Way More Than the Sum of Their Parts: From Statistical to Structural Mixtures

James P. Crutchfield*

Complexity Sciences Center and Physics and Astronomy Department
University of California at Davis,
One Shields Avenue, Davis, CA 95616
(Dated: November 24, 2025)

We show that mixtures comprised of multicomponent systems typically are much more structurally complex than the sum of their parts; sometimes, infinitely more complex. We contrast this with the more familiar notion of statistical mixtures, demonstrating how statistical mixtures miss key aspects of emergent hierarchical organization. This leads us to identify a new kind of structural complexity inherent in multicomponent systems and to draw out broad consequences for system ergodicity.

CONTENTS

* chaos@ucdavis.edu

I. INTRODUCTION

Multicomponent systems typically are much more structurally complex than the collection of their parts; even infinitely more so. This should be contrasted with statistical mixtures—such as arise in the Gibbs Paradox of thermodynamics [1, Secs. 2-3] where gases of distinct molecular species exhibit only a modest entropy increase upon formation due to the uncertainty in which species one has in hand. This contrast demonstrates how the ansatz of statistical mixtures misses key aspects of hierarchical organization. The result, as we show, is an awareness of a new kind of structural complexity of composite systems.

The development here focuses on the theoretical core of this basic phenomenon, arguing that it is, in fact, quite commonplace. To appreciate this, it will be helpful to address the motivating issues upfront.

The multicomponent systems of interest are found in several different domains, including the entropy of mixing in thermodynamics [2, 3], the change point problem in statistics [4], the attractor-basin portrait of a dynamical system [5], Smale's basic sets [6, 7],

spatially extended systems with multiple local attractors [8], chaotic crystallography [9, 10], evolutionary dynamics [11], and adaptive and learning systems with memory.

We introduce the concept of hidden multistationary processes to capture what is common across these domains—a system comprising multiple locally-competing behaviors and structures. The basic idea can be appreciated within an experimental paradigm: multistationarity models repeated experimental trials in which different initial conditions lead to statistically distinct behaviors.

In short, one goal is to provide a tractable model that quantitatively captures what is common among these domains, while providing an architectural, high-level view of the state-space organization of behaviors. In particular, we would like to analyze how unpredictable and how structurally complex hidden multistationary processes are, when given their components, whose unpredictability and complexity we know. Another goal is that the approach be constructive, allowing one to quantitatively determine essential properties and to determine precisely what gives rise to the emergent global complexity.

The development proceeds as follows. It first reviews statistical mixtures, briefly recalling stochastic processes, information theory, structural complexity, and mixed state processes. It then introduces the theory and construction of hidden multistationary processes. This includes a canonical minimal representation of hidden multistationary processes and a method to analyze their ergodic decompositions that determines how the latter affect information measures.

The sections following this explore a number of examples going from the simplest cases and familiar structured stationary component processes to the *Mother of All Processes* that subsumes them all. Taken altogether, these illustrate a new kind of structural hierarchy and make plain how infinite complexity naturally emerges. The development concludes drawing out parallels with related results and consequences in nonequilibrium thermodynamics and machine learning.

II. BACKGROUND

To get started, we give a minimal summary of the required background—a summary that assumes familiarity with computational mechanics [12, 13]

and with information theory for complex systems [14, 15].

A. Processes

A process, denoted $\mathcal{P}$, is specified by the joint distribution $\mathbb{P}(\overleftarrow{X}_t, \overrightarrow{X}_t)$ over its chain of random variables $\ldots X_{-1} X_0 X_1 \ldots$. We view $\mathcal{P}$ as a *communication channel* with a fixed input distribution $\mathbb{P}(\overleftarrow{X}_t)$: It transmits information from the *past* $\overleftarrow{X}_t = \ldots X_{t-3} X_{t-2} X_{t-1}$ to the *future* $\overrightarrow{X}_t = X_t X_{t+1} X_{t+2} \ldots$ by storing it in the present. X_t denotes the discrete random variable at time t taking on values x from a discrete alphabet $\mathcal{A}$. And, $X_t^\ell = X_t X_{t+1} \ldots X_{t+\ell-1}$ is the block of ℓ random variables starting at time t. A particular realization is denoted using lowercase: $X_t^\ell = x_t^\ell \in \mathcal{A}^\ell$. Often, we simply refer to a particular sequence $w = x_0 x_1 \ldots x_{\ell-1}$, $x_i \in \mathcal{A}$, as a *word*. If we have a symbol x and a word w, we form a new word by concatenation: e.g., wx or xw.

B. Information

Given a process, we form the block distributions $\{\mathbb{P}(X_t^\ell) : \text{for all } t \text{ and } \ell\}$ by marginalizing the given joint distribution:

$$\mathbb{P}(X_t^\ell) = \sum_{\{\overleftarrow{x}_t, \overrightarrow{x}_{t+\ell}\}} \mathbb{P}(\overleftarrow{x}_t, \overrightarrow{x}_{t+\ell}) \ .$$

(We ignore here the measure-theoretic construction of cylinder sets and their measures; for background see Ref. [16] and references therein.) A stationary process is one for which $\mathbb{P}(X_t^\ell) = \mathbb{P}(X_0^\ell)$ for all t and ℓ. For a stationary process, we drop the time index and thereby have the family of *word distributions* $\mathbb{P}(X^\ell)$ that completely characterizes the process.

The amount of Shannon information in words is measured by the *block entropy*:

$$H(\ell) = H[\mathbb{P}(X^\ell)] \ ,$$

where $H[\mathbb{P}(Y)] = -\sum_{\{y\}} \mathbb{P}(y) \log_2 \mathbb{P}(y)$ is the Shannon entropy of the random variable Y. A process' information production is given by its *entropy rate*:

$$h_\mu = \lim_{\ell \to \infty} \frac{H(\ell)}{\ell} \ .$$

It is often used to measure a process' degree of un-

predictability.

At a minimum, a good predictor—denote its random variables $\widehat{R}$—must capture *all* of a process' *excess entropy* $\mathbf{E}$ [15]—all of the information shared between past and future: $\mathbf{E} = I[\overleftarrow{X}; \overrightarrow{X}]$. Here, $I[Y; Z]$ is the mutual information between variables Y and Z. That is, for a good predictor $\widehat{R}$: $\mathbf{E} = I[\widehat{R}; \overrightarrow{X}]$.

These quantities are closely related. In particular, for finitary processes, those with $\mathbf{E} < \infty$, the block entropy has the linear asymptotic behavior:

$$H(\ell) \propto_{\ell \to \infty} \mathbf{E} + h_\mu \ell .$$

More precisely:

$$\mathbf{E} = \lim_{\ell \to \infty} [H(\ell) - h_\mu \ell] .$$

This shows that $\mathbf{E}$ controls the convergence of entropy rate estimates $h_\mu(\ell) = H(\ell) - H(\ell - 1)$. In fact, for one-dimensional processes, $\mathbf{E}$ can also be defined in terms of entropy convergence:

$$\mathbf{E} = \sum_{\ell=1}^{\infty} [h_\mu(\ell) - h_\mu] . \qquad (1)$$

An analogous quantity that controls the block entropy convergence to the linear asymptote is the *transient information*:

$$\mathbf{T} = \sum_{L=0}^{\infty} [\mathbf{E} + h_\mu L - H(L)] .$$

$\mathbf{T}$ measures the average amount of information an observer must extract in order to know a process' internal state. (For a review of these and related informations see Ref. [15].)

C. Structure

We refer to a model of a process—a particular choice of $\widehat{R}$—as a *presentation*. Note that building a model of a process is more demanding than developing a prediction scheme, since one wishes to go beyond sequence statistics to express a process' mechanisms and internal organization.

To do this, we first recall that a process' communication channel is determined by the conditional distributions $\mathbb{P}(\overrightarrow{X}_t | \overleftarrow{X}_t)$. Based on this, computational mechanics introduced an equivalence relation $\overleftarrow{x} \sim_\epsilon \overleftarrow{x}'$ that groups all of a process' histories which

give rise to the same prediction. The result is a map $\epsilon : \overleftarrow{X} \to \mathcal{S}$ from pasts to *causal states* defined by:

$$\epsilon(\overleftarrow{x}) = \{\overleftarrow{x}' : \mathbb{P}(\overrightarrow{X} | \overleftarrow{x}) = \mathbb{P}(\overrightarrow{X} | \overleftarrow{x}')\} . \qquad (2)$$

In other words, a process' causal states are equivalence classes—$\mathcal{S} = \mathbb{P}(\overleftarrow{X}, \overrightarrow{X})/\sim_\epsilon$—that partition the space $\overleftarrow{X}$ of pasts into sets which are predictively equivalent. With the causal states in hand, one determines the causal-state to causal-state transitions:

$$\{T_{\sigma,\sigma'}^{(x)} : x \in \mathcal{A}, \sigma, \sigma' \in \mathcal{S}\} .$$

The resulting model M, consisting of the causal states and transitions, is called the process' ϵ-*machine* [17]:

$$M(\mathcal{P}) \equiv \left\{ \mathcal{S}, \{T^{(x)}, x \in \mathcal{A}\} \right\} .$$

Informally, a process is *ergodic* if its statistics can be estimated from a single realization that is sufficiently long. If $\mathcal{P}$ is ergodic, then $M(\mathcal{P})$'s recurrent causal states are strongly connected and their asymptotic invariant distribution $\pi = \mathbb{P}(\mathcal{S})$ is unique and given by $\pi = \pi T$, where $T = \sum_{x \in \mathcal{A}} T^{(x)}$.

As described, an ϵ-machine is obtained from a process, but one can also simply define an ϵ-machine and consider its generated process. We will use both notions in the following, as they are equivalent [18]. But why should one use the ϵ-machine presentation of a process in the first place?

To summarize, out of all optimally predictive models $\widehat{R}$ resulting from a partition of the past—those such that $\mathbf{E} = I[\widehat{R}; \overrightarrow{X}]$—the ϵ-machine captures the amount of information that a process stores—the *statistical complexity* $C_\mu \equiv H[\mathbb{P}(\mathcal{S})]$. The excess entropy $\mathbf{E}$—the information explicitly observed in sequences—is only a lower bound on the information C_μ that a process stores [17]: $\mathbf{E} \leq C_\mu$. The difference $\chi = C_\mu - \mathbf{E}$, called the *crypticity*, measures how the process hides its internal state information from an observer [19].

A process' ϵ-machine is its minimal unifilar presentation. It is unique for the process. Moreover, it allows a number of the process' complexity measures to be directly and efficiently calculated [20]. The latter include the process' entropy rate, excess entropy, statistical complexity, and crypticity. In short, a process' ϵ-machine captures all of its informational and structural properties.

III.   MIXED STATE OPERATOR

Given an ϵ-machine M, its causal states can be treated as a standard basis $\{\mathbf{e}_j\}$ in a vector space. Then, any distribution $\mu = \mathbb{P}(\mathcal{S})$ over the states is a linear combination: $\mu = \sum_j c_j \mathbf{e}_j$. Following Ref. [13], these distributions are called *mixed states*. For an k state ϵ-machine, the mixed-state space is a $k-1$-dimensional simplex Δ^{k-1}, as the distributions $\mu \in \Delta^{k-1}$ are normalized.

Consider a special subset of mixed states. Define $\mu(w)$ as the distribution over M's states induced after observing sequence $w = x_0 \ldots x_{\ell-1}$, M having started with state distribution π:

$$\begin{aligned}
\mu_\pi(w) &\equiv \mathbb{P}(\mathcal{S}_\ell | X_0^\ell = w, \mathcal{S}_0 \sim \pi) \\
&= \frac{\mathbb{P}(X_0^\ell = w, \mathcal{S}_\ell, \mathcal{S}_0 \sim \pi)}{\mathbb{P}(X_0^\ell = w, \mathcal{S}_0 \sim \pi)} \\
&= \frac{\pi T^{(w)}}{\pi T^{(w)} \mathbf{1}} \ ,
\end{aligned} \tag{3}$$

where $\mathbf{1}$ is a column vector of 1s and $T^{(w)} = T^{(x_{\ell-1})} \ldots T^{(x_0)}$. Here, the notation $X \sim P$ serves to indicate that random variable X is governed by distribution P.

The last line gives the mixed-state $\mu_\pi(w)$ directly in terms of the initial state distribution π and M's transition matrices. One interpretation is that $\mu_\pi(w)$ represents an observer's best guess as to the process' causal-state distribution given that it saw word w and knows both the process' ϵ-machine and the initial distribution π.

To determine the set of mixed states allowed by a process, we simply calculate the set $\{\mu_\pi(w)\}$ of distinct $\mu_\pi(w)$ for all words $w \in \mathcal{A}^*$. This is most directly done by enumerating w in lexicographic order: e.g., for a binary alphabet successively choosing $w \in \{\lambda, 0, 1, 00, 01, 10, 11, \ldots\}$. Here, λ is the null word. As we will see, the mixed-state set can be finite or infinite.

If we consider the entire set of mixed states, then we construct a presentation of the process by specifying the transition matrices $\{\mathcal{T}^x : \mu_\pi(w) \to \mu_\pi(wx)\}$:

$$\begin{aligned}
\mathbb{P}(x, \mu_\pi(wx) | \mu_\pi(w)) &\equiv \frac{\mathbb{P}(wx | \mathcal{S}_0 \sim \pi)}{\mathbb{P}(w | \mathcal{S}_0 \sim \pi)} \\
&= \mu_\pi(w) T^{(x)} \mathbf{1} \ .
\end{aligned}$$

Note that many words can induce the same mixed state.

It is useful to define a corresponding operator $\mathcal{U}$ that acts on a machine M, returning its *mixed-state presentation* $\mathcal{U}_\pi(M) = \{\{\mathcal{T}^x\}, \{\mu_\pi(w)\}\}$ under initial distribution π. The examples to follow shortly illustrate how mixed states and $\mathcal{U}_\pi(M)$ are calculated.

IV.   CONSTRUCTING HIDDEN MULTISTATIONARY PROCESSES

Recall that a hidden multistationary nonergodic process is one that evolves, across successive realizations, to statistically distinct long-term behaviors. We now introduce our model of this by giving a construction procedure. This, in effect, defines what we mean by multistationary. We then develop several basic properties and analyze in detail a series of example constructions to illustrate them and their ergodic decompositions.

The main tool used to construct a hidden multistationary process is the mixed-state operator $\mathcal{U}_\pi$. We show that this results in a canonical presentation of a given set of stationary components. This is the multistationary process' ϵ-machine.

Definition 1. *A* hidden multistationary process *(HMSP) is defined by the presentation determined via the following procedure.*

1. *Identify an indexed family of component stationary ergodic processes $\{\mathcal{P}^i\}_{i \in I}$. Each is specified by its ϵ-machine presentation $M^i = M(\mathcal{P}^i)$. The ϵ-machines consist only of their recurrent states $\boldsymbol{\mathcal{S}}^i$ that, due to ergodicity, form a single, strongly connected set.*

2. *Specify the component's mixture distribution π—the probability with which each will be visited (sampled):*

$$\pi^i = \mathbb{P}(M^i) \ .$$

3. *Finally, calculate the mixed-state presentation of the multistationary process:*

$$M = \mathcal{U}_\pi \left(\bigsqcup_{i \in I} M^i \right) \ ,$$

where we take the nonoverlapping set of the measure semi-groups [21] specified by the component ϵ-machines. In this way, M's states and transitions are determined from the com-

ponent ϵ-machines and the mixture distribution π.

M, the result of the construction, determines the transient portion of a nonergodic ϵ-machine. M's recurrent components are essentially the same as those (M^i's) of the original component stationary processes $\mathcal{P}^i$. That is to say, what is new in M is the set of transient causal states.

Note that this construction is a stochastic analog of building recognizers for multiregular formal languages [22].

V. THE MULTISTATIONARY ϵ-MACHINE

With the background and definitions set, we are ready to explore the properties of multistationary nonergodic processes. We first establish the structural properties of their ϵ-machine presentations and then their informational properties via ergodic decompositions of various complexity measures.

Each component $M^i = \left\{ \boldsymbol{\mathcal{S}}^i, \{T_i^{(x)}, x \in \mathcal{A}^i\} \right\}$, considered as generating its own process $\mathcal{P}^i$, has a stationary distribution p^i over its states:

$$p_j^i = \mathbb{P}(\mathcal{S}_j) \ , \ \mathcal{S}_j \in \boldsymbol{\mathcal{S}}^i \ .$$

We will also write this as a vector over the multistationary process' recurrent states, when we have a finite number of components:

$$\pi = \left[\pi^1 \left(p_1^1 \dots p_{j_1}^1 \right) \dots \pi^k \left(p_1^k \dots p_{j_k}^k \right) \right] \ ,$$

where $k = |I|$ and $j_i = |\boldsymbol{\mathcal{S}}^i|$. The stationary state distribution π_{ij} for the multistationary process generated by M is, then:

$$\pi_{ij} = \pi^i \cdot p_j^i \ . \tag{4}$$

Consider the following properties of a multistationary process as just defined.

Lemma 1 (Stationarity). *The state distribution π_{ij} is stationary.*

Proof. *This follows from realizing that the recurrent portion of M's transition matrix is block diagonal. That is, asymptotically the components are independent and, by assumption, the component distributions are invariant.*

Lemma 2 (Unifilarity). *The hidden multistationary process machine M is unifilar.*

Lemma 3 (Minimality). *The hidden multistationary process machine M is minimal.*

Lemma 4 (Uniqueness). *The hidden multistationary process machine M is unique.*

The relevant definitions and proofs of these closely follow those given for ϵ-machines generally; see, for example, Ref. [23]. We leave the proofs for a sequel. This all noted, these remarks constitute a proof of the following claim.

Theorem 1. *The mixed state operator applied to a mixture of (finite, ergodic) ϵ-machines produces an ϵ-machine. That is, the ϵ-machine for the hidden multistationary process generated by:*

$$M = \mathcal{U} \left(\bigsqcup_{i \in I} M^i \right)$$

is an ϵ-machine.

Remark: Constructing HMSPs in this way one could start with other classes of presentation for the ergodic component processes, such as nonunifilar presentations—i.e., generic HMMs. However, the resulting M need not be an ϵ-machine. And, as a consequence, one could not directly calculate from such an M the various complexity measures nor, lacking minimality, draw structural conclusions about its architecture. This is one reason why we choose to specify the component processes using ϵ-machine presentations. Limiting the current construction to ergodic components specified by finite-state ϵ-machine presentations serves to simplify the discussion and highlight our main results.

However, lifting these various restrictions or generalizing the previous properties to address them would be a fruitful effort giving a much broader characterization of the complexity of multistationary processes.

So, from here on out we assume the ergodic components are ϵ-machines and ask what properties hold for the multistationary processes so constructed. We build processes consisting of either a finite number or countably infinite number of components.

VI. ERGODIC DECOMPOSITIONS

Since we are given the component processes $\{\mathcal{P}^i, i \in I\}$, what can we say about the resulting multistationary process generated by M? A first

step develops various kinds of ergodic decomposition that attempt to predict M's properties in terms of its ergodic components' properties. The basic question has a very long history in ergodic and information theories. The reader is referred to the review given in Ref. [24]. Our approach here is, on the one hand, to briefly give a flavor of several ergodic decompositions and, on the other, compensating for that lack of rigor, to analyze in detail a number of concrete examples.

The word distribution $\mathbb{P}(X^\ell)$ for $M = \mathcal{U}\left(\sqcup_{i \in I} M^i\right)$ is given by:

$$\mathbb{P}(X^\ell) = \sum_{i \in I} \pi^i \mathbb{P}(X^\ell | M^i) \ .$$

That is, for word w:

$$\mathbb{P}(w) = \sum_{i \in I} \pi^i \mathbb{P}_i(w) \ ,$$

where $\mathbb{P}_i(w)$ denotes the probability that $\mathcal{P}^i$ generates w.

Quantitatively, the HMSP's block entropy is upper bounded by the component block entropies:

$$\begin{aligned}
H(\ell) &= H\left[\mathbb{P}(X^\ell)\right] \\
&= H\left[\sum_{i \in I} \pi^i \mathbb{P}(X^\ell | M^i)\right] \\
&\leq \sum_{i \in I} \pi^i H\left[\mathbb{P}(X^\ell | M^i)\right] \\
&= \sum_{i \in I} \pi^i H^i(\ell) \ ,
\end{aligned}$$

where the second-to-last step employs Jensen's inequality [14] and $H^i(\ell)$ is component $\mathcal{P}^i$'s block entropy.

A more insightful upper bound, though, is developed by first imagining that the sequences generated by the ergodic components do not overlap—for example, the $\mathcal{P}^i$s have disjoint alphabets $\mathcal{A}^i$. Then we define an indicator function f of the process and an associated random variable θ: $\theta = f(X^\ell) = i$, if $X^\ell \in \mathcal{A}_i^\ell$. We have:

$$\begin{aligned}
H\left[X^\ell\right] &= H\left[X^\ell, f(X^\ell)\right] \\
&= H[\theta] + H\left[X^\ell | \theta\right] \\
&= H[\theta] + \sum_{i \in I} \mathbb{P}(\theta = i) H\left[X^\ell | \theta = i\right] \\
&= H[\pi] + \sum_{i \in I} \pi^i H\left[X^\ell | M^i\right] \\
&= H[\pi] + \sum_{i \in I} \pi^i H^i(\ell) \ .
\end{aligned}$$

In the general setting, however, the sequences generated by distinct components can overlap. This reduces the number of distinct positive-probability words and so, too, the block entropy. In this way, we see that the above equality is only an upper bound on the HMSP's block entropy:

$$H(\ell) \leq H[\pi] + \sum_{i \in I} \pi^i H^i(\ell) \ . \tag{5}$$

This bound highlights the contribution of the *mixture entropy* $H[\pi]$. We return to critique this notion of ergodic decomposition later on. For now, we draw out several useful consequences of this line of reasoning, relying on the bound Eq. (5). Elsewhere we explore tighter informational bounds on decomposition.

From this, we see that an HMSP's entropy rate h_μ is simply that of its ergodic components. Assuming the mixture entropy $H[\pi]$ is finite, we have:

$$\begin{aligned}
h_\mu &= \lim_{\ell \to \infty} \frac{H(\ell)}{\ell} \\
&\leq \lim_{\ell \to \infty} \frac{1}{\ell} \left\{ H[\pi] + \sum_{i \in I} \pi^i H^i(\ell) \right\} \\
&= \sum_{i \in I} \pi^i \lim_{\ell \to \infty} \frac{H^i(\ell)}{\ell} \\
&= \sum_{i \in I} \pi^i h_\mu^i \ ,
\end{aligned}$$

where we have the *component entropy rate* $h_\mu^i = h_\mu(M^i)$. Reference [24] originally established this decomposition.

What is less intuitive, though, are various complexity measures as they apply to HMSPs. As we will see, unlike the entropy rate, which component processes are selected and how they relate to one another play key roles. We first consider the ergodic decomposition for excess entropy, then for the tran-

sient information, and finally that for the statistical complexity.

The excess entropy **E** also has an ergodic decomposition. In this case, we have:

$$\mathbf{E} = \lim_{\ell \to \infty} \left(H(\ell) - h_\mu \ell \right)$$

$$\leq \lim_{\ell \to \infty} \left(H[\pi] + \sum_{i \in I} \pi^i H^i(\ell) - \ell \sum_{i \in I} \pi^i h_\mu^i \right)$$

$$= H[\pi] + \sum_{i \in I} \pi^i \left(\lim_{\ell \to \infty} \left[H^i(\ell) - h_\mu^i \ell \right] \right)$$

$$= H[\pi] + \sum_{i \in I} \pi^i \mathbf{E}^i \;,$$

where $\mathbf{E}^i$ is the excess entropy for ergodic component i. The excess entropy decomposition was explored in Refs. [25, 26].

Combining the entropy rate and excess entropy ergodic decompositions, we see that the block-entropy linear asymptotes—$H^i(\ell) \propto \mathbf{E}^i + h_\mu^i \ell$—have their own decomposition:

$$\mathbf{E} + h_\mu \ell \leq H[\pi] + \sum_{i \in I} \pi^i \mathbf{E}^i + \ell \cdot \sum_{i \in I} \pi^i h_\mu^i$$

$$= H[\pi] + \sum_{i \in I} \pi^i \left(\mathbf{E}^i + h_\mu^i \ell \right) \;.$$

It is a simple additional step to develop the ergodic decomposition for the transient information:

$$\mathbf{T} = \sum_{\ell=0}^{\infty} \left[\mathbf{E} + h_\mu \ell - H(\ell) \right]$$

$$\leq \sum_{\ell=0}^{\infty} \left[H[\pi] + \sum_{i \in I} \pi^i \left(\mathbf{E}^i + h_\mu^i \ell \right) \right.$$

$$\left. + \ell \sum_{i \in I} \pi^i h_\mu^i - H[\pi] - \sum_{i \in I} \pi^i H^i(\ell) \right]$$

$$= \sum_{\ell=0}^{\infty} \sum_{i \in I} \pi^i \left[\mathbf{E}^i + h_\mu^i \ell - H^i(\ell) \right]$$

$$= \sum_{i \in I} \pi^i \mathbf{T}^i \;.$$

Curiously, like the entropy rate decomposition, the mixture entropy $H[\pi]$ does not play a role.

The statistical complexity also has an ergodic de-

composition:

$$C_\mu = - \sum_{\sigma \in \boldsymbol{\mathcal{S}}} \mathbb{P}(\sigma) \log_2 \mathbb{P}(\sigma)$$

$$= - \sum_{i \in I} \sum_{\sigma^i \in \boldsymbol{\mathcal{S}}^i} \mathbb{P}(\sigma^i) \log_2 \mathbb{P}(\sigma^i)$$

$$= - \sum_{i \in I} \sum_{j=0}^{|\boldsymbol{\mathcal{S}}^i|-1} \pi_{ij} \log_2 \pi_{ij}$$

$$= - \sum_{i \in I} \sum_{j=0}^{|\boldsymbol{\mathcal{S}}^i|-1} \pi^i p_j^i \log_2 \pi^i p_j^i$$

$$= - \sum_{i \in I} \pi^i \sum_{j=0}^{|\boldsymbol{\mathcal{S}}^i|-1} p_j^i \left(\log_2 \pi^i + \log_2 p_j^i \right)$$

$$= - \sum_{i \in I} \pi^i \sum_{j=0}^{|\boldsymbol{\mathcal{S}}^i|-1} p_j^i \log_2 \pi^i - \sum_{i \in I} \pi^i \sum_{j=0}^{|\boldsymbol{\mathcal{S}}^i|-1} p_j^i \log_2 p_j^i$$

$$= - \sum_{i \in I} \pi^i \log_2 \pi^i - \sum_{i \in I} \pi^i C_\mu^i$$

$$= H[\pi] + \sum_{i \in I} \pi^i C_\mu^i \;,$$

where C_μ^i are the statistical complexities of the ergodic components. The decomposition for statistical complexity was first noted in Ref. [27]. Note that this decomposition does not rely on assuming an equality as in Eq. (5).

Finally, the multistationary crypticity χ, which measures how a process hides state information from an observer, is also unaffected by the mixture distribution:

$$\chi = C_\mu - \mathbf{E}$$

$$\geq H[\pi] + \sum_{i \in I} \pi^i C_\mu^i - \left(H[\pi] + \sum_{i \in I} \pi^i \mathbf{E}^i \right)$$

$$= \sum_{i \in I} \pi^i \left(C_\mu^i - \mathbf{E}^i \right)$$

$$= \sum_{i \in I} \pi^i \chi^i \;,$$

where χ^i is the crypticity of component M^i. In this, it is similar to the entropy rate and transient information decompositions.

VII. STRUCTURAL DECOMPOSITIONS—BEYOND STATISTICAL

To emphasize, what's notable in these kinds of informational decomposition is that, for nonergodic ϵ-machines, we have, for example:

$$C_\mu > \sum_{i \in I} \pi^i C_\mu^i \;.$$

That is, the global structural complexity C_μ of a multistationary process is strictly greater than that contained in its components $\{C_\mu^i\}$. In short, a multistationary process is *at least the sum of its parts.* Indeed, the above inequality leaves out the entropy of mixing. But this is too facile. As we will see, multistationary processes are much, much more.

We will see below, taking a more structural perspective going beyond the ergodic decompositions, that the transient causal state structure is key to a process' global organization and what sequences of observations reveal. This leads us to call into question the interpretation and use of the preceding kinds of ergodic decomposition.

We now show that the construction procedure can be used to answer a number of different questions about multistationary ergodic processes. Several questions are illustrated via particular examples; others via general constructions. The series of examples is developed incrementally to highlight the methods and particular results, as much in isolation as possible.

We first start with processes built from finite-state ergodic components that lead to a multistationary process that is itself finite-state. Then we analyze the case in which finite components lead to a multistationary process with an infinite number of states. We end with examples built from an infinite number of finitary ergodic processes. In each case, we explore the structure of the resulting multistationary process, its complexity measures, and its ergodic decomposition.

A. Finite Hidden Multistationary Processes

1. A Base Case

A simple but illustrative case is that of two period-1 component processes: all Heads and all Tails, selected with fair probability: $\pi = (1/2, 1/2)$.

The components observed separately have $h_\mu^0 = h_\mu^1 = 0$. But together $H(\ell) = 1$, $\ell \geq 1$. In this way, we see that the HMSP information $H(\ell)$ of the mixture is all mixing entropy $H(\pi)$.

2. Period-1 and Period-2 Process

Define the *Periodic Process* $\mathcal{P} \equiv \mathrm{P}(p)$ that repeats the word $w = 0^{p-1}1$. Let's construct the simplest multistationary process consisting of two such components:

1. Period-1 Process P(1), which has complexity measures $h_\mu^1 = 0$ bits per symbol, $C_\mu^1 = 0$ bits, $\mathbf{E}^1 = 0$ bits, $\mathbf{T}^1 = 0$ bit-symbols, and $\chi^1 = 0$ bits.
2. Period-2 Process P(2), which has complexity measures $h_\mu^1 = 0$ bits per symbol, $C_\mu^1 = 1$ bit, $\mathbf{E}^1 = 1$ bit, $\mathbf{T}^1 = 1$ bit-symbol, and $\chi^1 = 0$ bits.

The Period-1 component has a single recurrent state A and the Period-2, two recurrent states, label them B and C. The second step is to specify the mixture distribution π and we take this to be uniform: $\pi = \left(\frac{1}{2}, \frac{1}{2}\right)$. That is, $\mathbb{P}(M^1) = 1/2$ and $\mathbb{P}(M^2) = 1/2$. And, the final step is to use the mixed-state operator to construct $M = \mathcal{U}_\pi (\mathrm{P}(1) \bigsqcup \mathrm{P}(2))$. The resulting multistationary ϵ-machine is shown in Fig. 1(c).

The recurrent states of the component ϵ-machines show up as M's recurrent states, as claimed. What is new is the set of two transient states (solid circles). As a generator of the multistationary process, M begins in its start state (solid circle, with circumscribing circle) and, then, follows transitions according to the edge probabilities, emitting the corresponding symbols.

We can understand M's structure by calculating its mixed states $\mu(w) = (\mathbb{P}(A), \mathbb{P}(B), \mathbb{P}(C))$, $w \in \mathcal{A}$, using Eq. (3):

$$\begin{aligned}
\mu(\lambda) &= \left[\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{4}\right] \;, \\
\mu(0) &= [0, 1, 0] = B \;, \\
\mu(1) &= \left[\tfrac{1}{2}, 0, \tfrac{1}{2}\right] \;, \\
\mu(00) &= [0, 0, 0] = \emptyset \;, \\
\mu(01) &= [0, 0, 1] = C \;, \\
\mu(10) &= [0, 1, 0] = B \;, \text{ and} \\
\mu(11) &= [1, 0, 0] = A \;.
\end{aligned}$$

FIG. 2. Entropy growth $H(\ell)$ (top) and entropy convergence $h_\mu(\ell)$ (bottom) for the Period-1 and Period-2 HMSP, as function of word length $\ell = 0, \ldots, 5$.
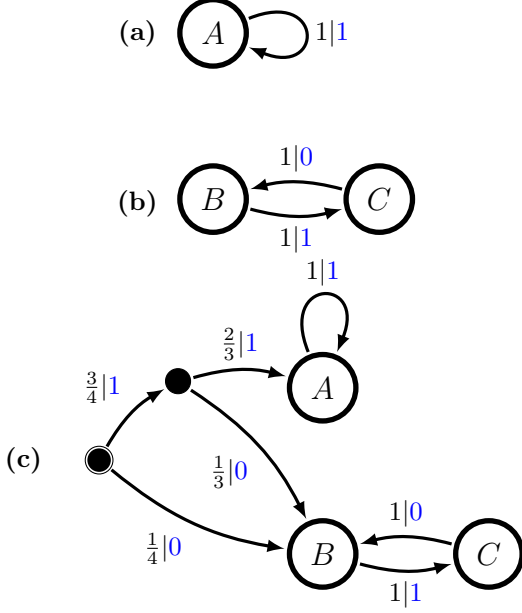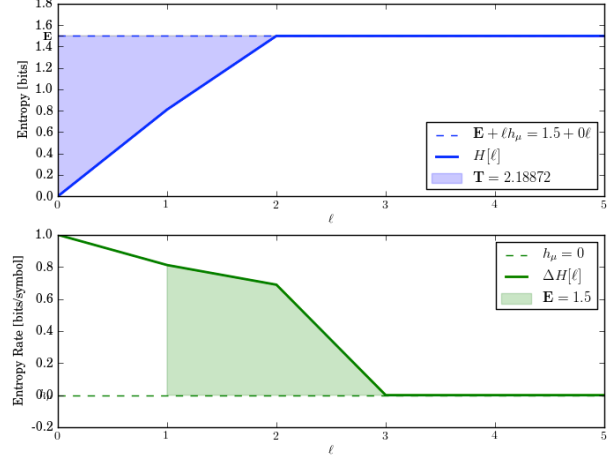
FIG. 1. The Period-1 and Period-2 Hidden Multistationary Process: (a) Component P(1), (b) Component P(2), and (c) $M = \mathcal{U}_\pi(\text{P}(1) \bigsqcup \text{P}(2))$ with $\pi = \left(\frac{1}{2}, \frac{1}{2}\right)$. Recurrent causal states are shown as hollow circles and transient causal states as small solid (black) circles. The start state sports a double circle. Transitions are labeled $p|x$ to indicate taking the transition with probability p and emitting symbol $x \in \mathcal{A}$.

In this, on the one hand, $\mu(\lambda)$ is the start state of the mixed state presentation and its distribution gives the asymptotic invariant distribution over the component recurrent states A, B, and C—the state probabilities before any symbols have been generated.

On the other hand, if $x = 0$ is generated, then we immediately know the process is in component $P(2)$, since $P(1)$ cannot produce a 0, and, in particular, it is in a specific state, B. This is reflected in the transient mixed state $\mu(0) = [0, 1, 0]$. In fact, any time a valid 0 is generated we know M is in state B. This is also seen in mixed state $\mu(10)$, in which the last symbol generated is a 0 and we again obtain a δ-function distribution concentrated on state B.

Now, there are also disallowed transitions and so disallowed words. This is shown in mixed state $\mu(00) = [0, 0, 0]$ for word $w = 00$.

More interestingly, though, is the transient mixed state $\mu(1) = \left[\frac{1}{2}, 0, \frac{1}{2}\right]$, which indicates that, having seen a 1 we know that M cannot be in state B. However, the best we can say is that it is either state A (the Period-1 component) or in state C (the Period-2 component) with fair probability. It is not

until we see another symbol that we are guaranteed to know with certainty in which component M is. If $w = 11$, then $\mathcal{P}$ is in A. Since we now know the state with certainty, we say that $w = 0$ and $w = 11$ are *synchronizing words*. In this case, they are the minimal synchronizing words.

The ergodic decompositions tell us that:

1. $h_\mu = \pi^1 h_\mu^1 + \pi^1 h_\mu^2 = 0$ bit per symbol,
2. $\mathbf{E} = H(\pi) + \pi^1 \mathbf{E}^1 + \pi^2 \mathbf{E}^2 = 1 + 0 + 1/2 = 1.5$ bits,
3. $C_\mu = H(\pi) + \pi^1 C_\mu^1 + \pi^2 C_\mu^2 = 1 + 0 + 1/2 = 1.5$ bits,
4. $\mathbf{T} = \pi^1 \mathbf{T}^1 + \pi^1 \mathbf{T}^2 = 0 + 1/2 = 1/2$ bit-symbols, and
5. $\chi = \pi^1 \chi^1 + \pi^1 \chi^2 = 0 + 0 = 0$ bits.

Let's check these by directly calculating the entropy growth $H(\ell)$ and convergence $h_\mu(\ell)$ for M. These are shown in Fig. 2.

The entropy growth plot (top) leads to an estimate of $\mathbf{E} \approx 1.5$ bits, which is predicted by the ergodic decomposition. Both entropy growth and entropy convergence (bottom) show that $h_\mu(\ell) = h_\mu = 0$ after $\ell = 2$. And, this too is correctly predicted by the corresponding entropy rate decomposition.

In fact, for lengths longer than the longest period, there are always three distinct sequences— $w \in \{1111\ldots, 0101\ldots, 1010\ldots\}$. And so, $\mathbf{E} \leq \log_2 3 \approx 1.585$ bits. This is roughly consistent with block entropy plots.

Let's analyze this exactly. One of those sequences is $w = 1^n$ and it occurs with probability $1/2$. The

two other sequences are $w = (01)^n$ and $w = (10)^n$ and they are generated equally often by their component. But since that component appears only half the time, they occur in the output sequences with probability 1/4 each. Thus, $\mathbf{E} = H[\mathbb{P}(w)] = H[(1/2, 1/4, 1/4)] = 1.5$ bits. And, this is what is seen in the plots.

The HMSP's statistical complexity is:

$$C_\mu = H\left[\mathbb{P}(\mathcal{S})\right] \tag{6}$$
$$= H\left[(1/2, 1/4, 1/4)\right] \tag{7}$$
$$= 3/2 \text{ bits .} \tag{8}$$

which agrees with the ergodic decomposition.

The ergodic decomposition, however, predicts $\mathbf{T} = 1/2$ bit-symbols, while the entropy growth plot shows that, in fact, $\mathbf{T} \approx 2.19$ bit-symbols. So, the ergodic decomposition for $\mathbf{T}$ is incorrect. In short, we see that the ergodic decomposition does not properly account for the state distribution's relaxation through the transient mixed states (solid circles) in M; Fig. 1(c). That relaxation takes longer than a single step (as the decomposition assumes) and that increased relaxation time increases $\mathbf{T}$.

Note that this is one of the simpler examples of the class of processes that have finite transients. Let's consider one that is more complex.

3. Isomorphic Golden Means Process

The No-Repeated-0s Golden Mean Process (GMP) generates all binary sequences except those with consecutive 0s. When a 0 is generated then the probability of a 0 or a 1 is fair. The GMP is an order-1 Markov process.

Let $\mathcal{P}^1$ be the No-Repeated 0s GMP, and let $\mathcal{P}^2$ be the No-Repeated-1s GMP. See Figs. 3(a) and 3(b). We define a nonergodic mixture $\mathcal{P}$ as follows:

$$\mathcal{P} = p\,\mathcal{P}^1 + (1-p)\,\mathcal{P}^2 \;,$$

with mixture distribution $\pi = (p, 1-p)$. The probability of any word w is, then:

$$\mathbb{P}(w) = p\,\mathbb{P}_1(w) + (1-p)\,\mathbb{P}_2(w) \;.$$

Using the mixed-state operator, we construct $M(\mathcal{P})$'s transient and recurrent states using this mixture distribution, finding:

$$\mu(w) = \begin{bmatrix} \mathbb{P}(A|w) & \mathbb{P}(B|w) & \mathbb{P}(C|w) & \mathbb{P}(D|w) \end{bmatrix} \;,$$
$$\mu(\lambda) = \begin{bmatrix} \frac{2p}{3} & \frac{p}{3} & \frac{2(1-p)}{3} & \frac{1-p}{3} \end{bmatrix} \;,$$
$$\mu(0) = \begin{bmatrix} 0 & \frac{p}{2-p} & \frac{2(1-p)}{2-p} & 0 \end{bmatrix} \;,$$
$$\mu(1) = \begin{bmatrix} \frac{2p}{1+p} & 0 & 0 & \frac{1-p}{1+p} \end{bmatrix} \;,$$
$$\mu(00) = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix} = C \;,$$
$$\mu(01) = \begin{bmatrix} p & 0 & 0 & 1-p \end{bmatrix} \;,$$
$$\mu(10) = \begin{bmatrix} 0 & p & 1-p & 0 \end{bmatrix} \;,$$
$$\mu(11) = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix} = A \;,$$
$$\mu(001) = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix} = D \;, \text{ and}$$
$$\mu(110) = \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix} = B \;.$$

Longer words can only lead to one of these mixed states and so the ϵ-machine is finite. The full multistationary ϵ-machine is shown in Fig. 3(c), as a function of the mixture parameter p. We see that the number of states, including the transients, is finite for all mixture probabilities.

The transition matrices for $M(\mathcal{P})$'s recurrent causal states are:

$$T^0 = \begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \text{ and } T^1 = \begin{bmatrix} \frac{1}{2} & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 \end{bmatrix} \;.$$

The stationary distribution is defined by the mixture of the two processes:

$$\pi(p) = \begin{pmatrix} p\pi^1 & (1-p)\pi^2 \end{pmatrix}$$
$$= \tfrac{1}{3}\begin{pmatrix} 2p & p & 2(1-p) & 1-p \end{pmatrix} \;,$$

recalling that $\pi^1 = \pi^2 = \begin{pmatrix} 2/3 & 1/3 \end{pmatrix}$.

Using methods from Refs. [12, 13], the excess entropy for each recurrent component is seen to be:

$$\mathbf{E}^1 = \mathbf{E}^2 = \frac{2}{3}\log_2\frac{3}{2} + \frac{1}{3}\log_2 3 - \frac{2}{3}$$
$$= \frac{2}{3}\log_2\frac{3}{4} + \frac{1}{3}\log_2 3$$
$$\approx 0.251629 \text{ bits.}$$

By the ergodic decomposition theorem, the excess entropy for the mixture, as a function of p is:

$$\mathbf{E}(p) = p\mathbf{E}^1 + (1-p)\mathbf{E}^2 + H(p)$$
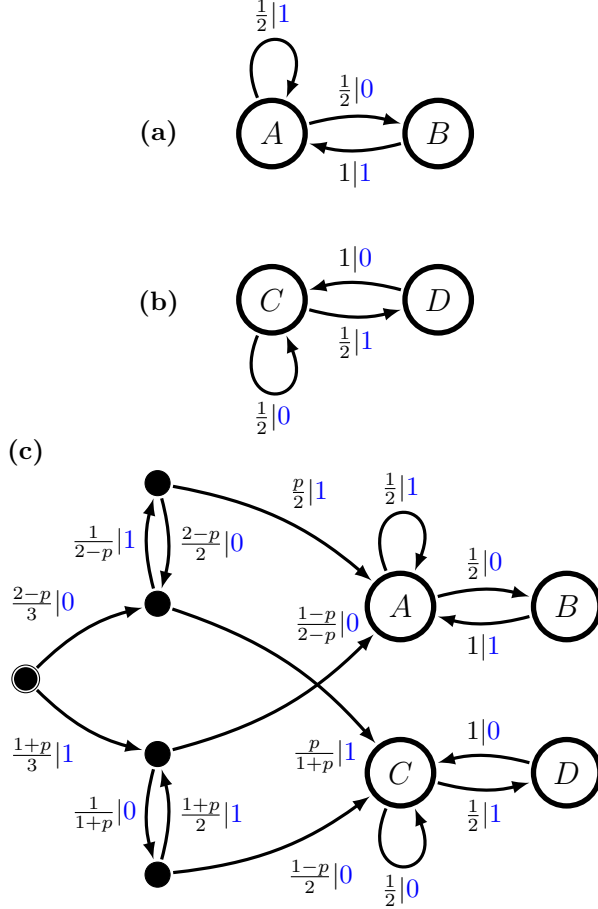$$= \mathbf{E}^1 + H(p) \;,$$

(a)

(b)

(c)

FIG. 3. Two Golden Mean Processes and their non-ergodic mixture: (a) M^1, (b) M^2, and (c) $M = \mathcal{U}_\pi(M^1 \bigsqcup M^2)$ with $\pi = (p, 1-p)$.
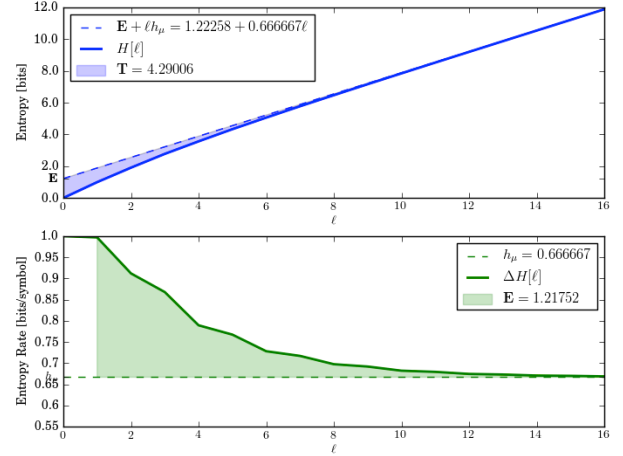

FIG. 4. Entropy growth $H(\ell)$ (top) and entropy convergence $h_\mu(\ell)$ (bottom) for the Two Isomorphic Golden Means HMSP, as function of word length $\ell = 0, \ldots, 16$ and mixture parameter $p = 1/2$.

So, from Eq. (6) the statistical complexity of the mixture as a function of p is:

$$C_\mu(p) = C_\mu^1 + H(p) . \qquad (9)$$

For $p = 1/2$, we expect $C_\mu \approx 1.9182958$ bits.

Let's check the decompositions by calculating the associated complexity measures from M's entropy growth and convergence. The latter are shown in Fig. 4.

The entropy growth plot estimates that $\mathbf{E} = 1.22258$ bits, which is low by 2%. And, the entropy convergence plot shows that the $\mathbf{E}$, calculated there using Eq. (1) as the area shown, is a bit lower still: $\mathbf{E} = 1.21753$. Although, due to the slow convergence and the finite number of terms taken in the approximation these errors are expected. Similarly, the ergodic decomposition of the entropy rate $h_\mu = 0.811278$ bits per symbol shows up correctly when estimated from $\mathcal{P}$'s entropy growth and convergence. And so, the predictions from the related ergodic decompositions are consistent.

The entropy growth, however, shows the transient information is substantially larger ($\mathbf{T} \approx 4.29$ bit-symbols) than that predicted from its ergodic decomposition ($\mathbf{T} = 0.251629$). This discrepancy is clearly not due to estimation errors. Rather, as noted above for the P1-P2 mixture, it arises from the decomposition not accounting for the five transient causal states of M; see Fig. 3(c).

Individually, GMPs are subshifts of finite type; finite Markov order. From the cycles in the tran-

since the two components are isomorphic. For $p = 1/2$, we expect $\mathbf{E} \approx 1.251629$ bits.

Again, the component transient information equals the excess entropy, since the GMP is order-1 Markov. So, the associated ergodic decomposition gives:

$$\begin{aligned} \mathbf{T}(p) &= p\mathbf{T}^1 + (1-p)\mathbf{T}^2 \\ &= \mathbf{T}^1 , \end{aligned}$$

since the two components are isomorphic. For $p = 1/2$, we expect $\mathbf{T} \approx 0.251629$ bits.

Similarly, the statistical complexity of each recurrent component is:

$$\begin{aligned} C_\mu^1 = C_\mu^2 &= \frac{2}{3}\log_2\frac{3}{2} + \frac{1}{3}\log_2 3 \\ &\approx 0.9182958 \text{ bits.} \end{aligned}$$

sient states we see that as components they make the multistationary process sofic—infinite Markov order. There are subsets of sequences—specifically $(01)^n$—for which one never synchronizes.

This means that mixtures of finite-order Markov chains, even "linear" mixtures that come from independently running them, are processes that are not finite Markovian. They require hidden Markov representations.

B. Infinite State

The preceding examples, chosen to explicitly illustrate methods and as harbingers of coming results, are rather special in that the led to finite-state multistationary processes. We now turn to more typical cases, still constructed from finite-state ergodic components, that lead to a multistationary process with an infinite number of states.

1. Period-1 and Fair Coin Process

The next example multistationary process mixes stochastic and periodic behaviors: We build it out of a period-1 process and a fair coin. In effect, we ask how difficult it is to distinguish these two simple, but extreme processes—one completely predictable, the other completely unpredictable.

For here and a bit later, define *Bernoulli Process* $B(p)$ which is a model of a coin flip with bias probability p.

The first step, then, is to select the two stationary components:

1. Period-1 Process P(1): $h_\mu^1 = 0$ bits per symbol, $C_\mu^1 = 0$ bits, $\mathbf{E}^1 = 0$ bits, $\mathbf{T}^1 = 0$ bit-symbols, and $\chi^1 = 0$ bits. See Fig. 5(a).

2. Fair Coin Process $B\left(\frac{1}{2}\right)$:, $h_\mu^1 = 1$ bit per symbol, $C_\mu^1 = 0$ bits, $\mathbf{E}^1 = 0$ bits, $\mathbf{T}^1 = 0$ bit-symbols, and $\chi^1 = 0$ bits. See Fig. 5(b).

Though at the two extremes of predictability, these are structurally trivial processes—$C_\mu^i = 0$.

The second step is to select the mixture distribution, which we take to be uniform: $\pi = \left(\frac{1}{2}, \frac{1}{2}\right)$. And the third step is to use the mixed-state operator to construct $M = \mathcal{U}_\pi\left(\text{P}(1) \bigsqcup B\left(\frac{1}{2}\right)\right)$. Several of the mixed states are:

$$\mu(\lambda) = \left[\tfrac{1}{2}, \tfrac{1}{2}\right] ,$$
$$\mu(0) = [0, 1] = B$$
$$\mu(1) = \left[\tfrac{3}{4}, \tfrac{1}{4}\right] ,$$
$$\mu(0\mathcal{A}^+) = [0, 1] = B ,$$
$$\mu(1^+0) = [0, 1] = B ,$$
$$\mu(1^k) = [1 - \alpha_k, \alpha_k] ,$$
$$\vdots , \text{ and}$$
$$\mu(1^\infty) = [1, 0] = A,$$

where $\alpha_k = (2k + 1)/2^{k+1}$. The resulting ϵ-machine is shown in Fig. 5(c).

In Fig. 5(c), and so too in the mixed states, we see our first surprising result for multistationary processes. Starting from two structurally trivial processes, the multistationary ϵ-machine has a countable infinity of transient causal states. Why? If, at any point, one sees a 0, then we know the process is in the Fair Coin component, since the other component cannot generate a 0. However, it is only after "seeing" an infinite sequence of 1s that one could determine that the process is in the All-1s component. In short, the effort required to distinguish between these two trivial processes is infinite and this is directly reflected in the infinite set of transient states.

The ergodic decompositions tell us that:

1. $h_\mu = \pi^1 h_\mu^1 + \pi^2 h_\mu^2 = 0 + 1/2 = 1/2$ bit per symbol,
2. $\mathbf{E} = H(\pi) + \pi^1 \mathbf{E}^1 + \pi^2 \mathbf{E}^2 = 1 + 0 + 0 = 1$ bits,
3. $C_\mu = H(\pi) + \pi^1 C_\mu^1 + \pi^2 C_\mu^2 = 1 + 0 + 0 = 1$ bits,
4. $\mathbf{T} = \pi^1 \mathbf{T}^1 + \pi^2 \mathbf{T}^2 = 0 + 0 = 0$ bit-symbols, and
5. $\chi = \pi^1 \chi^1 + \pi^2 \chi^2 = 0 + 0 = 0$ bits.

Note that the ergodic decompositions predict that the structural complexity measures are driven solely by the mixture entropy $H(\pi)$. Both components contribute nothing: $\mathbf{E}^i = C_\mu^i = 0$. Let's check these predictions by estimating the quantities from M's entropy growth and convergence, shown in Fig. 6.

The entropy growth plot shows that $\mathbf{E} = 1$ bit, as predicted by M's ergodic decomposition. And, the entropy convergence plot shows that the $\mathbf{E}$, calculated as the area shown, is also the same. Similarly, the ergodic decomposition of the entropy rate $h_\mu = \frac{1}{2}$ bits per symbol shows up correctly on the entropy plots.
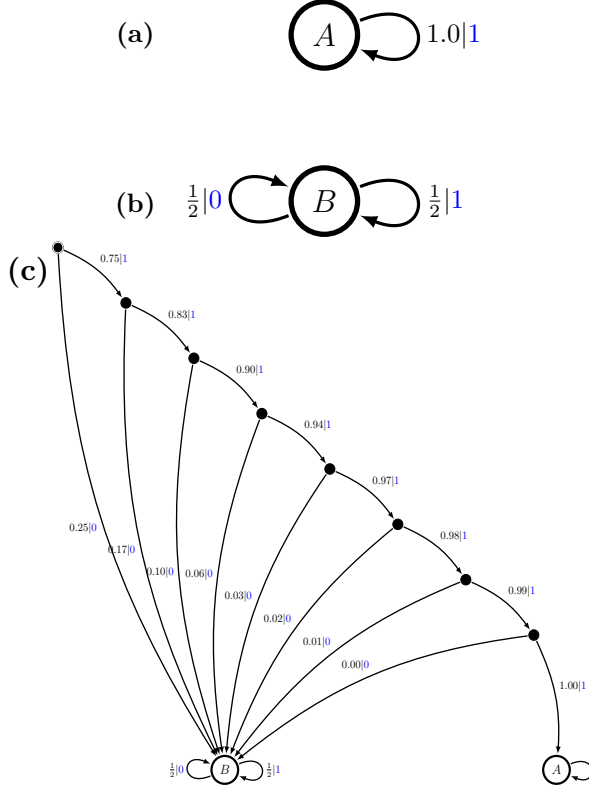
FIG. 5. Period-1 and Fair Coin HMSP: (a) P(1), (b) $B\left(\frac{1}{2}\right)$, and (c) $M = \mathcal{U}_\pi\left(\text{P}(1) \bigsqcup B\left(\frac{1}{2}\right)\right)$ with $\pi = \left(\frac{1}{2}, \frac{1}{2}\right)$. The latter is approximated by connecting the Period-1 component after 8 1s. In fact, the $P(1)$ component is never reached after any finite sequence. In contrast, $B\left(\frac{1}{2}\right)$ can be reached quickly and by many sequences. Recurrent causal states are shown as hollow circles and transient causal states as small solid (black) circles. The start state sports a double circle. Transitions are labeled $p|x$ to indicate taking the transition with probability p and emitting symbol $x \in \mathcal{A}$.

The entropy growth plot, however, shows the transient information is quite a bit larger ($\mathbf{T} \approx 2.78$ bit-symbols) than that predicted ($\mathbf{T} = 0$) from its ergodic decomposition.

Also, the informational ergodic decompositions, while indicating a role for the mixture entropy, miss entirely the existence of an infinite number of transient states and the attendant difficulty that confronts an observer trying to detect in which component the process is.

2. Two Biased Coins

Slightly increasing the level of sophistication, we now construct a multistationary process out of fully
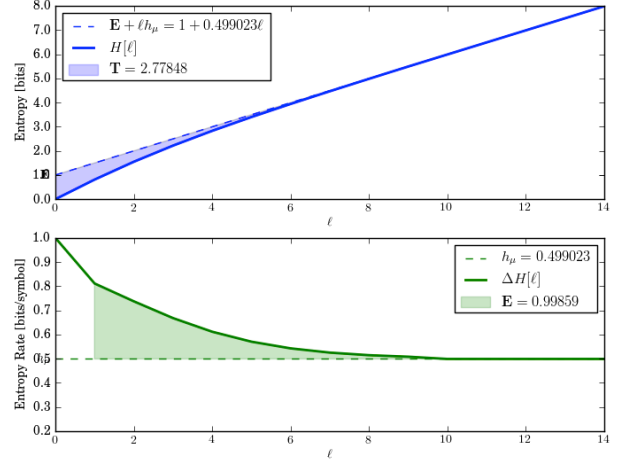

FIG. 6. Entropy growth $H(\ell)$ (top) and entropy convergence $h_\mu(\ell)$ (bottom) for the Period-1 and Fair Coin HMSP, as function of word length $\ell = 0, \ldots, 14$.
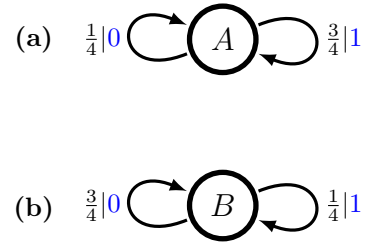

FIG. 7. Two Biased Coins HMSP Components: (a) $B\left(\frac{1}{4}\right)$ and (b) $B\left(\frac{3}{4}\right)$.

stochastic components: Two biased coins of unequal (but symmetric) biases $B\left(\frac{1}{4}\right)$ and $B\left(\frac{3}{4}\right)$. See Figs. 7(a) and (b).

We again take a uniform mixture distribution: $\pi = \left(\frac{1}{2}, \frac{1}{2}\right)$. The result of constructing $M = \mathcal{U}_\pi\left(B\left(\frac{1}{4}\right) \bigsqcup B\left(\frac{3}{4}\right)\right)$ with $\pi = \left(\frac{1}{2}, \frac{1}{2}\right)$ is shown in Fig. 8.

The mixed state presentation reveals two countably-infinitely-long chains of transient causal states. One leads to the ergodic component for $B\left(\frac{3}{4}\right)$ and the other for $B\left(\frac{1}{4}\right)$. In a simple sense these long transient chains show the mechanism by which one determines the coin biases. Interestingly, though, at any point statistical fluctuations can change the apparent bias and drive the state back up the long chains, heading for the complementary biased coin.

Consider, as above the h_μ, $\mathbf{E}$, C_μ, $\mathbf{T}$, and χ ergodic decompositions. The ergodic decompositions for excess entropy and statistical complexity give
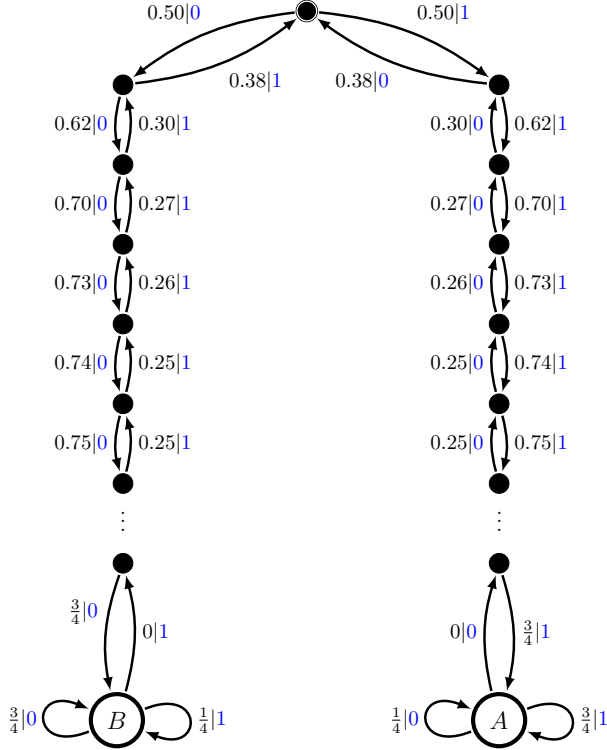
FIG. 8. Two Biased Coins HMSP: $M = \mathcal{U}_\pi \left(B \left(\frac{1}{4} \right) \bigsqcup B \left(\frac{3}{4} \right) \right)$ with $\pi = \left(\frac{1}{2}, \frac{1}{2} \right)$. The latter is approximated by connecting the transient states to $B \left(\frac{1}{4} \right)$ and $B \left(\frac{3}{4} \right)$ after 30 0s or 30 1s, respectively. In fact, $B \left(\frac{1}{4} \right)$ and $B \left(\frac{3}{4} \right)$ are never reached after any finite word.
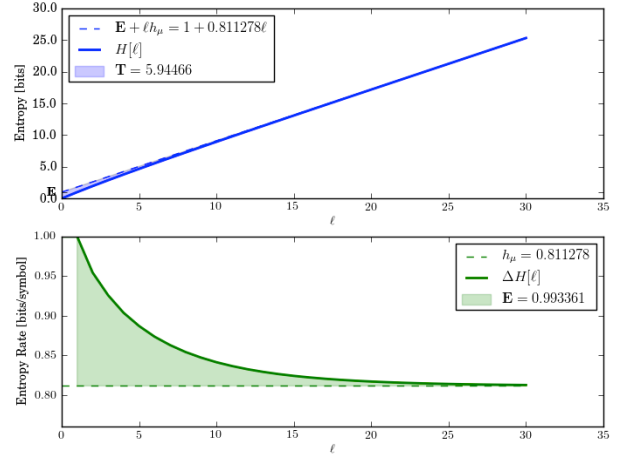

FIG. 9. Entropy growth $H(\ell)$ (top) and entropy convergence $h_\mu(\ell)$ (bottom) for the Two Biased Coins HMSP, as function of word length $\ell = 0, \ldots, 30$.

Entropy growth, using the y-intercept method, shows that $\mathbf{E} = 1$ bit, as predicted by M's ergodic decomposition. And, the entropy convergence plot shows that $\mathbf{E}$, as the area shown, is also the same, though it takes many terms and so shows slow convergence. The ergodic decomposition of the entropy rate $h_\mu = 0.811278$ bits per symbol shows up correctly on the entropy plots.

The entropy growth, however, shows the transient information is substantially larger ($\mathbf{T} \approx 5.95$ bit-symbols) than that predicted ($\mathbf{T} = 0$) from its ergodic decomposition. Again, the mixing entropy fails to account for the dominating transient causal state structure.

3. Pair of Isomorphic Even Processes

The Even Process (EP) generates all binary sequences such that pairs of 1s occur in blocks of even length bounded by 0s. Once a 0 is seen, a 0 or a 1 is generated with fair probability. The EP is closely related to the Golden Mean Process. They have the same entropy rates and statistical complexities. The main important difference, despite the close similarity and a simple relabeling of transitions, is that the EP is described by no finite-order Markov chain. It is infinite Markov order, though finite state.

To construct a multistationary process the first step, then, is to select two stationary components. One component $\mathcal{P}^1$ will be an EP with even number of 0s (Fig. 10(b)) and the other $\mathcal{P}^2$ an EP with an even number of 1s (Fig. 10(a)). The second

similar results; namely:

$$\mathbf{E} = H[\pi] + \mathbf{E}(M^1) + \mathbf{E}(M^2)$$
$$\mathbf{E} = H[\pi] \text{ , and}$$
$$C_\mu = H[\pi] + C_\mu(M^1) + C_\mu(M^2)$$
$$C_\mu = H[\pi] \text{ .}$$

That is, the complexities of the multistationary process is all in the mixture distribution. Even then, the mixture entropy, in this case a number upper bounded by 1, belies the infinite number of transients and the difficulty of determining in which ergodic component the process is. Quantitatively, it seems another measure of the global process complexity and a new decomposition are in order. We return to this shortly, after examining several more kinds of multistationary process.

Let's validate the ergodic decompositions' predictions vis a vis the process estimates of their various measures from M's entropy growth and convergence. The latter are shown in Fig. 9.

step is to choose mixture distribution: $\pi(M^1, M^2) = (1/2, 1/2)$.

Finally, Fig. 11 shows the ϵ-machine for the HMSP $M = \mathcal{U}_\pi\left(M^1 \bigsqcup M^2\right)$ with $\pi\left(M^1, M^2\right) = \left(\frac{1}{2}, \frac{1}{2}\right)$. The ϵ-machine displayed is estimated only up to words of length 8 and the transitions are set to give a well-formed ϵ-machine at this approximation.

There are several observations. First, the HMSP ϵ-machine is symmetric under 1-0 exchange, as it should be given this symmetry in the ergodic components. Second, and less obviously, there is an infinite number of transient causal states. This is due to the outside paths along 1^∞ and 0^∞. These two sequences arise from the 2-cycles in the respective ergodic component recurrent states: pairs of 1s in M^1 never synchronize; ditto for pairs of 0s in M^2. And so, in M there are infinitely long sequences that never reach M^1 or M^2.

Third, the HMSP is infinite Markov order. To see this, note that there are six cycles in the transient states—these cycles are the signature of infinite Markov order or, what is called "soficity". The HMSP is a shift of infinite type [28]. In particular, there is a two-cycle $(00)^+$ between states 32 and 38 and one $(11)^+$ between states 37 and 41. There are two four-cycles $(1100)^+$ between states, 10, 18, 24, and 26 and between states 24, 29, 34, and 36; and two $(0011)^+$ between states 11, 19, 25, and 23 and between states 25, 30, 35, and 33.

The ergodic decompositions tell us that:

1. $h_\mu = \pi^1 h_\mu^1 + \pi^1 h_\mu^2 = h_\mu^1 = \frac{2}{3}$ bit per symbol,
2. $\mathbf{E} = H(\pi) + \pi^1 \mathbf{E}^1 + \pi^2 \mathbf{E}^2 = H(\pi) + \mathbf{E}^1 = 1 + 0.918296 = 1.918296$ bits,
3. $C_\mu = H(\pi) + \pi^1 C_\mu^1 + \pi^2 C_\mu^2 = H(\pi) + C_\mu^1 = 1 + 0.918296 = 1.918296$ bits,
4. $\mathbf{T} = \pi^1 \mathbf{T}^1 + \pi^1 \mathbf{T}^2 = \mathbf{T}^1 = 3.16938$ bit-symbols, and
5. $\chi = \pi^1 \chi^1 + \pi^1 \chi^2 = \chi^1 = 0$ bits.

Let's check these by directly calculating the entropy growth and convergence for M. These are shown in Fig. 12.

Let's check the decompositions by comparing their predictions to estimates from M's entropy growth and convergence. These functions are shown in Fig. 12.

The entropy growth plot shows that $\mathbf{E} = 1.847$ bits which disagrees by about 4% with the prediction from the ergodic decomposition ($\mathbf{E} = 1.918$ bits). And, the entropy convergence plot shows that the
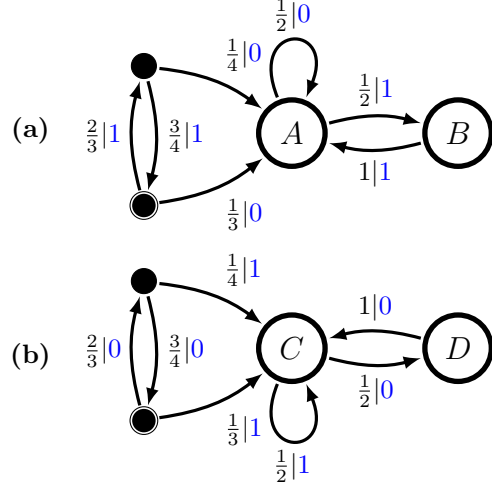

FIG. 10. Two Even Processes: (a) M^1, pairs of 1s, with its two transient causal states and (b) M^2, pairs of 0s, with its two transient causal states.

$\mathbf{E} = 1.847$ bits as the area shown. Although, due to the slow convergence, the finite number of terms taken in the numerical approximation, and the finite number of transient states taken in the approximation of M, this error is not surprising. Similarly, the ergodic decomposition of the entropy rate $h_\mu = 2/3$ bits per symbol shows up correctly on the entropy plots ($h_\mu \approx= 0.6666$).

Entropy growth, however, shows the transient information is three times larger ($\mathbf{T} \approx 9.77443$ bit-symbols) than that predicted from its ergodic decomposition ($\mathbf{T} = 3.16938$ bit-symbols). Again, this discrepancy follows from the mixture entropy missing the contributions from the (infinite) number of transient causal states.

C. Infinite Components

We end our selection of example multistationary processes by constructing several from an infinite number of finitary ergodic components.

1. Handbag of Necklaces

Fourier analysis of a signal assumes the generating process consists of (at most) periodic sequences. As an analog of this assumption in the present setting, consider the *Handbag of Necklaces* (HMSP) consisting of ergodic-component stationary processes $P(i)$ for all periods, $i \in I = 1, 2, 3, 4, \ldots$. That is, if we
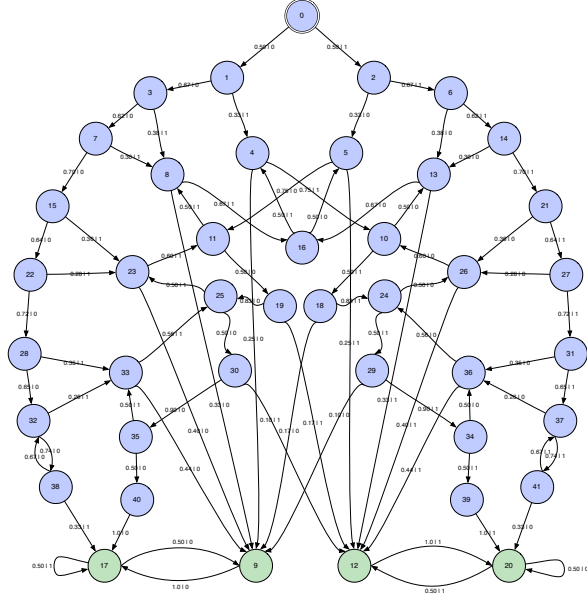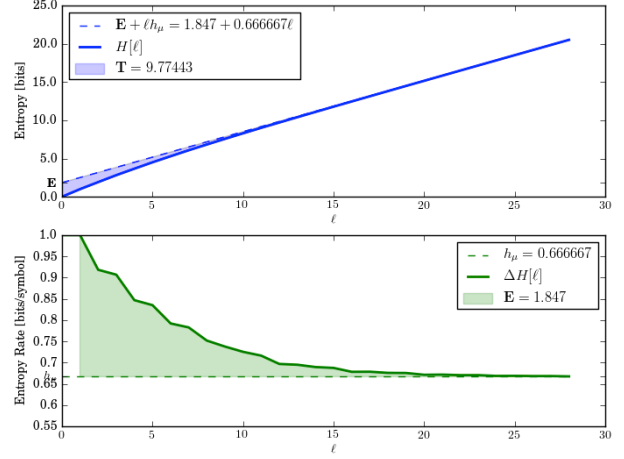
FIG. 12. Entropy growth $H(\ell)$ (top) and entropy convergence $h_\mu(\ell)$ (bottom) for the Two Isomorphic Even Multistationary Process, as function of word length $\ell = 0, \ldots, 28$.

$$H[\pi] = \sum_{i=1}^{\infty} \pi^i \log_2 \pi^i < \infty .$$

3. Infinitary ($\mathbf{E} \to \infty$) HMSP [15]: Must have $H[\pi] \to \infty$. See Ref. [29] for another example process in this family.

Consider the structure of the transitions in the HMSP's first row of states. The first transition probability for seeing an a is:

$$\mathbb{P}(a) = \sum_{i=1}^{\infty} \mathbb{P}(a|M^i)\mathbb{P}(M^i)$$
$$= \sum_{i=1}^{\infty} \pi^i / i ,$$

since the probability of seeing an a in the ith component is $1/i$. Let $p_a = \mathbb{P}(a)$. The probability for the succeeding transition emitting an a is:

$$\mathbb{P}(a|a) = \mathbb{P}(aa)/\mathbb{P}(a)$$
$$= \pi^1/p_a ,$$

since $\mathbb{P}(aa) = \pi^1$. These transitions determine those leaving the top row of states on a b. Note that $\mathbb{P}(a^n|aa) = 1, n = 1, 2, \ldots$.

Now, consider the second to the top row of transitions. First, $\mathbb{P}(b) = 1 - p_a$. Then, we have:

$$\mathbb{P}(a|b) = \mathbb{P}(ba)/\mathbb{P}(b)$$
$$= \frac{p_a - \pi^1}{1 - p_a} .$$

FIG. 11. Two Isomorphic Even Process HMSP: $M = \mathcal{U}_\pi \left(M^1 \bigsqcup M^2 \right)$ with $\pi \left(M^1, M^2 \right) = \left(\frac{1}{2}, \frac{1}{2} \right)$. Approximated with maximum word length of 8. Transient states are reconnected to mimic the component EP's transient states. Specifically, calculating M using word length 8 leaves M's states $38 - 41$ as dangling states—with no outgoing transitions. Transitions were added by noting that these states closely approximate the EPs' two transient states, seen in Figs. 10(a) and (b), at sufficiently large word length.

assume a binary process, the sequences emitted consist of words period-1 a^+, period-2 $(ab)^+$, period-3 $(bba)^+$, period-4 $(bbba)^+$, and so on. The HMSP ϵ-machine is shown in Fig. 13.

Note that there is infinite number of transient causal states. Overall the HMSP is a highly symmetric structure and dominated by the transient states. From this one can readily read-off how to synchronize—how to know in which ergodic component the process is. For example, to get to component M^i there are exactly i paths.

Now, consider the mixture measure π^i for the components. Then, the state probabilities are $\pi_{ij} = K\pi^i/j$, $j = 1, \ldots, i$, where $K = \sum_{i=1}^{\infty} \sum_{j=1}^{i} \pi^i/j$ is the normalization constant. Note this is the presentations' stationary invariant distribution.

There is some flexibility in setting the mixture distribution π^i. There are several criteria for choosing it for a countable number of states:

1. Normalization: $\sum_{i=1}^{\infty} \sum_{j=1}^{i} \pi_{ij} = \sum_{i=1}^{\infty} \pi^i = 1$. And so, π^i must decay faster than $1/i$.
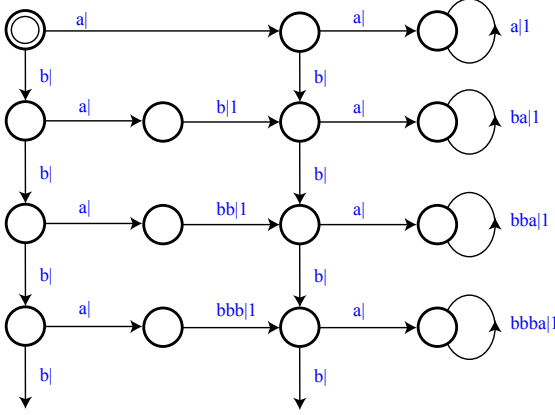2. Finitary ($\mathbf{E} < \infty$) HMSP [15]: Must have

FIG. 13. The ϵ-machine for HMSP presentation of the Handbag of Necklaces Process. Chains of probability-1 transitions between causal states are given as strings.
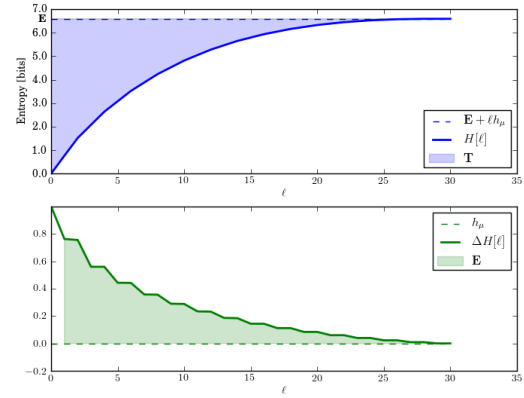

FIG. 14. Entropy growth $H(\ell)$ (top) and entropy convergence $h_\mu(\ell)$ (bottom) for the Handbag of Necklaces Multistationary Process, as function of word length $\ell = 0, \ldots, 30$ and up to period $p = 15$. (Note that the mixed states were approximated up to Max_Length = 30.)

since:

$$\mathbb{P}(ba) = \sum_{i=2}^{\infty} \mathbb{P}(M^i)\mathbb{P}(ba|M^i)$$
$$= \sum_{i=2}^{\infty} \pi^i/i$$
$$= p_a - \pi^1 .$$

Note that $\mathbb{P}(bab) = \mathbb{P}(ba)$. There is a second path to M^2 controlled by the transition:

$$\mathbb{P}(b|a) = \mathbb{P}(ab)/\mathbb{P}(a)$$
$$= \frac{p_a - \pi^1}{p_a} .$$

since $\mathbb{P}(ab) = \mathbb{P}(ab)$. That is, the appearance of ab and ba in each component occurs due to the same conditions.

The ergodic decompositions tell us that:

1. $h_\mu = \sum_{i=1}^{\infty} \pi^i h_\mu^i = 0$, as $h_\mu[P(i)] = 0$.
2. The excess-entropy ergodic decomposition for a process with p component periodic processes with periods $1, \ldots, p$ is:

$$\mathbf{E} = H[\pi] + \sum_{i=1}^{p} \pi^i \mathbf{E}^i \qquad (10)$$

$$= H[\pi] + \sum_{i=1}^{p} \pi^i \log_2 i \qquad (11)$$

$$= \log_2 p + 1/p \sum_{i=2}^{p} \log_2 i , \qquad (12)$$

where the second step follows assuming π^i is uniform.
3. $C_\mu = \mathbf{E}$.
4. $\mathbf{T} = \sum_{i=1}^{p} \pi^i \mathbf{T}^i = \sum_{i=2}^{p} \pi^i \log_2 i$ bit-symbols, and
5. $\chi = 0$: This is a bit surprising: No crypticity, no hidden information.

These are consistent with directly calculating the entropy growth for M, as shown in Fig. 14.

2. The Purse Process

The example of two biased coins suggests extending to an infinite number of biased coins in a purse— a bag of coins with different biases. As hinted at in the two-coin case, *all* of the (infinite) complexity is in the mixture and *none* comes from the components.

Moreover, we can chose π to be such that $H[\pi]$ is finite or infinite. Thus, the Purse Process is an extreme example in which infinite complexity comes from zero-complexity components. There is probably no simpler way to say that a multistationary process is way more than the sum of its (zero complexity) parts.

To get a brief sense of the Purse Process consider an HMSP consisting of three coins of unequal bias and compare this to the case of two coins of Sec. VII B 2. Figure 15 shows the HMSP for two completely biased coins and one fair coin. Its basic fea-
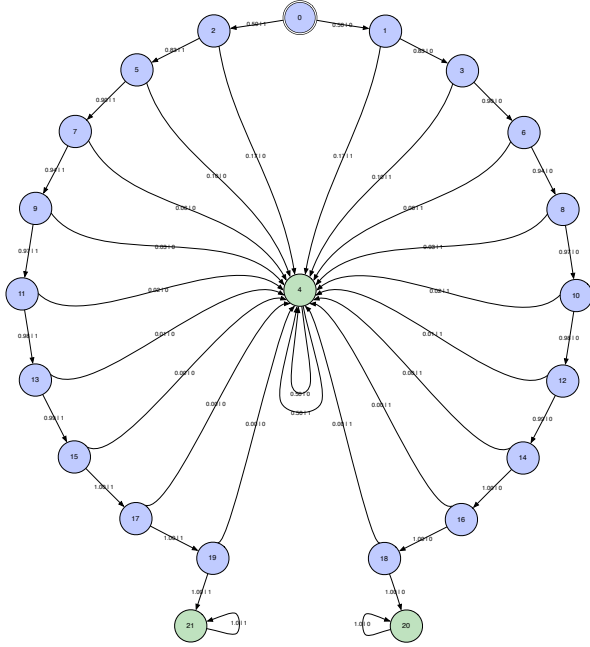
FIG. 15. The ϵ-machine HMSP presentation of two completely biased coins and one fair coin—all single-state ϵ-machines: $M = \mathcal{U}_\pi \left(\text{All-1s} \bigsqcup \text{All-0s} \bigsqcup B \left(\frac{1}{2} \right) \right)$ with $\pi = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right)$.

tures were already encountered above. And, it suggests a notable generalization to which we now turn.

3. Mother of All Processes

Finally, consider upping the complexity ante substantially. This generalization is to an HMSP consisting of a mixture of all processes. Let's step through its construction.

First, recall that every stationary process has a unique ϵ-machine presentation. That is, ϵ-machines and stationary processes are in 1-to-1 correspondence. Second, an efficient algorithm exists to list all ϵ-machines by the number of the recurrent causal states. Reference [30] shows how to systematically enumerate the ϵ-machine process library $\mathcal{L}_k$ for k-state ϵ-machines. See Table I there for the list of binary-alphabet topological ϵ-machines. There are $1,117,768,214$ such 8-state ϵ-machines.

In the current construction consider only topological ϵ-machines for which any branching transitions are taken with fair probability. We refer to each process by it's ϵ-machine's enumeration number—we call this the process' Gödel number.

Second, define the Process Urn (PU) as containing

the entire library of ϵ-machines. That is, we imagine a HMSP that is the result of reaching into the Urn, selecting one ϵ-machine, and having it generate a full realization. The repeatedly sampled PU is a HMSP—*The Mother of All Processes*. Certainly, one of the most nonergodic processes one could work with.

Definition 2. *The* Mother of All Processes *is:*

$$M = \mathcal{U}_\pi \left(\bigsqcup_{M \in \mathcal{L}} M \right) , \qquad (13)$$

with π being a chosen mixture distribution.

To simplify, let's examine the HMSP whose components are all one-state and all two-state ϵ-machines. There are now 10 components—three 1-state components and 7 2-state components.

There are 17 recurrent causal states altogether across the ergodic components. However, the many hundreds of mixed states are no longer usefully presented in a state-transition diagram, as done up to this point. Instead, we plot the mixed-states themselves as dots in the simplex Δ^{17}. This is shown in Fig. 16. This is a 2D projection in which the recurrent states are the vertices of Δ^{17} and so appear on its periphery. The start state, with uniform probability across the components and not across the recurrent states, is not in the simplex center.

One notes the concentration of mixed states that move near to Δ^{17}'s vertices, indicating close approaches to synchronization.

There are a number of notable properties:

1. The simplex vertices correspond to recurrent causal states.
2. There is an uncountably-infinite number of transient states. These fill out a complicated fractal measure within the Δ^{17}.
3. All mixed states that are not on vertices are transient states.
4. While it is clear that M is not exactly synchronizable [31] as it contains infinite Markov-order components, is it asymptotically synchronizable [32]? What about the synchronizability of approximations to it?

There are a number of open questions, as well:

1. What is M's statistical complexity dimension d_μ [33]?
2. What is the shortest synchronizing word to go from the Fair Coin to each other ergodic component?
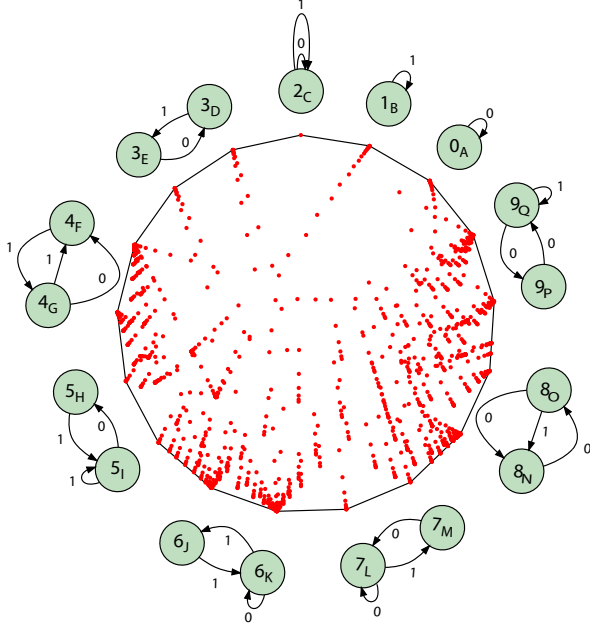
FIG. 16. The ϵ-machine presentation of the Mother of All 1-State and 2-State HMSP: $M = \mathcal{U}_\pi \left(\bigsqcup_{M \in \mathcal{L}_{\{1,2\}}} M \right)$ with $\pi = \left(\frac{1}{10}, \ldots, \frac{1}{10} \right)$. The start state is the 17-vector at $\left(\frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{20}, \ldots, \frac{1}{20} \right)$. The ϵ-machines for each ergodic component are placed around the periphery, dropping the transition probabilities which are either 1 or $\frac{1}{2}$. Their states are aligned to their associated vertex in Δ^{17}. The Fair Coin is the state at the top—the top-most vertex. The mixed states are approximated up to word lengths of $\ell = 28$.

3. How do informational measures grow with word-length approximation?

VIII. DISCUSSION

In addition to their particular application, the ergodic decompositions give important insight into basic questions about what structural complexity is and how to measure it. A number of previous efforts that address these definitional issues consider it a key property that complexity be additive over the components of a system [34]. This is often motivated by a parallel with Boltzmann entropy in thermodynamics. And, for that matter, additivity was also posited as an axiom by Shannon for his measure of surprise [35].

However, the ergodic decompositions here show that the manner in which a system's components relate to one another—specifically, the mixture distribution—plays a central role in the process' or-

ganization and contributes to quantitative measures of global complexity. The foregoing offered a different, more structural view that goes beyond the ergodic decompositions and statistical mixtures. Constructively, the transient state structure is key to a multistationary process' global organization and what observations can or cannot reveal.

The lessons here also suggest a skepticism in applying the ergodic decompositions of Sec. VI. One reason is that underlying them is the assumption of an IID sampling of components, which is not generally valid. Another is that they completely ignore how the internal structures of the components interrelate with each other. And, as shown, this brings out wholly new properties that are not part of any given component nor their sum nor their IID mixture. Indeed, the mixture entropy does not capture this, except in the most limited of cases.

Constructive responses to this will address the new kind of hierarchical structure explicitly represented by the multistationary process' ϵ-machine transient causal states and their complicated measure in Δ^k. Quantitatively, in contrast to the block entropy, entropy rate, and excess entropy, we demonstrated that the transient information is sensitive to this new kind of complexity in structural mixtures. It is this additional structure that makes the organization of multistationary processes way more than the sum of their parts. As a complementary metric, adapting the statistical complexity dimension d_μ suggests itself [33].

IX. CONCLUSIONS

Let's close exploring several wider implications for thermodynamics, on the one hand, and various attempts to introduce "universal modeling" schemes on the other.

First, we started out highlighting the colloquialism, made familiar by the social movements of the 1960s, that a system is more than the sum of its parts. Presumably, the social reaction then reflected an increasing awareness of the impact of technical systems humans were creating. The preceding development explored in which senses this could be true for truly complex systems—ones consisting many *structured* components—more akin the social subsystems than mere atoms. And, the various informational ergodic decompositions bolstered the popular understanding.

However, in emphasizing structure and analyzing the concrete process class of hidden multistationary processes it became abundantly clear—through all of the examples presented—that composite or heterogeneous (to use Gibbs' word [36]) systems are far more than the sum of their components. Specifically, beyond a mere entropic, missing contribution from increased disorder that arises from the random selection of components, composite systems are markedly more complex. And, they are more structured according to the relative interplay of the components' internal organization. It is that interplay which drives the explosive complexity of multicomponent systems.

On this score, the history of composite systems is perhaps a bit confusing; especially as they arose in the early foundations of thermodynamics. There is, for example, Gibbs' seemingly contradictory statement, as quoted by Jaynes [1, p. 13], that "The whole is simpler than the sum of its parts". The ergodic decompositions seemed to say the opposite. However, there is not really a confusion here. First, Gibbs was thinking of the correlations that would emerge between system components when coupled together. Here, we intentionally did not couple the components. Sequels address this. Second, at root, the issue turns on an ambiguous vocabulary for describing randomness and structure. Here, at least, by distinguishing between "randomness" in terms of Shannon's notion of the flatness of a probability distribution and "structure" in terms of statistical complexity, we shed some light on these important and still evolving issues.

Second, the HMSP construction procedure here gives a rather direct picture of one kind of hierarchical organization in how a stochastic process can be built from other processes. The constructive procedure uses the mixed state presentation. And, this generates a new kind of hierarchy that emerges due to the diverse combinatorial relationships between the components' internal organizations. Other related hierarchies can be similarly constructed; such as when using generalized hidden Markov models [37] as ergodic components.

Third and finally, modern statistical inference has been treated to a number of formalizations of general learning that make minimal assumptions. Consider for example:

1. *Universal Priors* [38–40]: In the computation-theoretic approach to modeling and statistical inference there are attempts to define a most-general prior over model space. However, these raise very natural questions, What kind of process would generate such a prior? Moreover, what kinds of difficulties are there in detecting processes drawn according to such a prior.

2. *No Free Lunch Theorem* [41]: This framing makes a number of implicit assumptions about the measure on the Process Urn simplex. Does the theorem hold? Not when you consider structure.

3. *Probably Almost Correct Learning* [42]: This "distribution-free" approach is a bold attempt within machine learning to identify the computational nature of evolution and learning. However, is not this the same thing as assuming any process is possible? If so, then it is analogous to assuming the Mother of All Processes. That is, rather than being "distribution-free" the assumption underlying PAC learning is "distribution-full".

In light of these, The Mother of All Processes suggests a construction for such assumption-free or minimal-assumption modeling. In this, one is sampling from the space of all processes and exploits the ϵ-machine representation to be specific about probability, on the one hand, and structure, on the other. The realization resulting from this was that the preceding development was able to demonstrate that transient-state structure made explicit the challenges in detecting component processes and that this was captured informationally via the transient information.

---

[1] E. T. Jaynes. The Gibbs paradox? In C. R. Smith, G. J. Erickson, and P. O. Neudorfer, editors, *Maximum Entropy and Bayesian Methods*, pages 1–22.

Kluwer Academic Publishers, Dordrecht, Holland, 19. 1, 20

[2] H. B. Callen. *Thermodynamics and an Introduction to Thermostatistics*. Wiley, New York, second edition, 1985a. 1

[3] D. Kondepudi. *Introduction to Modern Thermodynamics*. Wiley, Chichester, 2008. 1

[4] T. L. Lai. Sequential changepoint detection in quality control and dynamical systems. *J. Roy. Stat. Soc. Series B*, 57(4):613–658, 1995. 1

[5] R. Abraham and C. Shaw. *Dynamics: the Geometry of Behavior*. Addison-Wesley, second edition, 1992. 1

[6] S. Smale. Differentiable dynamical systems. *Bull. Amer. Math. Soc.*, 73:797–817, 1967. 1

[7] R. Bowen. One-dimensional hyperbolic sets for flows. *J. Differential Equations*, 12:173–179, 1972. 1

[8] J. E. Hanson and J. P. Crutchfield. The attractor-basin portrait of a cellular automaton. *J. Stat. Phys.*, 66:1415 – 1462, 1992. 2

[9] D. P. Varn, G. S. Canright, and J. P. Crutchfield. Discovering planar disorder in close-packed structures from X-Ray diffraction: Beyond the fault model. *Phys. Rev. B*, 66(17):174110–3, 2002. 2

[10] D. P. Varn, G. S. Canright, and J. P. Crutchfield. Inferring pattern and disorder in close-packed structures via ϵ-machine reconstruction theory: Structure and instrinsic computation in zinc sulphid. *Acta. Cryst. Sec. B*, 63(2):169–182, 2002. 2

[11] J. P. Crutchfield and E. van Nimwegen. The evolutionary unfolding of complexity. In L. F. Landweber and E. Winfree, editors, *Evolution as Computation: DIMACS Workshop, Princeton, 1999*, Natural Computing Series, page to appear, New York, 2001. Springer-Verlag. 2

[12] J. P. Crutchfield, C. J. Ellison, and J. R. Mahoney. Time's barbed arrow: Irreversibility, crypticity, and stored information. *Phys. Rev. Lett.*, 103(9):094101, 2009. 2, 10

[13] C. J. Ellison, J. R. Mahoney, and J. P. Crutchfield. Prediction, retrodiction, and the amount of information stored in the present. *J. Stat. Phys.*, 136(6):1005–1034, 2009. 2, 4, 10

[14] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, second edition, 2006. 2, 6

[15] J. P. Crutchfield and D. P. Feldman. Regularities unseen, randomness observed: Levels of entropy convergence. *CHAOS*, 13(1):25–54, 2003. 2, 3, 16

[16] S. Loomis and J. P. Crutchfield. Topology, convergence, and reconstruction of predictive states. *Physica D*, 445, 2021. 2

[17] J. P. Crutchfield and K. Young. Inferring statistical complexity. *Phys. Rev. Let.*, 63:105–108, 1989; J. P. Crutchfield, *Physica D* **75** 11–54, 1994; J. P.

Crutchfield and C. R. Shalizi, *Phys. Rev. E* **59**(1) 275–283, 1999. 3

[18] N. Travers and J. P. Crutchfield. Equivalence of history and generator ϵ-machines. *Symmetry*, 17(78):1–29, 2024. 3

[19] J. R. Mahoney, C. J. Ellison, R. G. James, and J. P. Crutchfield. How hidden are hidden processes? a primer on crypticity and entropy convergence. *CHAOS*, 21(3):037112, 2011. 3

[20] 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. 3

[21] B. Kitchens and S. Tuncel. Finitary measures for subshifts of finite type and sofic systems. *Memoirs of the AMS*, 58:no. 338, 1985. 4

[22] C. S. McTague and J. P. Crutchfield. Automated pattern discovery—An algorithm for constructing optimally synchronizing multi-regular language filters. *Theoretical Computer Science*, 359(1-3):306–328, 2006. 5

[23] C. R. Shalizi and J. P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *J. Stat. Phys.*, 104:817–879, 2001. 5

[24] R. M. Gray and L. D. Davisson. The ergodic decomposition of stationary discrete random processses. *IEEE Trans. Info. Th.*, 20(5):625–636, 1974. 6

[25] L. Debowski. Ergodic decomposition of excess entropy and conditional mutual information. *Technical Report*, 2007. 7

[26] L. Debowski. A general definition of conditional information and its application to ergodic decomposition. *Stat. Prob. Lett.*, 79(9):1260–1268, 2009. 7

[27] W. Lohr. Properties of the statistical complexity functional and partially deterministic hmms. *Entropy*, 11(3):385–401, 2009. 7

[28] D. Lind and B. Marcus. *An Introduction to Symbolic Dynamics and Coding*. Cambridge University Press, New York, 1995. 15

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

[30] B. D. Johnson, J. P. Crutchfield, C. J. Ellison, and C. S. McTague. Enumerating finitary processes. *Entropy*, 26(12):1105, 2024. 18

[31] N. Travers and J. P. Crutchfield. Exact synchronization for finite-state sources. *J. Stat. Phys.*, 145(5):1181–1201, 2011. 18

[32] N. Travers and J. P. Crutchfield. Asymptotic synchronization for finite-state sources. *J. Stat. Phys.*, 145(5):1202–1223, 2011. 18

[33] A. Jurgens and J. P. Crutchfield. Divergent predictive states: The statistical complexity dimension of stationary, ergodic hidden Markov processes. *Chaos*, 31(8):0050460, 2021. 18, 19

[34] C. H. Bennett. On the nature and origin of complexity in discrete, homogeneous locally-interacting systems. *Found. Phys.*, 16:585–592, 1986. 19

[35] C. E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423, 623–656, 1948. 19

[36] J. W. Gibbs. On the equilibrium of heterogeneous substances. *Trans. Connecticut Acad. Arts Sci.*, 3:108–248, 343–524, 1876. 20

[37] D. R. Upper. *Theory and Algorithms for Hidden Markov Models and Generalized Hidden Markov Models.* PhD thesis, University of California, Berkeley, 1997. Published by University Microfilms Intl, Ann Arbor, Michigan. 20

[38] J. Rissanen. Universal coding, information, prediction, and estimation. *IEEE Trans. Info. Th.*, IT-30:629, 1984. 20

[39] J. Rissanen. *Stochastic Complexity in Statistical Inquiry.* World Scientific, Singapore, 1989.

[40] M. Li and P. M. B. Vitanyi. *An Introduction to Kolmogorov Complexity and its Applications.* Springer-Verlag, New York, 1993. 20

[41] D. H. Wolpert and W. G. Macready. No free lunch theorems for optimization. *IEEE Trans. Evol. Comp.*, 1:67–82, 1997. 20

[42] L. Valiant. A theory of the learnable. *Comm. ACM*, 27(11):1134–1142, 1984. 20