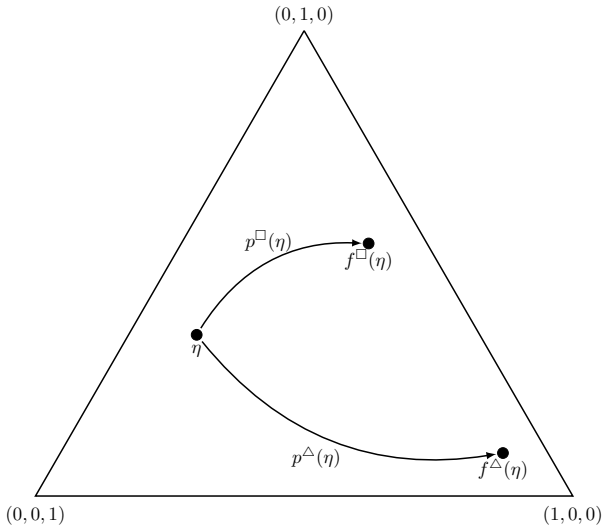


FIG. 2. How a hidden Markov-driven iterated function system (DIFS) generates a hidden Markov process: An initial state η—a distribution over three states: $(0, 0, 1)$, $(0, 1, 0)$, and $(1, 0, 0)$—in the 2-simplex is associated with a transition probability distribution over the alphabet $\mathcal{A} = \{\square, \triangle\}$. If the emitted symbol selected from this distribution is $\square$, the next state is generated according to the associated mapping function $f^{(\square)}(\eta)$ and the probability distribution is updated accordingly. The same steps are followed if the symbol is $\triangle$ using $f^{(\triangle)}(\eta)$, resulting in an emitted process $\mathcal{P}$ over symbols $\mathcal{A}$.

The development introduces and motivates the ambiguity rate h_a. Sections II and III review stochastic processes and information theory, respectively, and may be skipped by the familiar reader. Section IV introduces hidden Markov-driven iterated function systems. Section V then discusses the statistical complexity dimension d_μ and the *overlap problem*—a long-standing issue in the dimension theory of iterated function systems. Section VI introduces h_a from an information-theoretic perspective, motivating it as a solution to and a measure of the overlap problem. Various interpretations are explored, including an historical note on Shannon's original *dimension rate* from 1948. Finally, to illustrate our algorithm's effectiveness and the challenges for very complex processes, Section VII works through multiple examples, including processes generated by stationary and nonstationary ϵ-machines.

II. 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, 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's length-ℓ *word distributions* $\Pr(X_{0:\ell})$.

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 [25] whose input is a Markov process [14]. Somewhat surprisingly, though well known, the output process can have realizations with arbitrarily-long correlations.

III. INFORMATION THEORY

Beyond its vast technological applications to communication systems [25], Shannon's information theory [5] is a widely-used foundational framework that provides tools to describe how stochastic processes generate, store, and transmit information. In particular, we use information theory to study complex systems as it makes minimal assumptions as to the nature of correlations between random variables and handles multi-way, nonlinear correlations that are common in complex processes. Here, we now briefly recall several basic concepts needed in the following.

The most basic quantity within information theory is the Shannon *entropy*. Intuitively, it measures the amount of information that one learns when observing a sample of a random variable. (It is, equivalently modulo sign, also the amount of uncertainty one faces when predicting the sample.) The entropy $H[X]$ of the random variable X is:

$$H[X] = -\sum_{x \in \mathcal{A}} \Pr(X = x) \log_2 \Pr(X = x) \ . \qquad (1)$$

In addition to focusing on individual samples, we can probe the relationship between two jointly-distributed random variables, say, X and Y. There is the *joint entropy* $H[X, Y]$, of the same functional form but applied to the joint distribution $\Pr(X, Y)$. And, there is *conditional entropy* that gives the amount of information learned from observation of one random variable X given another Y:

$$H[X|Y] = H[X, Y] - H[Y] \ . \qquad (2)$$

Conditional entropy can be generalized to describe processes in terms of the *intrinsic* randomness—the amount of information one learns upon observing the next emitted symbol X_0, given complete knowledge of the infinite past. This is the Shannon *entropy rate*:

$$h_\mu = \lim_{\ell \to \infty} H[X_0|X_{-\ell:0}] \ , \qquad (3)$$

the irreducible amount of information gained in each time step.

The fundamental measure of correlation between random variables is the *mutual information*. It can be written in terms of Shannon entropies:

$$I[X; Y] = H[X, Y] - H[X|Y] - H[Y|X] \ . \qquad (4)$$

As should be clear by inspection, the mutual information between to variables is symmetric. When X and Y are independent, the mutual information between them vanishes. As with entropy, we may condition the mutual information on another random variable Z, giving the *conditional mutual information*:

$$I[X; Y|Z] = H[X|Z] + H[Y|Z] - H[X, Y|Z] \ . \qquad (5)$$

The conditional mutual information is the amount of information shared by X and Y, given we know the third Z. Note that X and Y can share mutual information, but be conditionally independent. Moreover, conditioning on a third variable Z can either increase or decrease mutual information [25]. That is, the two variables can appear more or less dependent, given additional data.

IV. DRIVEN ITERATED FUNCTION SYSTEM

Our main objects of study are hidden Markov processes. The following introduces driven iterated function system as a class of predictive models for them.

Definition 1. *An N-dimensional hidden Markov*-driven iterated function system *(DIFS)* $\left(\mathcal{A}, \mathcal{V}, \boldsymbol{\mathcal{R}}, \{T^{(x)}\}, \{p^{(x)}\}, \{f^{(x)}\} : x \in \mathcal{A}\right)$ *consists of:*

1. *a finite alphabet $\mathcal{A}$ of k* symbols $x \in \mathcal{A}$,
2. *a set $\mathcal{V}$ of N* presentation states,
3. *a set of* states $\boldsymbol{\mathcal{R}} \subset \Delta^{N-1}$, *over N-dimensional presentation-state distributions $\eta \in \boldsymbol{\mathcal{R}}$,*
4. *a finite set of N by N* symbol-labeled substochastic matrices $T^{(x)}$, $x \in \mathcal{A}$,
5. *a set of k* symbol-labeled probability functions $p^{(x)} = \langle \eta | \, T^{(x)} \, | \mathbf{1} \rangle$, *and*
6. *a set of k* symbol-labeled mapping functions $f^{(x)} = \frac{\langle \eta | T^{(x)}}{p^{(x)}(\eta)}$.

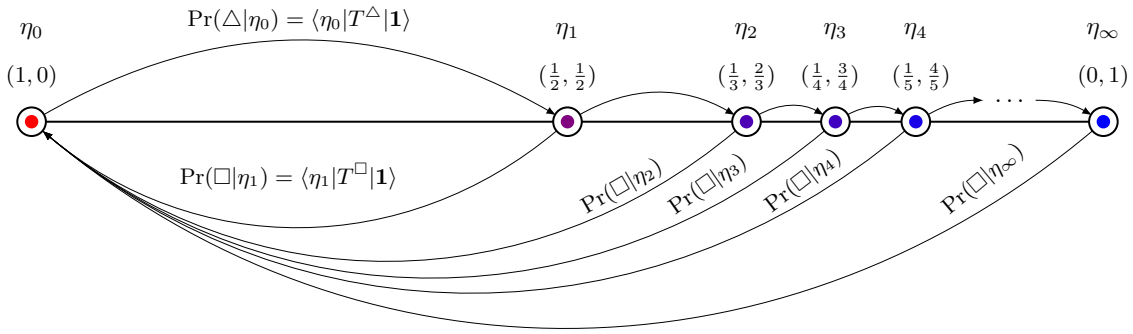


FIG. 3. The states and transitions of a hidden Markov-driven iterated function system (DIFS) discussed in Section VII A embedded in the 1-simplex. In this case, the set $\boldsymbol{\mathcal{R}}$ of states is countable and each subsequent application of $f^{(\triangle)}$ brings η nearer to $\eta_\infty = (0, 1)$. The latter is reached only after observing infinitely many $\triangle$s. The countable nature of $\boldsymbol{\mathcal{R}}$ arises from the fact that one of the mapping functions is a constant: $f^{(\square)} = (1, 0)$.

The *(N-1)-simplex* Δ^{N-1} is the set of presentation-state probability distributions 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 presentation-state vector η to avoid confusion with temporal indexing.

Note that the set of substochastic matrices must sum to the nonnegative, row-stochastic matrix $T = \sum_{x \in \mathcal{A}} T^{(x)}$—the transition matrix for the presentation-state Markov chain. This ensures that $\sum_{x \in \mathcal{A}} p^{(x)}(\eta) = 1$ for all $\eta \in \Delta^{N-1}$.

The transition probability between states is equivalent to the probability of seeing the symbol that leads to that state:

$$\Pr(X_t = x, R_{t+1} = \eta_{t+1} | R_t = \eta_t) = \begin{cases} p^{(x)}(\eta_t), & \eta_{t+1} = f^{(x)}(\eta_t) \\ 0 & \eta_{t+1} \neq f^{(x)}(\eta_t) \end{cases} . \tag{6}$$

Each symbol must by definition lead to a unique state, although two symbols may lead to the same state. Figure 2 shows how a DIFS generates a hidden Markov process: Given an initial state $\eta_0 \in \Delta^{N-1}$, the probability distribution $\{p^{(x)}(\eta_0) : x = 1, \ldots, k\}$ is sampled. According to the realization x_0, apply the mapping function to map η_0 to the next state $\eta_1 = f^{(x_0)}(\eta_0)$. According to the new probability distribution defined by η_1, draw x_1 and repeat. This action generates our emitted process $\mathcal{P}$: $x_0, x_1, x_2, \ldots$.

This describes the random dynamical system—the DIFS—that generates the hidden state sequence $\eta_0, \eta_1, \eta_2, \ldots$. As we previously showed, the attractor of this dynamical system is the invariant set $\boldsymbol{\mathcal{R}}$ of states and their evolution is ergodic [11, 26]. Additionally, the attractor has a unique, attracting, invariant measure known as the *Blackwell measure* $\mu_B(\boldsymbol{\mathcal{R}})$ [27]. Although $\boldsymbol{\mathcal{R}}$ may be countable, as for the DIFS depicted in Fig. 3, in general, $\boldsymbol{\mathcal{R}}$ is uncountably infinite and fractal in nature, as in the examples in Fig. 4.

DIFS states are *predictive* in the sense that they are functions of the prior observables. Consider an infinitely-long past that, in the present, has induced some state η. It is not guaranteed that this infinitely-long past induce a unique state, but it is the case that any state induced by this past must have the same conditional future distribution $\Pr(X_{0:\infty}|\cdot)$. Indeed, for the task of prediction, knowing the previous state is as good as knowing the infinite past: $\Pr(X_{0:\ell}|\mathcal{R}_0 = \eta) = \Pr(X_{0:\ell}|X_{-\infty:0})$ for all $\ell \in \mathbf{N}^+$.

Therefore, the DIFS is a predictive model of the process $\mathcal{P}$ it generates. (This is in contrast to it being merely a generative model—whose only requirement is to produce all and only the set of realizations.) Borrowing from the language of automata theory, we refer to the set of states $\boldsymbol{\mathcal{R}}$ plus its transition dynamic—$\Pr(x_t|\eta_t)$ and $\Pr(\eta_{t+1}|\eta_t, x_t)$—as a *state machine* or, simply *machine* that optimally predicts $\mathcal{P}$. When each state is associated with a unique future distribution, we have a canonical predictive model that is unique: a process' *ϵ-machine* [1].

Definition 2. *An* ϵ-machine *is a DIFS with* probabilistically distinct states*: For each pair of distinct states* $\eta, \zeta \in \boldsymbol{\mathcal{R}}$ *there exists a finite word* $w = x_{0:\ell-1}$ *such that:*

$$\Pr(X_{0:\ell} = w | \mathcal{R}_0 = \eta) \neq \Pr(X_{0:\ell} = w | \mathcal{R}_0 = \zeta) \ .$$

A process' ϵ-machine is its optimally-predictive, minimal model, in the sense that the set $\boldsymbol{\mathcal{R}}$ of predictive states is minimal compared to all of the process' other predictive models. By capturing a process' structure and

not merely being predictive, an ϵ-machine's states are called *causal states*. Unless otherwise noted, we assume that all DIFS discussed here are ϵ-machines.

Calculating the Shannon entropy rate for a process generated by a DIFS was the focus of our first discussion of DIFSs [11]. Due to the associated process' ergodicity, h_μ may be written as a time average:

$$\widehat{h_\mu}^B = -\lim_{\ell\to\infty}\frac{1}{\ell}\sum_{t=0}^{\ell}\sum_{x\in\mathcal{A}}\Pr(x|\eta_\ell)\log_2\Pr(x|\eta_\ell)\ . \quad (7)$$

This tracks the uncertainty in the next symbol x given our current causal state η_t, averaged over the Blackwell measure. It quantifies the intrinsic randomness of the process $\mathcal{P}$.

V. STATISTICAL COMPLEXITY DIMENSION

Images of the self-similar state sets $\boldsymbol{\mathcal{R}}$ of DIFS are evocative (again see Fig. 4) and lead naturally to questions about how $\boldsymbol{\mathcal{R}}$'s geometric properties relate to intrinsic properties of the underlying process $\mathcal{P}$. To begin to answer this, we identify a process' memory with the information required to specify its ϵ-machine states; i.e., the minimal amount of information needed to predict $\mathcal{P}$. This may be measured either in terms of the cardinality $|\boldsymbol{\mathcal{R}}|$ of causal states or the amount of historical Shannon entropy they store—that is, the *statistical complexity* C_μ.

Definition 3. *A process'* statistical complexity *is the Shannon entropy stored in its ϵ-machine's causal states:*

$$\begin{aligned} C_\mu &= \mathrm{H}[\Pr(\boldsymbol{\mathcal{R}})] \\ &= -\sum_{\eta\in\boldsymbol{\mathcal{R}}} p(\eta)\log_2 p(\eta)\ . \end{aligned} \quad (8)$$

From the definitions above, a process' ϵ-machine is its smallest predictive model, in the sense that both $|\boldsymbol{\mathcal{R}}|$ and C_μ are minimized by a process' ϵ-machine, compared to all other predictive models. Due to the ϵ-machine's unique minimality, we identify the ϵ-machine's C_μ as the process' memory.

However, when the set $\boldsymbol{\mathcal{R}}$ of causal states is infinite, the statistical complexity may diverge. In this case, C_μ is no longer an appropriate complexity measure to distinguish processes. Despite this, a need remains: It is clear that processes with infinite state sets differ significantly in internal structure, as shown in Fig. 4. In this case, we turn to the *statistical complexity dimension*, defined as the rate of divergence of the statistical complexity, to serve as a measure of structural complexity. This leaves us with an abiding question, though, What does it mean that a finitely-specified process' state information (memory) diverges?

A. Dimension and Causal State Divergence

A set's *dimension*, construed most broadly, gives the rate at which a chosen size-metric diverges with the scale at which the set is observed [28–32]. Fractional dimensions, in particular, are useful to probe the "size" of sets when cardinality alone is not informative. "Fractal dimension", said in isolation, is often taken to refer to the box-counting or Minkowski-Bouligand dimension. The following, though, determines the *information dimension*—a dimension that accounts for the scaling of a measure on a fractional dimension set. In this case, our measure of interest is the Blackwell measure μ_B over our causal states $\boldsymbol{\mathcal{R}}$.

Consider the state set $\boldsymbol{\mathcal{R}}$ on the $(N-1)$-simplex for a DIFS that generates a process $\mathcal{P}$. 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 state. Now, let each cell C in $\mathcal{F}(\epsilon)$ itself be a (coarse-grained) state and approximate the ϵ-machine dynamic by grouping all transitions to and from states encompassed by the same cell. This results in a finite-state Markov chain that generates an approximation of the original process $\mathcal{P}$ and has a stationary distribution $\mu(\mathcal{F}(\epsilon))$. Then $\mu_B(\boldsymbol{\mathcal{R}})$'s *information dimension* is:

$$d_1(\mu_B(\boldsymbol{\mathcal{R}})) = \lim_{\epsilon\to 0}\frac{H_\mu[\mathcal{F}(\epsilon)]}{\log\left(\frac{1}{\epsilon}\right)}\ , \quad (9)$$

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

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

$$H_\mu[\mathcal{F}] \sim d_1(\mu_B)\cdot\log\frac{1}{\epsilon}\ . \quad (10)$$

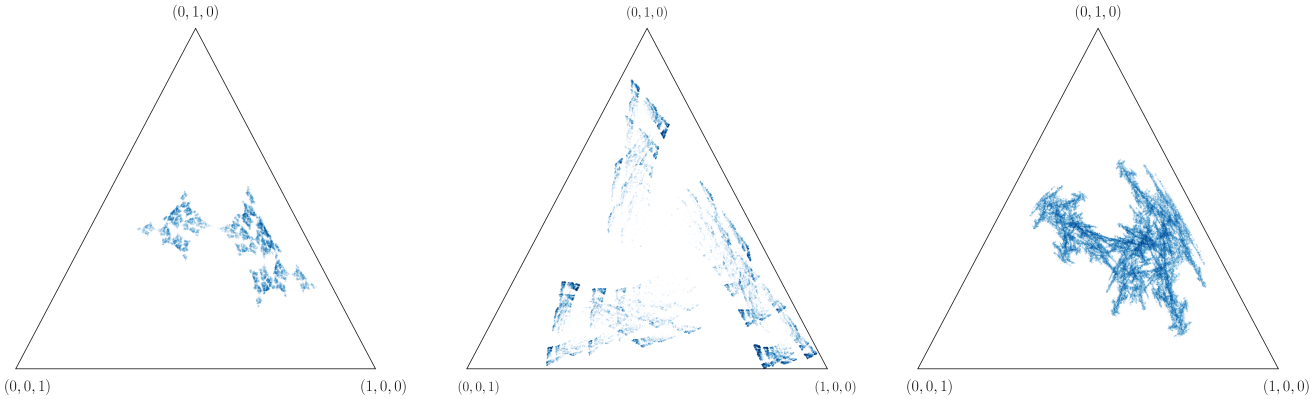
Applied to a process $\mathcal{P}$'s ϵ-machine, d_1 describes the divergence rate of statistical complexity C_μ:

$$C_\mu(\epsilon) \sim d_\mu\cdot\log\frac{1}{\epsilon}\ . \quad (11)$$

In this way, we refer to the ϵ-machine's information dimension $d_1(\mu_B)$ as $\mathcal{P}$'s *statistical complexity dimension* d_μ.

B. Determining Statistical Complexity Dimension

Directly calculating the statistical complexity dimension using Eq. (9) is nontrivial, as it often requires estimating a fractal measure. Fortunately, as our two previous works discussed and as we now show, the intractability can be circumvented by leveraging the process's asso-



(a) Set of states generated by the "delta" DIFS.

(b) Set of states generated by the "Nemo" DIFS.

(c) Set of states generated by the "gamma" DIFS.

FIG. 4. Hidden Markov driven iterated function system (DIFS) may generate state sets with a wide variety of structures, many fractal in nature. Each subplot displays 10^5 states of a different DIFS. The DIFSs themselves are specified in Appendix A.

ciated generating dynamical system—the DIFS—to calculate d_μ [11, 12].

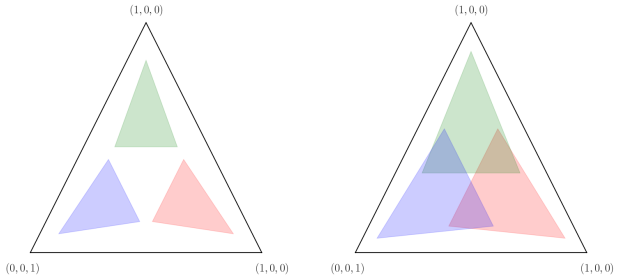
For a dynamical system, the *spectrum of Lyapunov characteristic exponents* $\Gamma = \{\lambda_1, \ldots, \lambda_N : \lambda_i \geq \lambda_{i+1}\}$ [33, 34] measures expansion and contraction as the average local growth or decay rate, respectively, of orbit perturbations. The result is a list of rates that indicate long-term orbit instability ($\lambda_i > 0$) and orbit stability ($\lambda_i < 0$) in complementary directions.

Consider covering an attractor generated by a dynamical system f with hypercubes of side length ϵ. After applying f to a hypercube k times, the side lengths are approximately $\epsilon e^{\lambda_1 k}, \epsilon e^{\lambda_2 k}, \ldots$, assuming that the hypercube orientation is chosen appropriately. This property allows combining the elements of the spectrum Γ into an expression approximating the growth rate of hypercubes needed to cover the attractor, as $\epsilon \to 0$. In turn, this implies a natural relationship between the Γ and dimensional quantities, such as Eq. (9). The *Lyapunov dimension* [35] has been conjectured to be equivalent to the information dimension for "typical systems" f.

Our previous work showed how to calculate Γ for DIFSs [12]. However, since DIFSs are *random* dynamical systems, additional orbit expansion arises from the stochastic selection of the maps $f^{(x)}$. Indeed, for DIFSs, *all* expansion arises from this stochastic choice, which is measured by the Shannon entropy rate h_μ of the generated process $\mathcal{P}$. That is to say, for DIFSs, $\lambda_i < 0$ for all i while h_μ monitors the expansive exponent.

With this in mind, we adapt the Lyapunov dimension expression to DIFSs as follows:

$$\widetilde{d_\Gamma} = \begin{cases} k + \dfrac{\Lambda(k) + h_\mu}{|\lambda_{k+1}|}, & -\Lambda(N) > h_\mu \\ N\ , & -\Lambda(N) \leq h_\mu \end{cases} , \qquad (12)$$



(a) (b)

FIG. 5. Overlap problem on the 2-simplex Δ^2: Two distinct DIFSs (given in Appendix A) are considered, each with three mapping functions. Images of the mapping functions over the entire simplex are depicted as regions in red, blue, and green. (a) Images of the mapping functions $f^{(\triangle)}$, $f^{(\square)}$, and $f^{(\circ)}$ do not overlap—every possible state has a unique preimage. (b) Images of the mapping functions overlap—there exist $\eta_1, \eta_2 \in \Delta^2$ such that $f'^{(\triangle)}(\eta_1) = f'^{(\square)}(\eta_2) = \eta_3$. This case is an *overlapping DIFS*.

where we introduce the Lyapunov spectrum partial sum $\Lambda(m) = \sum_{i=1}^{m} \lambda_i$ and $k = 0, 1, 2, \ldots, N-1$ is the largest index for which $-\Lambda(k) < h_\mu$. Note $\Lambda(m) < 0$ for $m = 1, 2, \ldots, N$ and we take $\Lambda(0) = 0$. Readers familiar with the Lyapunov dimension should take care as we reindexed from the traditional presentation of d_Γ for readability.

Under specific technical conditions, d_Γ is exactly the information dimension of the DIFS's attractor: $d_\Gamma = d_\mu$ [36]. Generally, relaxing the conditions, $\widetilde{d_\Gamma}$ only upper bounds the statistical complexity dimension:

$$\widetilde{d_\Gamma} \geq d_\mu\ . \qquad (13)$$

The extent to which the bound is not saturated is in large part determined by the *open set condition*, which we now discuss. We, then, turn to solve the associated "overlap problem". This leads to an exact expression for DIFS attractor information dimension d_μ.

C. The Overlap Problem

The *overlap problem* is a long-standing concern for iterated function systems that arises from the overlap of the ranges of the symbol-labeled mapping functions $f^{(x)}$. Figure 5 illustrates the issue. Specifically, to quantitatively count system orbits we must properly monitor orbit divergence and convergence. This then requires distinguishing between iterated function systems that meet the open set condition (OSC) and those that do not.

Definition 4. *An iterated function system with mapping functions $f^{(x)} : \Delta \to \Delta$ satisfies the* open set condition *(OSC) if there exists an nonempty open set $U \in \Delta$ such that for all $x, x' \in \mathcal{A}$:*

$$\bigcup_{x=1}^{k} f^{(x)}(U) \subset U \ ,$$
$$f^{(x)}(U) \cap f^{(x')}(U) = \emptyset \ , \quad x \neq x' \ .$$

IFSs that meet the OSC are nonoverlapping IFSs*.*

When the OSC is not met, the inequality in the d_μ bound Eq. (13) becomes strict. This is a consequence of using h_μ as our measure of state space expansion in Eq. (12). The Shannon entropy rate tracks the uncertainty in the next symbol x given our current causal state η_t, averaged over the Blackwell measure. From a dynamical systems point of view, we identify this as the typical growth rate of orbits (words) in symbol space.

When the OSC is met, the Shannon entropy rate also measures the typical growth rate of orbits in the $(N-1)$-simplex. Observing x, current state η_t transitions to the next state η_{t+1} via application of the mapping function $\eta_{t+1} = f^{(x)}(\eta_t)$. Since images of the map do not overlap, this is guaranteed to be a distinct new state—thus the number of distinct state sequences grow at the same rate as the number of words. Then, we may use h_μ to measure state-space expansion.

However, when the OSC is not met, it is possible for two distinct states $\eta_t, \zeta_t \in \Delta$ to map to the same next state on different symbols, by occupying the "overlapping region", as depicted in Fig. 5. In this case, $\eta_{t+1} = \zeta_{t+1}$ has no unique pre-image. This introduces ambiguity about the past, given knowledge of the current state. As a consequence, using the Shannon entropy rate as a proxy for the state expansion rate implies a more rapid expansion in state space than is actually occurring. This indicates the need to adjust the use of h_μ in determining d_μ.

VI. AMBIGUITY RATE

We propose that the *ambiguity rate* properly corrects $\widetilde{d_\Gamma}$ of Eq. (12) from overcounting orbits. Since the problem at hand is an overestimation in uncertainty in our state space, we must identify and quantify mechanisms of state uncertainty reduction when the OSC is not met. Consider that when the OSC is met, every state η_t has a unique pre-image η_{t-1} that can only be reached via a single, specific observed symbol. When the OSC is not met, for a subset of $\eta \in \boldsymbol{\mathcal{R}}$ there is uncertainty about the previous state, the previous symbol, or both. Quantifying this ambiguity about the past is the goal in constructing the ambiguity rate h_a.

Intuitively, it would seem that generating uncertainty in reverse time is equivalent to reduction of uncertainty in forward time. The following shows that this is the case and that the ambiguity rate is the necessary correction to the DIFS dimension formula Eq. (12).

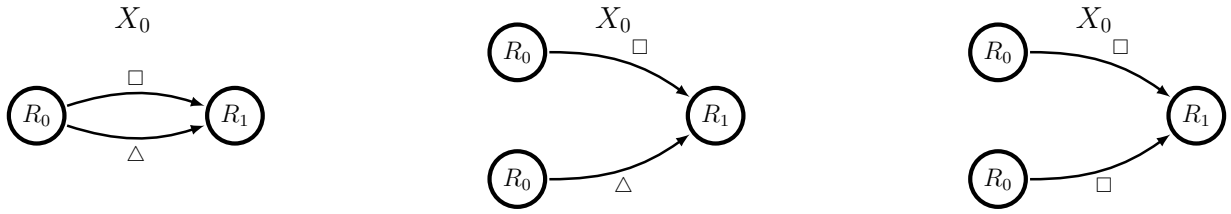
A. Sources of State Uncertainty Reduction

For ϵ-machines represented as DIFSs, there are three distinct mechanisms that contribute to the ambiguity rate, as depicted in Fig. 6.

The first is *identical mapping functions*, depicted in Fig. 6 (a). When for $x, x' \in \mathcal{A}$, $f^{(x)}(\eta) = f^{(x')}(\eta)$ for all $\eta \in \boldsymbol{\mathcal{R}}$, we say that x and x' have identical mapping functions. In this case, the distinction between x and x' is not reflected in state sequences and produces ambiguity in the symbol sequence. We quantify this as the Shannon entropy in our current symbol, conditioned on the previous state and the next state: $H[X_t|\mathcal{R}_t, \mathcal{R}_{t+1}]$.

The second is *overlapping mapping functions*, which motivated this investigation and have already been defined. Their impact on the state machine is shown in Fig. 6 (b). In this case, two distinct symbols $x, x' \in \mathcal{A}$ map two distinct states $\eta, \zeta \in \boldsymbol{\mathcal{R}}$ to the same next state. Although the previous state affects the probability distribution over the observed symbol, the next state "forgets" that distinction. This is quantified by the mutual information shared by the current symbol and the previous state, conditioned on the next state: $I[X_t; \mathcal{R}_t|\mathcal{R}_{t+1}]$.

Finally, there is *noninvertibility in the mapping functions* themselves. If a single mapping function maps distinct states $\eta, \zeta \in \boldsymbol{\mathcal{R}}$ to the same next state, the pasts that led to η and ζ can no longer be distinguished. Figure 6 (c) shows this in general. However, it may also be observed in $f^{(\square)}$ from Fig. 3, which maps every state to $\eta_0 = (1, 0)$. Numerically, the reduction via this mechanism is measured by the Shannon entropy in the previous state, given our next state and current symbol: $H[\mathcal{R}_t|X_t, \mathcal{R}_{t+1}]$.



(A) $H[X_0|\mathcal{R}_0, \mathcal{R}_1]$ (B) $I[X_0; \mathcal{R}_0|\mathcal{R}_1]$ (C) $H[\mathcal{R}_0|X_0, \mathcal{R}_1]$

FIG. 6. Sources of ambiguity rate depicted in state machines: (A) $H[X_0|\mathcal{R}_0, \mathcal{R}_1] > 0$—Previous state $\mathcal{R}_0$ is mapped to the next state $\mathcal{R}_1$ by two distinct symbols. This occurs when two symbols have identical mapping functions. (B) $I[X_0; \mathcal{R}_0|\mathcal{R}_1] > 0$—Two distinct previous states $\mathcal{R}_0$ map to the same next state by distinct symbols, due to overlapping mapping functions. (C) $H[\mathcal{R}_0|X_0, \mathcal{R}_1] > 0$—Two distinct previous states $\mathcal{R}_0$ map to the same next state by the same symbol. This occurs when a mapping function is noninvertible.

Combining these three sources of uncertainty reduction defines the *ambiguity rate*:

$$\begin{aligned} h_a &= H[X_t|\mathcal{R}_t, \mathcal{R}_{t+1}] + I[X_t; \mathcal{R}_t|\mathcal{R}_{t+1}] \\ &\quad + H[\mathcal{R}_t|X_t, \mathcal{R}_{t+1}] \\ &= H[X_t, \mathcal{R}_t|\mathcal{R}_{t+1}] \ . \end{aligned} \tag{14}$$

This can be rewritten as a integral over $\boldsymbol{\mathcal{R}}$:

$$h_a = -\int_{\eta \in \boldsymbol{\mathcal{R}}} \mathrm{d}\mu_B(\eta) \sum_{\substack{x \in \mathcal{A}, \\ \zeta \in (f^{(x)})^{-1}(\eta)}} \Pr(x, \zeta|\eta) \log_2 \Pr(x, \zeta|\eta). \tag{15}$$

In this, we must be careful about the pre-images of η, due to the possibility of noninvertible mapping functions. The probability distribution inside the summation is given by the relationship:

$$\begin{aligned} &\Pr(X_0 = x, \mathcal{R}_0 = \zeta|\mathcal{R}_1 = \eta) = \\ &\qquad \frac{\mu_B(\mathcal{R}_0 = \zeta)}{\mu_B(\mathcal{R}_1 = \eta)} \times \Pr(X_0 = x|\mathcal{R}_0 = \zeta) \ . \end{aligned} \tag{16}$$

Calculating this distribution requires calculating or estimating the Blackwell measure, which may be nontrivial. Section VII discusses this in greater depth.

B. Correcting d_μ

The information-theoretic decomposition of ambiguity rate facilitates combining h_a and h_μ. Recall that for prediction, the states of a predictive model are equivalent to knowledge of the infinite past. Due to this, the Shannon entropy rate may be written $H[X_t|\mathcal{R}_t]$. Combining this with the ambiguity rate gives:

$$\begin{aligned} h_\mu - h_a &= H[X_t|\mathcal{R}_t] - H[X_t, \mathcal{R}_t|\mathcal{R}_{t+1}] \\ &= H[\mathcal{R}_{t+1}|\mathcal{R}_t, X_t] + H[\mathcal{R}_{t+1}] - H[\mathcal{R}_t] \\ &= \Delta H[\mathcal{R}_t] \ . \end{aligned}$$

Moving to the third line called on the fact that the symbol and state transitions are defined by functions. So, the difference between the Shannon entropy rate and the ambiguity rate gives the rate of growth of our causal state set $\mathcal{R}$.

Recall that the information dimension, as defined in Eq. (9), compares the average growth of occupied cells $\mathcal{F}$—taking into account the measure over those cells—as the cell size ϵ shrinks. To adhere to the main development, here we will not walk through the heuristic for how a dimensional quantity is determined from the Γ. (Though, this is briefly discussed in Section V B.) Nonetheless, we will show how the relationship between d_1, $h_\mu - h_a$, and Γ is intuitive for DIFSs in one dimension.

When the DIFS states lie in the 1-simplex, Γ consists of only one exponent $\lambda_1 < 0$, which is the weighted average of the Lyapunov exponents of each map:

$$\lambda_1 = \int \sum_x p^{(x)}(\eta) \log \left| \frac{df^{(x)}(\eta)}{d\eta} \right| d\mu \ ,$$

where μ is the Blackwell measure.

Now, consider a line segment in Δ^1 of length ϵ. Mapping this line forward k times by the DIFS produces, averaging over several iterations of this action, $2^{(h_\mu - h_a)k}$ new lines of length $\epsilon e^{\lambda_1 k} < \epsilon$. (Note that the use of base-two for Shannon entropy rather than base e follows convention; retained here for familiarity. In numerical calculation of d_μ, we recommend a consistent base be chosen for h_μ, h_a, and the Γ.) The logarithmic ratio of the growth rate of lines (as averaged over the Blackwell measure) compared to the shrinking of these lines is the simple ratio:

$$d_\mu = -\frac{h_\mu - h_a}{\lambda_1} \ .$$

This, of course, is exactly the definition of the information dimension Eq. (9) and is, assuming the DIFS is an ϵ-machine, the statistical complexity dimension d_μ.

For higher-dimensional DIFSs, we conjecture that the ambiguity rate is the adjustment to the IFS Lyapunov dimension formula that gives the information dimension:

$$\widetilde{d_\mu} = \begin{cases} k + \dfrac{\Lambda(k) + h_\mu - h_a}{|\lambda_{k+1}|}, & -\Lambda(N) > h_\mu - h_a \\ N \ , & -\Lambda(N) \leq h_\mu - h_a \end{cases} , \quad (17)$$

where, as in Eq. (12), $\Lambda(m)$ is the Lyapunov spectrum partial sum $\Lambda(m) = \sum_{i=1}^{m} \lambda_i$ and $k = 0, 1, 2, \ldots, N-1$ is the largest index for which $-\Lambda(k) < h_\mu - h_a$.

C. Interpreting Ambiguity Rate

Up to this point, we motivated ambiguity rate as correcting over counting in the DIFS statistical complexity dimension d_μ. It is worth discussing the quantity in more depth.

On the one hand, note that when $h_\mu - h_a = 0$, the causal-state process is stationary and C_μ time-independent: $\Delta H[\mathcal{R}_t] = 0$. This occurs for finite-state DIFSs, as well as many with countably-infinite states; see Section VII A. When this occurs, applying Eq. (17) returns a vanishing statistical complexity dimension $d_\mu = 0$, as expected.

On the other hand, when ambiguity rate vanishes, C_μ grows at the Shannon entropy rate: $\Delta H[\mathcal{R}_t] = h_\mu$. This occurs when there are no identical maps, no overlap, and no noninvertibility in the mapping functions. In short, $h_a = 0$ when the process is "perfectly self-similar" and every new observed symbol produces a new, distinct state.

With this in mind, we can use the ambiguity rate, and specifically $h_\mu - h_a$, to describe the stationarity of the model's internal state process. The state set is time independent. When $h_a > 0$, however, to optimally predict the process $\mathcal{P}$ requires a nonstationary model (temporally-growing state set $\boldsymbol{\mathcal{R}}$), even though $\mathcal{P}$ is itself stationary. This is a consequence of modeling "out of class". That is, predicting a perfectly self-similar $\mathcal{P}$ requires differentiating every possible infinite past. This is only possible with a DIFS by storing new states at the rate new pasts are being created. (Moving to a more powerful model class by, say, imbuing our states with counters or stacks, may make it possible to model $\mathcal{P}$ with a stationary model.)

This perspective naturally leads to another that probes the efficacy of the causal-state mapping. Considering the space of all possible infinite pasts $\overleftarrow{X}$, the causal-state mapping $f_\epsilon(\overleftarrow{X}) \rightarrowtail \boldsymbol{\mathcal{R}}$ is defined such that:

$$f_\epsilon(\overleftarrow{X} = \overleftarrow{x}) = f_\epsilon(\overleftarrow{X} = \overleftarrow{x'}) = \eta_i \ ,$$

if $\Pr(X_{0:\ell}|\overleftarrow{X} = \overleftarrow{x}) = \Pr(X_{0:\ell}|\overleftarrow{X} = \overleftarrow{x'})$ for all $\ell \in \mathbf{N}^+$. When the process is perfectly self-similar, the causal-state mapping is one-to-one and $h_a = 0$. In this case, storing the causal states is no better for prediction than simply tracking the space of all pasts. (Although the causal-state set $\boldsymbol{\mathcal{R}}$ is still informative in characterizing how we might approximate the process with a finite state machine [37].) The number of pasts each state "contains" is stationary and given by $2^{h_a} = 1$.

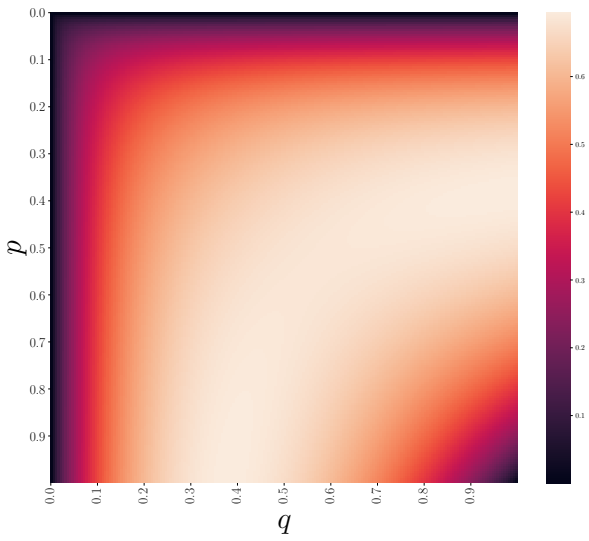


FIG. 7. The entropy rate h_μ, which in this case is equivalent to the ambiguity rate h_a, is plotted for the DIFS depicted in Fig. 3 for $p, q, \in (0, 1)$.

In general, for a stationary process $\mathcal{P}$, the average number of pasts contained by a given causal state grows at the rate 2^{h_a}. When the process has a stationary state set, the number of pasts each state contains must necessarily grow at the rate new pasts are being generated, and so $2^{h_\mu} = 2^{h_a}$.

Finally, let's close with a short historical perspective. The development of d_μ was partially inspired by Shannon's definition in the 1940s of *dimension rate* [5]:

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

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.

Therefore, it appears the first modern discussion of a dimensional quantity of this nature for stochastic processes motivated the development using resource theory [37], noting that the d_1 of the causal-state set Eq. (9) characterizes the *distortion rate* when coarse-graining an uncountably-infinite state set. Starting from the dimen-

sional quantity, the relationship to statistical complexity was then forged.

In this light, developing ambiguity rate and calling out its easy mathematical connection to $\Delta H[\boldsymbol{\mathcal{R}}_t]$ flips this motivation. The quantity $h_\mu - h_a$ can be defined purely in terms of $\mathcal{P}$ and has an intuitive relationship to the causal-state mapping. The dimensional quantity d_μ naturally falls out when we compare this rate of model-state growth to the dynamics of the causal states in the mixed-state simplex. Therefore, we may motivate d_μ as not only a resource-theoretic tool for coarse-graining infinitely-complex state machines, but also as an intrinsic measure of a process' structural complexity.

VII. EXAMPLES

We now consider two examples. The first is a parametrized discrete-time renewal process that has a countably-infinite state space for all parameters. This allows us to explicitly write down the Blackwell measure and calculate ambiguity rate exactly using Eq. (15). The second is a parametrized machine with three maps, which has an uncountably-infinite state space for nearly all parameters. Calculating the ambiguity rate in this case requires us to approximate the Blackwell measure using Ulam's method.

A. Example: Discrete-Time Renewal Process

The DIFS depicted in Fig. 3 has the alphabet $\mathcal{A} = \{\triangle, \square\}$ and the substochastic matrices:

$$T^{(\triangle)} = \begin{pmatrix} 1-q & q \\ 0 & 1-p \end{pmatrix} \text{ and } T^{(\square)} = \begin{pmatrix} 0 & 0 \\ p & 0 \end{pmatrix} \ , \qquad (18)$$

where $p = q = \frac{1}{2}$. Recall that the states η will take the form of length-two vectors such that $\langle \eta \, | \mathbf{1} \rangle = 1$, so $f^{(x)}$ and $p^{(x)}$ will depend only on a single variable—which we will take to be the first component of η. We write this component $\langle \eta | \delta_1 \rangle$ to avoid confusion with temporal indexing.

For the general case where $p, q \in (0, 1)$ are left unspecified, we have the probability function set:

$$\begin{aligned} p^{(\triangle)}(\eta) &= 1 - p + \langle \eta | \delta_1 \rangle p \ , \\ p^{(\square)}(\eta) &= p - \langle \eta | \delta_1 \rangle p \ , \end{aligned}$$

and the mapping function set:

$$\begin{aligned} f^{(\triangle)}(\eta) &= \left(\frac{\langle \eta | \delta_1 \rangle (1-q)}{1 - p + \langle \eta | \delta_1 \rangle p}, \frac{\langle \eta | \delta_1 \rangle (p+q-1) + 1 - p}{1 - p + \langle \eta | \delta_1 \rangle p} \right) \ , \\ f^{(\square)}(\eta) &= (1, 0) \ . \end{aligned}$$

Due to the temporal indexing, this may appear complicated, but note that the denominator of $f^{(\triangle)}$ is simply $p^{(\triangle)}$. When $p = q = 1/2$, the functions reduce to:

$$\begin{aligned} p^{(\triangle)}(\eta) &= \frac{1}{2} \left(1 + \langle \eta | \delta_1 \rangle \right) \ , \\ p^{(\square)}(\eta) &= \frac{1}{2} \left(1 - \langle \eta | \delta_1 \rangle \right) \ , \end{aligned}$$

and

$$\begin{aligned} f^{(\triangle)}(\eta) &= \left(\frac{\langle \eta | \delta_1 \rangle}{1 + \langle \eta | \delta_1 \rangle}, \frac{1}{1 + \langle \eta | \delta_1 \rangle} \right) \ , \\ f^{(\square)}(\eta) &= (1, 0) \ . \end{aligned}$$

It is simple to confirm that the mixed state set is countable:

$$\boldsymbol{\mathcal{R}} = \left\{ \left(\frac{1}{n+1}, \frac{n}{n+1} \right) : n = \mathbb{Z}^{0+} \right\} \ ,$$

where $\eta_n = \left(\frac{1}{n+1}, \frac{n}{n+1} \right)$ is the state induced after observing $\triangle$'s since the last $\square$. Compare this to Fig. 3. From here, we could find the transition probabilities $p^{(x)}(\eta_n)$ and compute the Blackwell measure, but let us do so in the general case.

When $p \neq q$, $\boldsymbol{\mathcal{R}}$ becomes:

$$\eta_n = \left[\frac{(p-q)(1-q)^n}{p(1-q)^n - q(1-p)^n}, \frac{q(1-q)^n - q(1-p)^n}{p(1-q)^n - q(1-p)^n} \right] \ .$$

This simple structure allows us to give the Blackwell measure explicitly:

$$\mu_B(n) = \frac{p(1-q)^n - q(1-p)^n}{p - q} \times \frac{pq}{p+q} \ ,$$

where $\mu_B(n)$ is the asymptotic invariant measure over the state induced after seeing n $\triangle$s since the last $\square$.

With the Blackwell measure in hand, the entropy rate can be explicitly calculated as the infinite sum:

$$\begin{aligned} h_\mu &= \sum_{n=1}^{\infty} \mu_n H[X_n | \mathcal{R}_n = \eta_n] \\ &= -\sum_{n=1}^{\infty} \mu_n \Big(p^{(\triangle)}(\eta_n) \log_2 p^{(\triangle)}(\eta_n) \\ &\qquad + p^{(\square)}(\eta_n) \log_2 p^{(\square)}(\eta_n) \Big) \ . \end{aligned}$$

Figure 7 plots h_μ for $p, q \in (0, 1)$. In calculating h_μ, there is a contribution from every state except the first—η_0—since the first state transitions to the second with probability one and there is no branching uncertainty. Every other state transitions on a coin flip of a determined bias between $(\triangle, \square)$, generating uncertainty with each transition.

In contrast to how h_μ averages over all mixed states, ambiguity rate accumulates in only one state—η_0. From Fig. 3, we see that $H[x_n, \eta_n | \eta_{n+1}] = 0$ for all n other

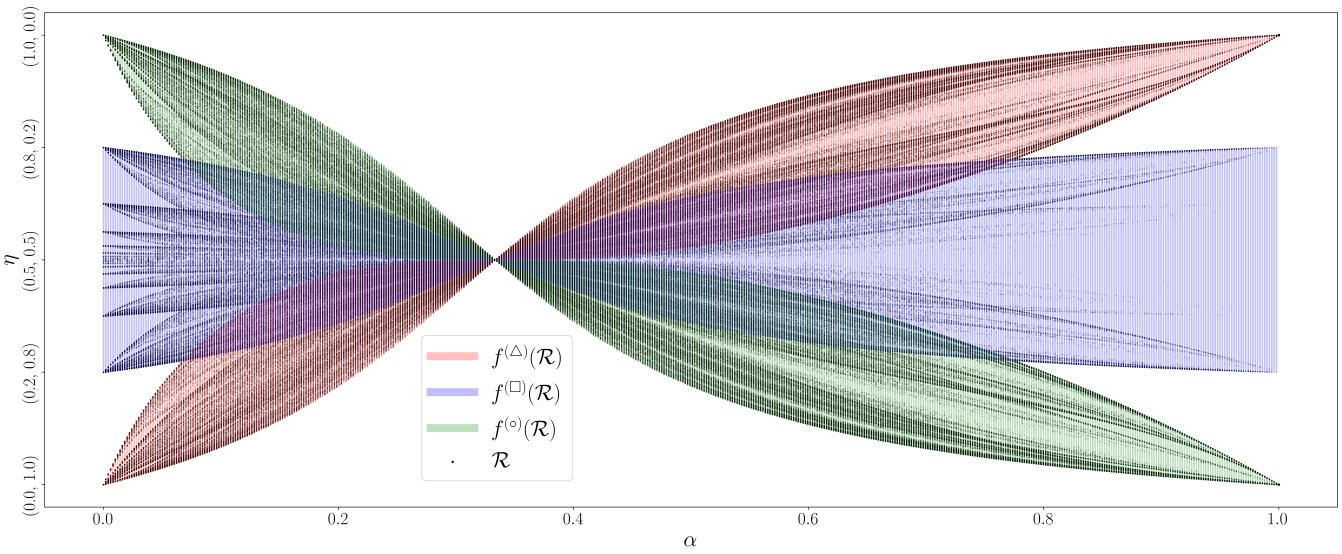


FIG. 8. One-dimensional attractor for DIFS given in Eq. (19) with $x = 0.25$ and horizontally varying $\alpha \in (0, 1)$. State set $\eta \in \boldsymbol{\mathcal{R}}$ plotted (red, blue green) on top of the images of the mapping functions $\left\{f^{(\triangle)}, f^{(\square)}, f^{(\circ)}\right\}$ applied to $\boldsymbol{\mathcal{R}}$. For $\alpha \in (0.07, 0.78)$, there is overlap in the images of the maps.

than $n = 0$. That is, each state η_n is only accessed via the prior state η_{n-1}, except for η_0, which may be accessed from every other state. So, ambiguity in the past can only be introduced by visiting η_0. Since these transitions only occur on a $\square$, we must find the probability distribution $\Pr(X_0 = \square, \mathcal{R}_0 = \eta_n | \mathcal{R}_1 = \eta_0)$.

Applying Eq. (15) and Eq. (16), we explicitly write down the ambiguity rate as:

$$h_a = -\mu_0 \sum_{n=1}^{\infty} \left(\frac{\mu_n}{\mu_0} p^{(\square)}(\eta_n)\right) \log_2 \left(\frac{\mu_n}{\mu_0} p^{(\square)}(\eta_n)\right) \ .$$

Both h_μ and h_a are infinite summations, but when calculating the ambiguity rate, the sum refers to calculating a single Shannon entropy over the infinite, discrete distribution representing the probability distribution over prior states when arriving in η_0.

Since the state space does not grow—$\Delta H[\boldsymbol{\mathcal{R}}_t] = 0$—the entropy rate $h_\mu = h_a$ as $n \to \infty$. Therefore, d_μ vanishes for all values of p and q. This will always be the case for finite-state DIFSs.

B. Example: 1-D Ambiguity Rate

Now, let's turn to the more general case, those DIFSs with uncountably-infinite state spaces. For the moment we restrict to one-dimensional DIFSs, so that the states lie in the 1-simplex. Consider a DIFS with the alphabet $\mathcal{A} = \{\triangle, \square, \circ\}$ and the associated substochastic matrices:

$$T^{(\triangle)} = \begin{pmatrix} \alpha y & \beta x \\ \alpha x & \beta y \end{pmatrix}, \ T^{(\square)} = \begin{pmatrix} \beta y & \beta x \\ \beta x & \beta y \end{pmatrix}, \text{ and}$$
$$T^{(\circ)} = \begin{pmatrix} \beta y & \alpha x \\ \beta x & \alpha y \end{pmatrix} \ , \tag{19}$$

with $\alpha = 1 - 2\beta$, $\alpha \in (0, 1)$, and $x = 1 - y$, $x \in (0, 1)$.

Figure 8 depicts all of the DIFSs for the slice of the parameter space where $x = 0.25$. The vertical axis is the 1-simplex and each vertical slice plots the state space $\boldsymbol{\mathcal{R}}(\alpha)$ at the appropriate value of α, given on the horizontal axis. Additionally, the images of the functions $\left\{f^{(\triangle)}(\boldsymbol{\mathcal{R}}), f^{(\square)}(\boldsymbol{\mathcal{R}}), f^{(\circ)}(\boldsymbol{\mathcal{R}})\right\}$ are shaded in red, blue, and green, respectively.

At $\alpha = 1/3$, the mixed state set $\boldsymbol{\mathcal{R}}$ contracts to a finite set, and h_μ must equal to h_a, making $d_\mu = 0$. At this point in parameter space, $\boldsymbol{\mathcal{R}}$ consists of only a single state; $\boldsymbol{\mathcal{R}}(\alpha = 1/3) = \{(1/2, 1/2)\}$. At every other value of α, $h_a < h_\mu$. There is overlap in the images of the maps for, approximately, $\alpha \in (0.07, 0.78)$. In this regime, $h_a > 0$.

To calculate the ambiguity rate and therefore the statistical complexity dimension d_μ we use a modified Ulam's method to approximate the Blackwell measure and then approximate the integral equation Eq. (15). This method is not the only way to find the ambiguity rate, but does have several advantages, including speed and the ability to control the accuracy of our approximation. Appendix B discusses the method in depth.

The top plot in Fig. 9 gives the entropy rate, ambiguity rate, and $h_\mu - h_a$ for the DIFSs pictured in Fig. 8. As α is increased, h_a smoothly increases from zero as overlap

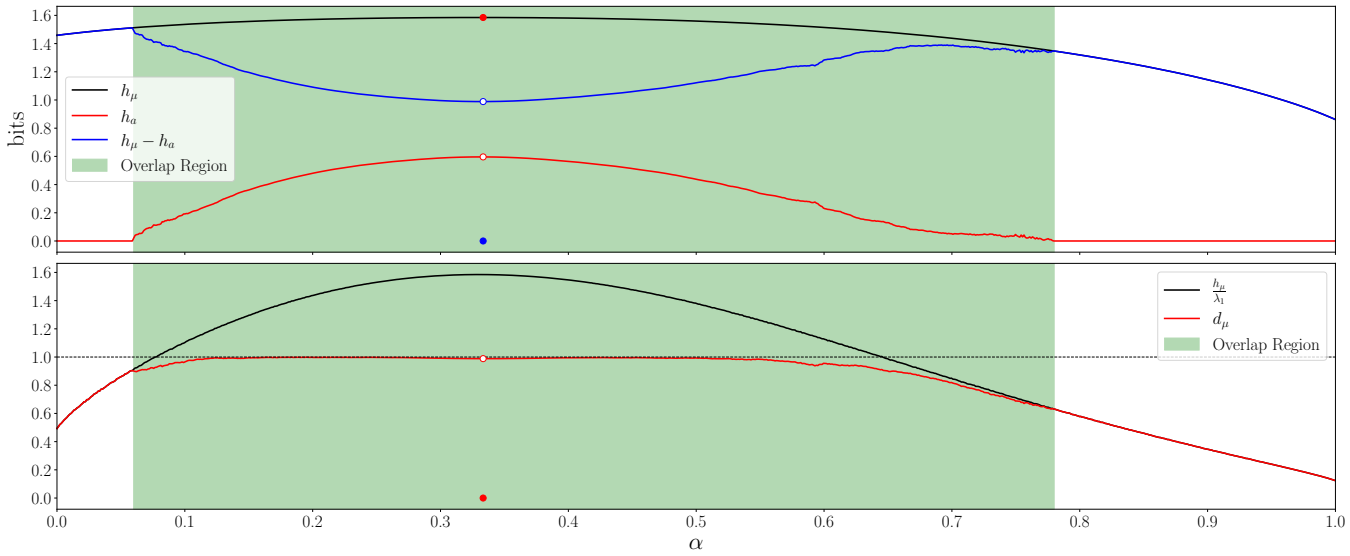


FIG. 9. Entropies and dimensions for the DIFS given in Eq. (19) with $\alpha \in (0, 1)$ and $x = 0.25$: (Top) Entropy rate h_μ, ambiguity rate h_a, and $h_\mu - h_a$. (Bottom) Comparing $d_\Gamma = h_\mu/\lambda_1$ to $d_\mu = (h_\mu - h_a)/\lambda_1$. The latter smoothly departs from the former. It is approximately 1 for much of the overlap region, except where it discontinuously jumps to zero at $\alpha = 1/3$.

begins to occur. It approaches ≈ 0.6 around $\alpha = 1/3$, but is discontinuously equal to h_μ at this point. The reason for this is an instantaneous equality in the fixed points of the mapping functions, causing the state space to collapse. As α increases to 1, h_a smoothly decreases back to zero. The roughness seen in the plot is due to numerical precision, as analyzed in Appendix B.

For a large portion of the overlap region, d_Γ saturates at 1.0 due to its threshold condition on the exponent sum. Figure 9(Bottom) instead plots h_μ/λ_1 to show how this quantity smoothly changes across parameter space, reaching a maximum at around 1.6. In comparison, $d_\mu = (h_\mu - h_a)/\lambda_1$ smoothly departs from the Lyapunov dimension when overlap begins and instead hugs the underside of the dimension 1 line for much of the overlap. Again, at $\alpha = 1/3$ there is the discontinuous drop to $d_\mu = 0$, followed by d_μ smoothly rejoining with the Lyapunov dimension as the overlap region ends.

Unsurprisingly, calculating the ambiguity rate in higher dimensions is more challenging. Although, in principle, Ulam's method still applies and we may in principle follow the algorithm laid out in Appendix B, higher-dimensional mapping functions introduce additional error sources in the approximation. Developing an algorithm to efficiently and accurately calculate the ambiguity rate in higher dimensions is of great interest, but we leave this task to future work.

VIII. CONCLUSION

Stepping back from our development of ambiguity rate and statistical complexity dimension, let's position the new results in the context of the prior two works in this series [11, 12]. In the first, motivated by needing a general solution to the Shannon entropy rate for processes generated by finite-state hidden Markov chains, we showed how an optimal predictor can be constructed for any such process, at the cost of a potentially uncountably-infinite state space. To address the resulting challenge, we introduced hidden Markov-driven iterated function systems and showed that the attractor of a properly defined DIFS is equivalent to the ϵ-machine for the process generated by its substochastic matrices.

The result gave benefits beyond a finite-dimensional description of an infinite-state model. The identification allowed us to adopt several rigorous results on IFSs, including an ergodic theorem that allows us to sample the DIFS to accurately and efficiently calculate the Shannon entropy rate of the underlying process. With this, our original goal was completed.

However, identifying these ϵ-machines as IFSs allowed us to show that the dimension of the mixed-state set, a quantity well studied for IFSs, is a structural complexity measure for stochastic processes. Reference [12] then introduced the statistical complexity dimension—the DIFS attractor information dimension. A longstanding conjecture in dynamical systems theory states that the Lyapunov dimension, a dimensional quantity calculated using a system's Lyapunov spectrum, is equivalent to the information dimension. We showed that for many DIFSs this is indeed the case, connecting the information dimension of the ϵ-machine's state space to the ϵ-machine's statistical complexity dimension—the rate of divergence of the statistical complexity. This related a DIFS's dynamics to the information-theoretic properties of the underlying process. Additionally, it gave a new

and meaningful measure of structural complexity—one that differentiates between stochastic processes with divergent state spaces.

That was not the end of the story, since calculating d_μ is difficult due to longstanding problems in the field of IFS dimension theory. In particular, the overlap problem posed a significant hurdle—restricting the preceding results to only nonoverlapping IFSs. This restricted us to the class of stochastic processes with a one-to-one past-to-causal state mapping. In a sense, these processes are the most complex but with structure that is the least interesting. That is, for DIFSs we simply store every past to build an optimally-predictive model.

This state of affairs led directly to the present development and to introducing the ambiguity rate. The latter allows smoothly varying between ϵ-machines with countable state spaces ($h_a = h_\mu$ and $\Delta H[\boldsymbol{\mathcal{R}}] = 0$) and those with perfectly self-similar state spaces ($h_a = 0$ and $\Delta H[\boldsymbol{\mathcal{R}}] = h_\mu$), including the vast majority lying in between, with $h_\mu > h_a > 0$ and $\Delta H[\boldsymbol{\mathcal{R}}] = h_\mu - h_a$. This model class is much more general, generating an exponentially larger family of stochastic processes. As such, we anticipate that this class will be of great interest and likely to lead to significant further progress in analyzing the randomness and structure generated by hidden Markov chains.

To close, we note that the structural tools and the entropy-rate method introduced by this trilogy were put to practical application in two other previous works. One diagnosed the origin of randomness and structural complexity in quantum measurement [38]. The other exactly determined the thermodynamic functioning of Maxwellian information engines [39], when there had been no previous method for this kind of detailed and accurate analysis. The lesson from these applications of finite-state-generated processes is that the resulting effectively-infinite state processes are very likely generic. That said, for now we must leave to the future investigating infinite-state machines and developing the required algorithmic tools.

ACKNOWLEDGMENTS

The authors thank Alec Boyd, Sam Loomis, 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 Grant number FQXi-RFP-IPW-1902, and U.S. Army Research Laboratory and the U.S. Army Research Office under grants W911NF-18-1-0028 and W911NF-21-1-0048.

[1] J. P. Crutchfield. Between order and chaos. *Nature Physics*, 8(January):17–24, 2012. 1, 2, 4

[2] A. Turing. On computable numbers, with an application to the Entschiedungsproblem. *Proc. Lond. Math. Soc.*, 42, 43:230–265, 544–546, 1937. 1

[3] 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.

[4] M. Minsky. *Computation: Finite and Infinite Machines*. Prentice-Hall, Englewood Cliffs, New Jersey, 1967. 1, 2

[5] C. E. Shannon. A mathematical theory of communication. *Bell Sys. Tech. J.*, 27:379–423, 623–656, 1948. 1, 3, 9

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

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

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

[9] 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.

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

[11] A. Jurgens and J. P. Crutchfield. Shannon entropy rate of hidden Markov processes. *J. Stat. Physics*, 183(32):1–18, 2020. 2, 4, 5, 6, 12

[12] A. Jurgens and J. P. Crutchfield. Divergent predictive states: The statistical complexity dimension of stationary, ergodic hidden Markov processes. *Chaos*, in press, 2021. 2, 6, 12

[13] 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. 2

[14] Y. Ephraim and N. Merhav. Hidden Markov processes. *IEEE Trans. Info. Th.*, 48(6):1518–1569, 2002. 2, 3

[15] J. Bechhoefer. Hidden Markov models for stochastic thermodynamics. *New. J. Phys.*, 17:075003, 2015. 2

[16] L. R. Rabiner and B. H. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, January:4–16, 1986. 2

[17] E. Birney. Hidden Markov models in biological sequence analysis. *IBM J. Res. Dev.,*, 45(3.4):449–454, 2001. 2

[18] S. Eddy. What is a hidden Markov model? *Nature Biotech.*, 22:1315–1316, Oct 2004. 2

[19] 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. 2

[20] 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. 2

[21] T. Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, 85:461–464, 2000. 2

[22] J. Sun and E. M. Bollt. Causation entropy identifies indirect influences, dominance of neighbors and anticipatory couplings. *Physica D*, 267:49–57, 2014. 2

[23] R. G. James, N. Barnett, and J. P. Crutchfield. Information flows? A critique of transfer entropies. *Phys. Rev. Lett.*, 116(23):238701, 2016. SFI Working Paper 16-01-001; arxiv.org:1512.06479 [cond-mat.stat-mech]. 2

[24] R. G. James, C. J. Ellison, and J. P. Crutchfield. Anatomy of a bit: Information in a time series observation. *CHAOS*, 21(3):037109, 2011. 2

[25] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, second edition, 2006. 3

[26] J. H. Elton. An ergodic theorem for iterated maps. *Ergod. Th. Dynam. Sys.*, 7:481–488, 1987. 4

[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. 4

[28] A. Renyi. On the dimension and entropy of probability distributions. *Acta Math. Hung.*, 10:193, 1959. 5

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

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

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

[32] Ya. B. Pesin. *Dimension Theory in Dynamical Systems: Contemporary Views and Applications*. University of Chicago Press, 1997. 5

[33] I. Shimada and T. Nagashima. A numerical approach to ergodic problem of dissipative dynamical systems. *Prog. Theo. Phys.*, 61:1605, 1979. 6

[34] 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. 6

[35] 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. 6

[36] B. Barany. On the Ledrappier-Young formula for self-affine measures. *Math. Proc. Cambridge Phil. Soc.*, 159(3):405–432, 2015. 6

[37] S. E. Marzen and J. P. Crutchfield. Nearly maximally predictive features and their dimensions. *Phys. Rev. E*, 95(5):051301(R), 2017. 9

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

[39] 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. 13

Supplementary Materials

Ambiguity Rate of Hidden Markov Processes

Alexandra Jurgens and James P. Crutchfield

arXiv:2002.XXXXX

The Supplementary Materials to follow give a suite of example hidden Markov chains.

Appendix A: Hidden Markov-Driven Iterated Function System Examples

We reproduce here the hidden Markov-driven iterated function systems (DIFS) used to create Fig. 4.

First, the "delta" DIFS, from Fig. 4a, is given by a three symbol alphabet and the substochastic symbol-labeled matrices:

$$T^{(\square)} = \begin{pmatrix} 0.112 & 0.355 & 3.901 \times 10^{-2} \\ 0.434 & 7.685 \times 10^{-2} & 2.333 \times 10^{-2} \\ 0.215 & 2.518 \times 10^{-2} & 0.220 \end{pmatrix}, \tag{S1}$$

$$T^{(\triangle)} = \begin{pmatrix} 1.778 \times 10^{-2} & 0.113 & 0.220 \\ 6.465 \times 10^{-2} & 0.272 & 2.413 \times 10^{-2} \\ 0.400 & 8.697 \times 10^{-3} & 9.892 \times 10^{-3} \end{pmatrix}, \text{ and}$$

$$T^{(\circ)} = \begin{pmatrix} 8.312 \times 10^{-2} & 2.867 \times 10^{-2} & 3.096 \times 10^{-2} \\ 4.690 \times 10^{-2} & 5.625 \times 10^{-2} & 1.807 \times 10^{-3} \\ 0.114 & 1.095 \times 10^{-3} & 7.522 \times 10^{-4} \end{pmatrix},$$

Second, the "Nemo" DIFS, from Fig. 4b, is given by a two symbol alphabet and the substochastic symbol-labeled matrices:

$$T^{(\square)} = \begin{pmatrix} 0.409 & 0.0 & 0.091 \\ 0.5 & 0.0 & 0.0 \\ 0.0 & 0.182 & 0.0 \end{pmatrix}, \text{ and} \tag{S2}$$

$$T^{(\triangle)} = \begin{pmatrix} 0.091 & 0.0 & 0.409 \\ 0.5 & 0.0 & 0.0 \\ 0.0 & 0.818 & 0.0 \end{pmatrix},$$

Finally, the "gamma" DIFS, from Fig. 4c, is given by a three symbol alphabet and the substochastic symbol-labeled matrices::

$$T^{(\square)} = \begin{pmatrix} 2.479 \times 10^{-2} & 0.355 & 1.745 \times 10^{-2} \\ 0.410 & 1.878 \times 10^{-2} & 2.388 \times 10^{-4} \\ 0.204 & 2.472 \times 10^{-3} & 0.215 \end{pmatrix}, \tag{S3}$$

$$T^{(\triangle)} = \begin{pmatrix} 1.672 \times 10^{-3} & 0.133 & 0.235 \\ 3.377 \times 10^{-2} & 0.272 & 8.277 \times 10^{-2} \\ 0.426 & 1.498 \times 10^{-2} & 4.286 \times 10^{-3} \end{pmatrix}, \text{ and}$$

$$T^{(\circ)} = \begin{pmatrix} 8.870 \times 10^{-2} & 3.059 \times 10^{-2} & 0.114 \\ 6.918 \times 10^{-2} & 0.112 & 1.804 \times 10^{-3} \\ 0.131 & 1.165 \times 10^{-3} & 8.005 \times 10^{-4} \end{pmatrix},$$

Due to finite numerical accuracy, reproducing the attractors using these specifications may differ slightly from Fig. 4.

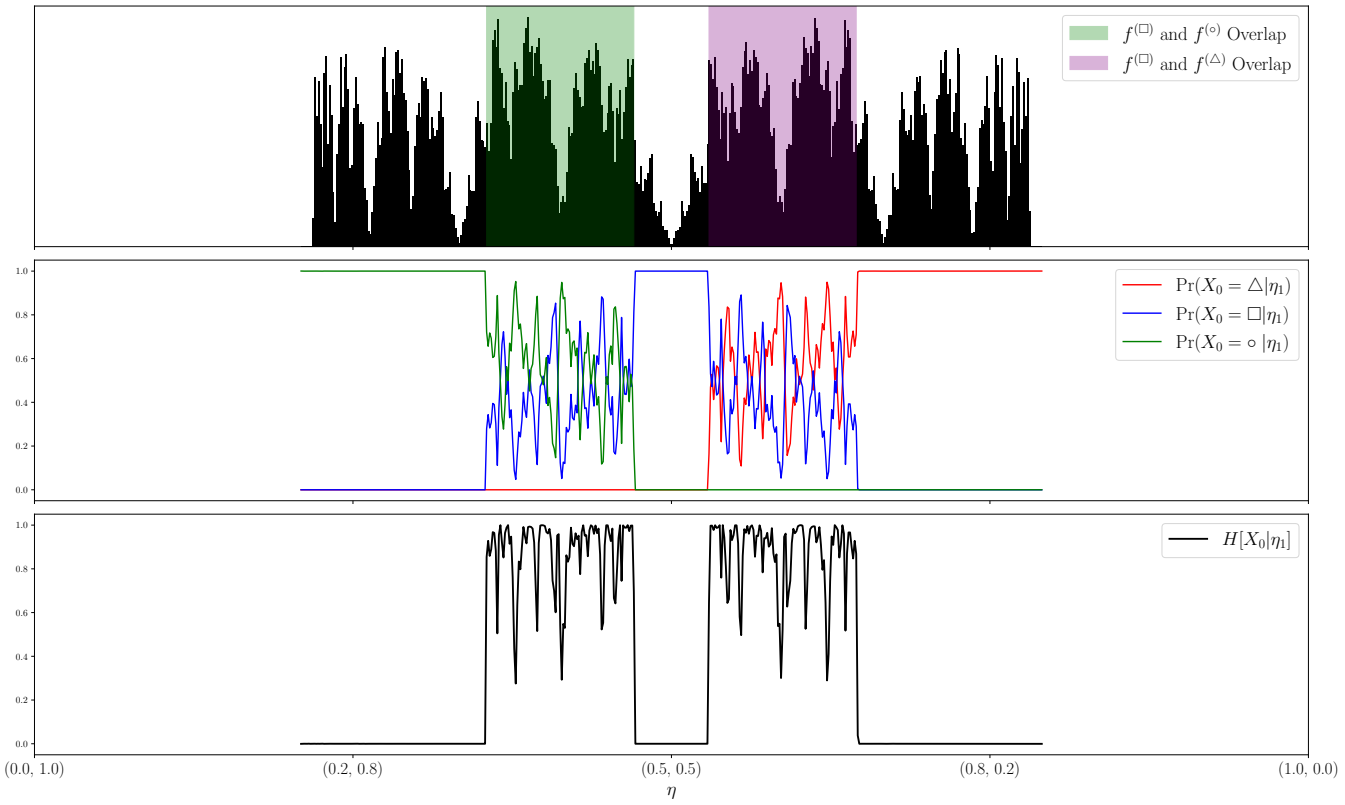


FIG. S1. DIFS for $x = 0.25$ and $\alpha = 0.5$: (Top) Blackwell Measure μ_B approximated by Ulam's method with $k = 400$. The two overlapping regions are overlaid and may be compared with Fig. 8. (Middle) Probability of each prior map is plotted. In nonoverlapping regions, only one prior map is possible. In the overlapping regions, there are complicated, fractal-like distributions over multiple prior maps. (Bottom) Shannon entropy over the prior map: Nonzero only in the overlapping regions.

The mapping images shown in Fig. 5 are produced by the following three symbol DIFS:

$$T^{(\square)} = \begin{pmatrix} \alpha y & \beta x & \beta x \\ \alpha x & \beta y & \beta x \\ \alpha x & \beta x & \beta y \end{pmatrix}, \; T^{(\triangle)} = \begin{pmatrix} \beta y & \alpha x & \beta x \\ \beta x & \alpha y & \beta x \\ \beta x & \alpha x & \beta y \end{pmatrix}, \text{ and}$$
$$T^{(\circ)} = \begin{pmatrix} \beta y & \beta x & \alpha x \\ \beta x & \beta y & \alpha x \\ \beta x & \beta x & \alpha y \end{pmatrix}, \tag{S4}$$

with $\alpha = 0.63, x = 0.2$ for the overlapping example in Fig. 5a and $\alpha = 0.6, x = 0.15$ for the nonoverlapping example in Fig. 5b.

Appendix B: Numerical Approximation of Ambiguity Rate

To numerically approximate the ambiguity rate for a DIFSlying in the 1-simplex, we may use Ulam's method to approximate the Blackwell measure, then compute Eq. (15). Given a partition $\{A_1, \ldots A_k\}$ of the simplex, define:

$$P_{ij}^{(x)} = \frac{m(f^{(x)}(A_i) \cap A_j)}{m(f^{(x)}(A_i))} \times p^{(x)}\left(\overline{A_i}\right) \; ,$$

where m is the Lebesgue measure over Δ and $\overline{A_i}$ is the center of a partition element. Let $P = \sum P^{(x)}$ and find the left eigenvalue $p = pP$. Then, the invariant-measure approximation is:

$$\mu_n(A) = \sum_i p_i \frac{m(A \cap A_i)}{m(A_i)} \ .$$

For this example, let's walk through estimating the ambiguity rate for one DIFS–setting $x = 0.25$ and $\alpha = 0.5$. The partition $\{A_1, \ldots A_k\}$ is created by dividing the 1-simplex into k boxes of equal length. The approximated Blackwell measure $\widehat{\mu_B}$ for the DIFS, using $k = 400$, is shown in the top plot of Fig. S1. The overlay indicates the region (green) of the state space that exhibits overlap. Compare the two regions depicted in Fig. S1 to the overlap shown in Fig. 8, for the vertical slice at $\alpha = 0.5$.

Note that the partition may be defined as desired. We have found that defining the partition by calculating the set of fixed points of the mapping functions $\left\{p^x : f^{(x)}(p^x) = p^x\right\}$. Then, as many times as is desired, find all possible iterates of each fixed point, constructing a new set $\left\{f^{(w)}(p^x) : x \in \mathcal{A}, w \in \bigcup_{n=0}^{N} \mathcal{A}^n\right\}$, where $N \in \mathbb{Z}^+$. Removing duplicates and ordering the set gives a list of endpoints for a partition of the 1-simplex. Increasing N produces increasingly fine partitions. This method of defining partitions has advantages when calculating h_a across parameter space as we have in Section VII B, since the position of the fixed point iterates in the simplex are smooth functions of α.

Regardless, once the partition is selected and $\widehat{\mu_B}$ is determined, we again use the partition. For each cell A_i, we find the probability distribution over the maps that could have transitioned into A_i: by applying Eq. (16) and assuming invertibility of the mapping functions:

$$\Pr(X_0 = x | \mathcal{R} \in A_i) = \frac{\widehat{\mu_B}\left(\left(f^{(x)}\right)^{-1}(A_i)\right)}{\widehat{\mu_B}(A_i)} \times p^x\left(\overline{\left(f^{(x)}\right)^{-1}(A_i)}\right) \ .$$

For our example DIFS, the probability of the previous map given current location in the simplex is plotted in the middle figure of Fig. S1. For parts of the simplex outside the overlapping regions, only one prior map is possible and it has probability one. Within the overlapping regions, the distribution over the possible prior maps may be very complicated. The Shannon entropy over the prior map distribution $H\left[X_0 = x | \mathcal{R} \in A_i\right]$ is shown in the third plot of Fig. S1. Once these entropies are calculated, the final step is to approximate the integral equation Eq. (15) with a summation over cells in the partition:

$$h_a = \sum_i \widehat{\mu_B}(A_i) \sum_{x \in \mathcal{A}} H\left[X_0 = x | \mathcal{R} \in A_i\right] \ .$$

In our example, the ambiguity rate is found to be $h_a = 0.4499$. Since the DIFS entropy rate is $h_\mu = 1.5596$, this gives an adjusted state space expansion rate of $h_\mu - h_a = 1.1098$. Calculating the DIFS's Γ and applying Eq. (17) results in a statistical complexity dimension of $d_\mu = 0.9815$.

The advantage of Ulam's method is its relative ease and speed. Additionally, it is deterministic given the partition. We may may increase the accuracy of our approximation simply by tuning our partition, although increasingly fine partitions increase computation time. Additionally, when the set becomes highly rarefied, noisiness will be observed in the calculation of h_a. This can be seen in our example DIFS on either end of the overlap region, although it is worst when $\alpha \in (0.6, 0.78)$. This may be understood when comparing Fig. 8 to Fig. 9— from $\alpha \in (0.6, 0.78)$ are bands of high density in the overlapping region that increase in probability as the overlapping region itself is shrinking. Calculating the h_a accurately in this region requires increasingly fine partitioning. An immediate improvement may be made by adapting the method to use adaptive partitioning as it sweeps parameter space, taking into account the structure of the state set. The method may be applied to any DIFS in the 1-simplex with overlaps.