Santa Fe Institute Working Paper 10-08-015
arxiv.org:1007.5354 [cond-mat.stat-mech]

Synchronization and Control in Intrinsic and Designed Computation: An Information-Theoretic Analysis of Competing Models of Stochastic Computation

James P. Crutchfield,[1,2,*] Christopher J. Ellison,[2,†] Ryan G. James,[2,‡] and John R. Mahoney[2,§]

[1] *Complexity Sciences Center and Physics Department, University of California at Davis, One Shields Avenue, Davis, CA 95616*
[2] *Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501*

(Dated: August 23, 2010)

We adapt tools from information theory to analyze how an observer comes to synchronize with the hidden states of a finitary, stationary stochastic process. We show that synchronization is determined by both the process's internal organization and by an observer's model of it. We analyze these components using the convergence of state-block and block-state entropies, comparing them to the previously known convergence properties of the Shannon block entropy. Along the way, we introduce a hierarchy of information quantifiers as derivatives and integrals of these entropies, which parallels a similar hierarchy introduced for block entropy. We also draw out the duality between synchronization properties and a process's controllability. The tools lead to a new classification of a process's alternative representations in terms of minimality, synchronizability, and unifilarity.

Keywords: controllability, synchronization information, stored information, entropy rate, statistical complexity, excess entropy, crypticity, information diagram, presentation, minimality, gauge information, oracular information

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

CONTENTS

*chaos@ucdavis.edu
†cellison@cse.ucdavis.edu
‡rgjames@ucdavis.edu
§jrmahoney@ucdavis.edu

Nonlinear dynamical systems store and generate information—they *intrinsically compute*. Real computing devices use nonlinearity to do the same, except that they are *designed to compute*—the information serves some utility or function determined by the designer. Intuitively, useful computing devices must be constructed out of (physical, chemical, or biological) processes that have some minimum amount of intrinsic computational capability. However, the exact relationship between intrinsic and designed computation remains elusive. In fact, bridging intrinsic and designed computation requires solving a number of intermediate problems. One is to understand the diversity of intrinsic computations of which nonlinear dynamical systems are capable. Another is to determine if one can practically manipulate these systems in the service of functional information generation and storage.

Here, we address both of these problems from the perspective of information theory. We describe new information processing characteristics of dynamical systems and the stochastic processes they generate. We focus particularly on two key aspects that impact design: synchronization and control. Synchronization concerns how we come to know the hidden states of a process through observations; while control concerns how we manipulate a process into a desired internal condition.

I. INTRODUCTION

Given a model of a stationary stochastic process, how much information must one extract from observations to exactly know which state the process is in? With this, an observer is said to be synchronized to the process. (For an introduction to the problem, see Ref. [1].)

Given that one has designed a stochastic process, is there a series of inputs that reliably drive it to a desired internal condition? If so, the designed process is said to be controllable.

Synchronization and control are dual to each other: In synchronization, an observer attempts to predict the process's internal state from incomplete and indirect observations, typically starting with complete ignorance and hopefully ending with complete certainty. In control, one must extract from the design a series of manipulations, typically indirect, that will drive the process to a desired state or set of states. The duality is simply that the observer's measurements can be interpreted as the designer's control inputs.

Synchronization and control are key aspects in intrinsic and designed computation, both for detecting intrinsic computation in dynamical systems and for leveraging a dynamical system's intrinsic computation into useful computation. For the latter, the circuit designer attempts to build circuits, themselves dynamical systems, that synchronize to incoming signals.

For example, even the most mundane initial operation is essential: When power is first applied, a digital computer must predictably reach a stable and repeatable state, without necessarily being able to perform even small amounts of digital intelligent control or analysis of its changing environment. Without reliably reaching a stable condition—now a quite elaborate operation in modern microprocessors—no useful information processing can be initiated. The device is still a dynamical system, of course, but it fails at raising itself from that prosaic condition to the level of a computing device.

Once digital computing operations have commenced, similar concerns arise in the timing and control of information being loaded from memory into a register. Not only must each data bus line synchronize properly or risk misconstruing the voltage level offered up by the wires, but this must happen simultaneously across a number of component devices— busses as wide as 128 or 256 lines are not uncommon today.

Stepping back a bit, one must wonder what tools dynamical systems theory itself provides to analyze and design computation. Indeed, many of the properties often used to characterize and classify dynamical systems are time-asymptotic—the Kolmogorov-Sinai entropy or Shannon entropy rate, the spectrum of Lyapunov characteristic exponents, the fractal and information dimensions (which rely on the asymptotic invariant measure), come to mind. However, real computing is not asymptotic. Individual logic gates, as dynamical systems, deliver their results on the short term. Indeed, the faster they do this, the better.

How can we bridge the gap between dynamical systems theory and the need to characterize the short term properties of dynamical systems? A suggestive example is found in the analysis of escape rates [2], a property of transient, short-term behavior. Another answer is found in synchronization and controllability, as they too are properties of the short term behavior of dynamical systems. We will show that there is a connection between these properties and the more typical asymptotic view of dynamical behavior: Synchronization and control are determined by the nature of convergence to the asymptotic—this convergence will be explored.

Given the duality between synchronization and control, in the following we present results in terms of only one notion—synchronization. The results apply equally

well to control, though with different interpretations.

A. Precis

Analyzing informational convergence properties is the main strategy we will use. However, as we will see, different properties have a variety of convergence behaviors. Moreover, we will consider a variety of representations for any given process. The result, while giving insight into informational properties and how representations can distort them, ends up being a rather elaborate classification scheme. To reduce the apparent complication, it will be helpful to give a detailed summary of the steps we employ in the development.

After describing related work, we review the use of Shannon block entropy and related quantities, analyzing their asymptotic behavior and aspects of convergence. We introduce a single framework—the convergence hierarchy—to call out the systematic nature of convergence properties.

We then take a short tour introducing the range of possible descriptions a process can have, noting their defining properties. One description, the ϵ-machine, plays a particularly central role, as it allows one to calculate all of a process's intrinsic properties. Other descriptions typically do not allow this to the same broad extent.

With a model in hand, one can start to discuss how one synchronizes to its states. When the model is the ϵ-machine, one can speak of synchronizing to the process itself. To do this, we analyze the convergence properties of two new entropies: the state-block entropy and the block-state entropy. We establish their general asymptotic properties, introducing convergence hierarchies of their own, paralleling that for the block entropy. For finitary processes, the latter converges from below, but the new block-state entropy converges from above to the same asymptote. One benefit is that estimation methods can be improved through use of bounds from above and below.

When we specialize to the ϵ-machine, we establish a direct connection between synchronization and convergence of block entropies. We provide an informational measure—synchronization information—that summarizes the total uncertainty encountered during synchronization. We relate this back to the transient information introduced previously, which derives only from the observed sequences, not requiring a model or a notion of state. Along the way, we discuss a process's Markov order—the scale at which "asymptotic" statistics set in—and its cryptic order—the length scale over which internal state information is spread. These scales control synchronization.

The development then, step-by-step, relaxes the ϵ-machine's defining properties in order to explore an increasingly wide range of models. A particular emphasis in this is to show how nonoptimal models bias estimates of a process's informational properties. Conversely, we learn how certain classes of models, some widely used in mathematical statistics and elsewhere, make strong assumptions and, in some cases, preclude the estimation of important process properties.

Starting with the class of minimal, optimally predictive models that synchronize (finitary ϵ-machines), we first relax the minimality assumption. We show that needless model elaborations—such as more, but redundant states—can affect synchronization. We identify that class which still does synchronize. Then, we consider nonminimal unifilar, nonsynchronizing models. Finally, we relax the unifilarity assumption. At each stage, we see how the convergence properties of the various entropies change. These changes, in turn, induce a number of informational measures of what the models themselves contribute to a process's now largely-apparent information processing.

A key tool in the analysis takes advantage of the fact that the various multivariable information quantities form a signed measure [3]. Their visual display, a form of Venn diagram called an information diagram, brings some order to the notation and classification chaos.

B. Synchronization and Control: Related Work

Controlling dynamical systems and stochastic processes has an extensive history. For linear dynamical systems see, for example, Ref. [4] and for hidden Markov models see, for example, Ref. [5]. More recently, there has been much work on controlling nonlinear dynamical systems, a markedly more difficult problem in its full generality; see Refs. [6–8].

Synchronization, too, has been very broadly studied and for much longer, going back at least to Huygens [9]. It is also an important property of symbolic dynamical systems [10]. It has even become quite popularized of late, being elevated to a general principle of natural organization [11].

Here, we consider a form of synchronization that is, at least at this point, different from the dynamical kind. Moreover, we take a complementary, but distinct approach—that of information theory—to address control and synchronization. This was introduced in Ref. [12] and several applications are given in Refs. [1, 13]. A roughly similar problem setting for synchronization is found in Ref. [14]. We note that the closely related topics of state estimation and control are addressed in information theory [15, 16], nonlinear dynamics [17–19], and

Markov decision processes [20].

Adapting the present approach to continuous dynamical systems and stochastic processes remains a future effort. For the present, the closest connections found will be to the works cited above on hidden Markov models and symbolic dynamical systems.

II. BLOCK ENTROPY AND ITS CONVERGENCE HIERARCHY

It is an interesting fact, perhaps now intuitive, that to estimate even the randomness of an information source, one must also estimate its internal structure. Reference [12] gives a review of this interdependence and it serves as a starting point for our analysis of synchronization, which is a question about coming to know the source's states from observations. Indeed, if one has to make estimates of internal organization just to get to randomness, then one, in effect and without too much more effort, can also address issues of synchronization. This is an intimate relationship that we hope to establish.

We briefly review Ref. [12], largely to introduce notation and highlight the main ideas needed for synchronization. This review and our development of synchronization requires the reader to be facile with information theory at the level of the first half of Ref. [21], signed information measures and information diagrams of Ref. [3], and their uses in Refs. [22–24].

A. Stationary Stochastic Processes

The approach in Ref. [12] starts simply: Any stationary process $\mathcal{P}$ is a joint probability distribution $\Pr(\overleftarrow{X}, \overrightarrow{X})$ over past and future observations. This distribution can be thought of as a *communication channel* with a specified input distribution $\Pr(\overleftarrow{X})$. It transmits information from the *past* $\overleftarrow{X} = \ldots X_{-3}X_{-2}X_{-1}$ to the *future* $\overrightarrow{X} = X_0X_1X_2\ldots$ by storing it in the present. X_t is the random variable for the measurement outcome at time t; the lowercase x_t denotes a particular value. Throughout, we always use $\overleftarrow{X}$ and $\overrightarrow{X}$ in the limiting sense. That is, we work with length-L sequences or *blocks* of random variables $X_t^L = X_tX_{t+1}\cdots X_{t+L-1}$ and take the limit as L approaches infinity.

In the following, we consider only discrete measurement outcomes—$x \in \mathcal{A} = \{1, 2, \ldots, k\}$—and stationary processes—$\Pr(X_t^L) = \Pr(X_0^L)$, for all times t and block lengths L. Unlike some definitions of stationarity, this makes no assumptions about the process's internal starting conditions, as such knowledge obviates the very question of synchronization.

Such processes include those found in the field of stochastic processes, of course, but one also has in mind the symbolic dynamics of continuous-state continuous-time or continuous-state discrete-time dynamical systems on their invariant sets. The notions also apply equally well to one-dimensional spatial configurations of spin systems and of deterministic and probabilistic cellular automata, where one interprets the spatial coordinate as a "time".

B. Block Entropy

One measure of the diversity of length-L sequences generated by a process is its *Shannon block entropy*:

$$H(L) \equiv H[X_0^L] \tag{1}$$

$$= -\sum_{w \in \mathcal{A}^L} \Pr(w) \log_2 \Pr(w) \ , \tag{2}$$

where $w = x_0x_1\ldots x_{L-1}$ is a *word* in the set $\mathcal{A}^L$ of length-L sequences. It has units of [bits] of information. One can think of the block entropy as a kind of transform that reduces a process's distribution over the (typically infinite) number of sequences to a function of a single variable L. In this view, Ref. [12] focused on a simple question: What properties of a process can be determined solely from its $H(L)$?

C. Source Entropy Rate

One of those properties, and historically the most widely used and technologically important, is *Shannon's source entropy rate*:

$$h_\mu = \lim_{L \to \infty} \frac{H(L)}{L} \ . \tag{3}$$

The entropy rate is the irreducible unpredictability of a process's output—the intrinsic randomness left after one has extracted all of the correlational information from past observations. The difference between it and the alphabet size, $\log_2 |\mathcal{A}| - h_\mu$, indicates how much the raw measurements can be compressed. More precisely, Shannon's First Theorem states that the output sequences x^L from an information source can be compressed, without error, to Lh_μ bits [21]. Moreover, Shannon's Second Theorem gives operational meaning to the entropy rate [21]: A communication channel's capacity must be larger than h_μ for error-free transmission.

D. Excess Entropy

As noted, any process—chaotic dynamical system, spin chain, cellular automata, to mention a few—can be considered a channel that communicates its past to its fu-

ture. The messages to be transmitted in this way are the pasts which the process can generate. Thus, the "capacity" of this channel is not something that one optimizes as done in Shannon's theory to engineer channels and construct error-free encodings. Rather, we think of it as how much of the process's channel is actually used.

A process's channel utilization is another property that can be determined from the block entropy. It is called the *excess entropy* and is defined, closely following Shannon's channel capacity definition, by:

$$\mathbf{E} = I[\overleftarrow{X}; \overrightarrow{X}] \ , \tag{4}$$

where $I[Y; Z]$ is the mutual information between random variables Y and Z. It has units of [bits] and tells one how much information the output (the future) shares with the input (the past) and so measures how much information is transmitted through a, possibly noisy, channel.

E. Block Entropy Asymptotics

It has been known for quite some time now that the entropy rate and excess entropy control the asymptotic behavior ($L \to \infty$) of a *finitary* process's block entropy. Specifically, it scales according to the linear asymptote:

$$H(L) \propto \mathbf{E} + h_\mu L \ , \tag{5}$$

where

$$\mathbf{E} = \lim_{L \to \infty} \left(H(L) - L h_\mu \right) \ . \tag{6}$$

$\mathbf{E}$ is the sublinear part of $H(L)$. This gives important general insight into the block entropy's behavior. It is also quite practical, though: If $H(L)$ actually meets the asymptote at some finite sequence length R, then the process is effectively an order-R Markov chain [12, 24]: $\Pr(X_0|\overleftarrow{X}) = \Pr(X_0|X_{-R}^R)$. Interestingly, many finitary processes do not reach the asymptote at finite lengths and so cannot be recast as Markov chains of any order. Roughly speaking, they have various kinds of infinite-range correlation.

F. The Convergence Hierarchy

The study of how the block entropy converges, or does not, is a tool for classifying processes. Reference [12] showed that the entropy rate and excess entropy are merely two players in an infinite hierarchy that determines the shape of $H(L)$. The central idea is to take L-derivatives and integrals of $H(L)$.

To start, one has the block entropy difference:

$$\Delta H[X_0^L] \equiv H[X_0^L] - H[X_0^{L-1}] \ , \tag{7}$$

where Δ is the discrete derivative with respect to block length L. It is easy to see that the right-hand side is the conditional entropy $H[X_{L-1}|X_0^{L-1}]$ and that, in turn,

$$h_\mu = \lim_{L \to \infty} H[X_{L-1}|X_0^{L-1}] \tag{8}$$

$$= H[X_0|\overleftarrow{X}_0] \ , \tag{9}$$

recovering the entropy rate. It is often useful to directly refer to the length-L approximation to the entropy rate as $h_\mu(L) \equiv H[X_{L-1}|X_0^{L-1}]$. $h_\mu(L) \geq h_\mu$ and so it converges from above.

The excess entropy, for its part, controls the convergence speed, as it is the discrete integral:

$$\mathbf{E} = \sum_{L=1}^{\infty} (h_\mu(L) - h_\mu) \ . \tag{10}$$

It requires only a few steps to see that this form is equivalent to that of Eq. (4).

Following a similar strategy, the discrete integral

$$\mathbf{T} = \sum_{L=0}^{\infty} \left[\mathbf{E} + h_\mu L - H(L) \right] \tag{11}$$

measures how $H(L)$ itself reaches its linear asymptote $\mathbf{E} + h_\mu L$. $\mathbf{T}$ is called the *transient information* and it is implicated in determining the Markov order and, as we will show, synchronization.

The pattern should be clear now: At the lowest level, the transient information indicates how quickly the block entropy reaches its asymptote. Then, that asymptote grows at the rate h_μ and has y-intercept $\mathbf{E}$. It might be helpful to refer to the graphical summary of block-entropy convergence and the associated information measures given in Ref. [12, Fig. 2]. Analogous diagrams will appear shortly.

All this can be compactly summarized by introducing two operators: a derivative and an integral that operate on $H(L)$. The derivative operator at the n^{th}-level is:

$$\Delta^n H[X_0^L] \equiv \Delta^{n-1} H[X_0^L] - \Delta^{n-1} H[X_0^{L-1}] \ , \tag{12}$$

for $L \geq n = 1, 2, \ldots$ and for $L \geq n = 0$,

$$\Delta^0 H[X_0^L] \equiv H[X_0^L] \ . \tag{13}$$

The integrals are:

$$\mathcal{I}_n \equiv \sum_{L=n}^{\infty} \left[\Delta^n H[X_0^L] - \lim_{\ell \to \infty} \Delta^n H[X_0^\ell] \right] , \quad (14)$$

$n = 0, 1, 2, \ldots$. (This is a slight deviation from Ref. [12], when $n = 2$. See App. A.)

To make the connection with what we just discussed, in this notation we have:

$$h_\mu = \lim_{L \to \infty} \Delta^1 H[X_0^L] , \quad (15)$$

$$\mathbf{E} = \mathcal{I}_1 , \text{ and} \quad (16)$$

$$\mathbf{T} = -\mathcal{I}_0 . \quad (17)$$

Additionally, $\mathcal{I}_2$ is a process's *total predictability* $\mathbf{G}$ and $\Delta^2 H[X_0^L]$ is its predictability gain—the rate at which predictions improve by examining statistics of longer sequences.

The two operators, Δ_n and $\mathcal{I}_n$, define the *entropy convergence hierarchy* for a process, capturing those properties reflected in the process's block entropy. Given a process's specification, one attempts to calculate the hierarchy analytically; given data, to estimate it empirically. In addition to systematizing a process's informational properties, the hierarchy has a number of uses. For example, structural classes of processes can be distinguished by the n^* at which the hierarchy becomes trivial; for example, when $\Delta^n H[X_0^L] = 0,\ n > n^*$. Other classifications turn on bounded $\mathcal{I}_{n^*}$. The finitary processes, for example, are defined by $n^* = 1$: $\mathcal{I}_1 = \mathbf{E} < \infty$. Or, conversely, there are well known processes for which some integrals diverge; they include the onset of chaos through period-doubling, where the excess entropy diverges. Reference [12, Sec. VII.A] introduces a classification of processes along these lines.

III. PROCESS PRESENTATIONS

A. The Causal State Representation

Prediction is closely allied to the view of a process as a communication channel: We wish to predict the future using information from the past. At root, a prediction is probabilistic, specified by a distribution of possible futures $\overrightarrow{X}$ given a particular past $\overleftarrow{x}$: $\Pr(\overrightarrow{X}|\overleftarrow{x})$. At a minimum, a good predictive model needs to capture *all* of the information I shared between the past and future: $\mathbf{E} = I[\overleftarrow{X}; \overrightarrow{X}]$.

Consider now the goal of modeling—building a representation that allows not only good prediction but also expresses the mechanisms producing a system's behavior. To build a model of a structured process (a memoryful channel), computational mechanics [25] introduced an equivalence relation $\overleftarrow{x} \sim \overleftarrow{x}'$ that clusters all histories which give rise to the same prediction:

$$\epsilon(\overleftarrow{x}) = \{\overleftarrow{x}' : \Pr(\overrightarrow{X}|\overleftarrow{x}) = \Pr(\overrightarrow{X}|\overleftarrow{x}')\} . \quad (18)$$

In other words, for the purpose of forecasting the future, two different pasts are equivalent if they result in the same prediction. The result of applying this equivalence gives the process's *causal states* $\boldsymbol{\mathcal{S}} = \Pr(\overleftarrow{X}, \overrightarrow{X})/\sim$, which partition the space $\overleftarrow{\mathbf{X}}$ of pasts into sets that are predictively equivalent. So $\epsilon(\overleftarrow{x})$ is an equivalence class, and we call it a causal state. The set of causal states [26] can be discrete, fractal, or continuous; see, e.g., Figs. 7, 8, 10, and 17 in Ref. [27]. There is a unique start state.

State-to-state transitions are denoted by matrices $T^{(x)}_{\mathcal{S}\mathcal{S}'}$ whose elements give the probability $\Pr(X = x, \mathcal{S}'|\mathcal{S})$ of transitioning from one state $\mathcal{S}$ to the next $\mathcal{S}'$ on seeing measurement x. The resulting model, consisting of the causal states and transitions, is called the process's *ϵ-machine*. Given a process $\mathcal{P}$, we denote its ϵ-machine by $M(\mathcal{P})$.

Causal states have a Markovian property that they render the past and future statistically independent; they *shield* the future from the past [28]:

$$\Pr(\overleftarrow{X}, \overrightarrow{X}|\mathcal{S}) = \Pr(\overleftarrow{X}|\mathcal{S}) \Pr(\overrightarrow{X}|\mathcal{S}) . \quad (19)$$

Moreover, they are optimally predictive [25] in the sense that knowing which causal state a process is in is just as good as having the entire past: $\Pr(\overrightarrow{X}|\mathcal{S}) = \Pr(\overrightarrow{X}|\overleftarrow{X})$. In other words, causal shielding is equivalent to the fact [28] that the causal states capture all of the information shared between past and future: $I[\mathcal{S}; \overrightarrow{X}] = \mathbf{E}$.

ϵ-Machines have an important structural property called *unifilarity* [25, 29]: From the start state, each symbol sequence corresponds to exactly one sequence of causal states [30]. The importance of unifiliarity, as a property of any model, is reflected in the fact that representations without unifilarity, such as generic hidden Markov models, *cannot* be used to directly calculate important system properties—including the most basic, such as h_μ—how random a process is. Practically, though, unifilarity is easy to verify: For each state, each measurement symbol appears on at most one outgoing transition [31]. Thus, the signature of unifilarity is that on knowing the current state $\mathcal{S}_t$ and measurement X_t, the uncertainty in the next state $\mathcal{S}_{t+1}$ vanishes: $H[\mathcal{S}_{t+1}|\mathcal{S}_t, X_t] = 0$.

Out of all optimally predictive models $\widehat{\boldsymbol{\mathcal{R}}}$—for which $I[\widehat{\mathcal{R}}; \overrightarrow{X}] = \mathbf{E}$—the ϵ-machine captures the minimal amount of information that a process must store in or-

der to communicate all of the excess entropy from the past to the future. This is the Shannon information contained in the causal states—the *statistical complexity* [28]: $C_\mu \equiv H[\mathcal{S}] \leq H[\widehat{\mathcal{R}}]$. It turns out that statistical complexity upper bounds the excess entropy [25, 29]: $\mathbf{E} \leq C_\mu$. In short, $\mathbf{E}$ is the effective information transmission rate of the process, viewed as a channel, and C_μ is the memory stored in that channel.

Combined, these properties mean that the ϵ-machine is the basis against which modeling should be compared, since it captures all of a process's information at maximum representational efficiency.

Importantly, due to unifilarity one can calculate the entropy rate directly from a process's ϵ-machine:

$$\begin{aligned} h_\mu &= H[X|\mathcal{S}] \\ &= -\sum_{\{\mathcal{S}\}} \Pr(\mathcal{S}) \sum_{\{x\mathcal{S}'\}} T^{(x)}_{\mathcal{S}\mathcal{S}'} \log_2 \sum_{\{\mathcal{S}'\}} T^{(x)}_{\mathcal{S}\mathcal{S}'} \ . \end{aligned} \tag{20}$$

$\Pr(\mathcal{S})$ is the asymptotic probability of the causal states, which is obtained as the normalized principal (left) eigenvector of the transition matrix $T = \sum_{\{x\}} T^{(x)}$. A process's statistical complexity can also be directly calculated from its ϵ-machine:

$$\begin{aligned} C_\mu &= H[\mathcal{S}] \\ &= -\sum_{\{\mathcal{S}\}} \Pr(\mathcal{S}) \log_2 \Pr(\mathcal{S}) \ . \end{aligned} \tag{21}$$

Thus, the ϵ-machine directly gives two important properties: a process's rate (h_μ) of producing information and the amount (C_μ) of historical information it stores in doing so. Moreover, Refs. [22, 23] showed how to calculate a process's excess entropy directly from the ϵ-machine.

B. General Presentations

The ϵ-machine is only one possible description of a process. There are many alternatives: Some larger, some smaller; some with the same prediction error, some with larger prediction error; some that are unifilar, some not; some that do an excellent job of capturing $\Pr(\overleftarrow{X}, \overrightarrow{X})$, many (or most) doing only an approximate job; some allowing for the direct calculation of the process's properties, some precluding such calculations.

The ϵ-machine, compared to all other possible descriptions, is arguably the best. The results in the following, as an ancillary benefit, strengthen this conclusion considerably showing in which ways it is preferred. However, it is important to keep in mind that due to implementation constraints or intended use or specified performance criteria, alternative models may be desirable and preferred to the ϵ-machine. Reference [27], for example, compares the benefits and disadvantages of different kinds of nonunifilar models that are smaller than the ϵ-machine. We return to elaborate on this in Sec. VII D.

One refers to a process's possible descriptions as *presentations* [32]. Specifically, these are state-based models—using states and state transitions—that exactly describe $\Pr(\overleftarrow{X}, \overrightarrow{X})$. That is, given a finitary process $\mathcal{P}$, we consider the set of all presentations that generate the same *process language*: $\{(w, \Pr(X^L = w)) : w \in \mathcal{A}^L, L = 1, 2, \ldots\}$. The set of $\mathcal{P}$'s presentations is the focus of our work here. (That is, we do not address models that give only approximations to the process language. For this, see Ref. [33].)

To be consistent with historical usage, we also refer to presentations as *rivals*. A rival consists of a set $\boldsymbol{\mathcal{R}}$ of states and state-to-state transitions $T^{(s)}_{\mathcal{R}\mathcal{R}'}$ over the symbols s in the process's measurement alphabet $\mathcal{A}$. There is an associated mapping $\eta : \overleftarrow{x} \to \boldsymbol{\mathcal{R}}$ from pasts to rival states. Generally, this mapping is multivalued in the sense that a past can map to multiple rival states [34]. When we refer to the rival's state as a random variable, we will denote this as $\mathcal{R}$. We use lower case ρ when we refer to a particular realization: $\mathcal{R} = \rho$, $\rho \in \boldsymbol{\mathcal{R}}$. Just as with the ϵ-machine, given a rival presentation, we can refer to the amount of information the rival states contain—this is the *presentation state entropy* $H[\mathcal{R}]$.

Above, we noted that a process's ϵ-machine is its minimal unifilar presentation. But, how are the rivals related, if at all, to the ϵ-machine? To explore the organization of the space of rivals, in the following we relax those properties that make the ϵ-machine unique, working with presentations that are nonminimal unifilar and those that are not even unifilar. And so, we must distinguish several kinds of presentations. First, we define unifilarity.

Definition 1. *A presentation is* unifilar *if and only if* $H[\mathcal{R}_{t+1}|\mathcal{R}_t, X_t] = 0$.

Second, we introduce the notion of reverse-time unifilarity.

Definition 2. *A presentation is* counifilar *if and only if* $H[\mathcal{R}_t|X_t, \mathcal{R}_{t+1}] = 0$.

Third, we will consider prescient presentations, those whose states are as good at predicting as the ϵ-machine's causal states [28, 29].

Definition 3. *A presentation is* prescient *if and only if, for every past* $\overleftarrow{x} \in \overleftarrow{\mathbf{X}}$ *and every* $\rho \in \eta(\overleftarrow{x})$*:*

$$\Pr(\overrightarrow{X}^L|\mathcal{R} = \rho) = \Pr(\overrightarrow{X}^L|\mathcal{S} = \epsilon(\overleftarrow{x})) \tag{22}$$

for all $L \geq 1, 2, 3, \ldots$.

We will also shortly discuss presentations to which one

can or cannot synchronize—that are or are not controllable.

IV. STATE-BLOCK AND BLOCK-STATE ENTROPIES

Now, we introduce two block entropies and discuss their properties, but first, we recall some well known results from information theory [21, Sec. 4.2].

For any stationary stochastic process, $\Delta H[X_0^L]$ is a nonincreasing sequence of nonnegative terms that converges, from above, to the entropy rate h_μ. There is a complementary result which provides an estimate of the entropy rate that converges from below. It is typically stated in terms of the Moore (state-output) type of hidden Markov model [21, Thm. 4.5.1], so we recast the theorem in terms of the Mealy (edge-output) type of hidden Markov models, used exclusively here.

Theorem 1. *If $\mathcal{R}_0, \mathcal{R}_1, \ldots$ form a stationary Markov chain and $(X_i, \mathcal{R}_{i+1}) = \phi(\mathcal{R}_i)$, then*

$$H[X_L|\mathcal{R}_0, X_0^L] \leq h_\mu \leq H[X_L|X_0^L] \ , \tag{23}$$

$L = 0, 1, 2, \ldots$, and

$$H[X_\infty|\mathcal{R}_0, \overrightarrow{X}_0] = h_\mu \ . \tag{24}$$

ϕ need not be a deterministic mapping.

Appendix B provides the proof details. Henceforth, we refer to $H[\mathcal{R}_0, X_0^L]$ as the *state-block entropy*.

We also define the *block-state* entropy to be $H[X_0^L, \mathcal{R}_L]$. As with the state-block entropy, there is a corresponding convergence result.

Theorem 2. *If $\mathcal{R}_0, \mathcal{R}_1, \ldots$ form a stationary Markov chain and $(X_i, \mathcal{R}_{i+1}) = \phi(\mathcal{R}_i)$, then*

$$H[X_0^L, \mathcal{R}_L] - H[X_0^{L-1}, \mathcal{R}_{L-1}] \leq h_\mu \leq H[X_L|X_0^L] \ , \tag{25}$$

$L = 1, 2, 3, \ldots$, and

$$\lim_{L\to\infty} \Big(H[X_0^L, \mathcal{R}_L] - H[X_0^{L-1}, \mathcal{R}_{L-1}] \Big) = h_\mu \ . \tag{26}$$

Again, ϕ need not be a deterministic mapping.

Ref. [35] provides the proof of this theorem and discusses related results in the context of crypticity and cryptic order [24].

Note, both of these theorems hold for general presentations—not just ϵ-machines—and this fact serves as the motivation for our later generalizations.

A. Convergence Hierarchies

Just as with the block entropy $H[X_0^L]$, we will consider L-derivatives and integrals of the state-block and block-state entropies. At the first level,

$$\Delta H[\mathcal{R}_0, X_0^L] \equiv H[\mathcal{R}_0, X_0^L] - H[\mathcal{R}_0, X_0^{L-1}] \ , \tag{27}$$

$$\Delta H[X_0^L, \mathcal{R}_L] \equiv H[X_0^L, \mathcal{R}_L] - H[X_0^{L-1}, \mathcal{R}_{L-1}] \ . \tag{28}$$

Higher-order derivatives are defined similarly to Eq. (12). As before, the $n = 0$ case is an identity operator. So, for example, $\Delta^0 H[\mathcal{R}_0, X_0^L] = H[\mathcal{R}_0, X_0^L]$.

We already know—Thms. 1 and 2—that both of these quantities tend to h_μ in the large-L limit, ensuring that all higher-order derivatives tend to zero.

Now, consider the n^{th} state-block and block-state integrals:

$$\mathcal{K}_n = \sum_{L=n}^{\infty} \big(\Delta^n H[\mathcal{R}_0, X_0^L] - \lim_{\ell\to\infty} \Delta^n H[\mathcal{R}_0, X_0^\ell] \big) \ , \tag{29}$$

$$\mathcal{J}_n = \sum_{L=n}^{\infty} \big(\Delta^n H[X_0^L, \mathcal{R}_L] - \lim_{\ell\to\infty} \Delta^n H[X_0^\ell, \mathcal{R}_\ell] \big) \ . \tag{30}$$

Note that both $\mathcal{K}_0 \geq 0$ and $\mathcal{J}_0 \geq 0$ while, in contrast, $\mathcal{I}_0 \leq 0$. Also, $\mathcal{K}_1 \leq 0$ and $\mathcal{J}_1 \leq 0$ while $\mathcal{I}_1 \geq 0$. These differences are due to the fact that the block entropy is concave in L while the state-block and block-state entropies are convex.

Consider the partial sums of $\mathcal{K}_1$—the state-block integral:

$$\begin{aligned} \mathcal{K}_1(L) &= \sum_{\ell=1}^{L} \big(\Delta H[\mathcal{R}_0, X_0^\ell] - h_\mu \big) \\ &= H[\mathcal{R}_0, X_0^L] - H[\mathcal{R}_0, X_0^0] - Lh_\mu \\ &= H[X_0^L|\mathcal{R}_0] - Lh_\mu \ . \end{aligned} \tag{31}$$

Note that if the presentation is unifilar, then $H[X_0^L|\mathcal{R}_0] = Lh_\mu$ and $\mathcal{K}_1(L) = 0$. Thus, unifilarity is a sufficient condition for $\mathcal{K}_1 = 0$, but it is not a necessary condition.

Now, consider the partial sums of $\mathcal{J}_1$—the block-state integral:

$$\begin{aligned} \mathcal{J}_1(L) &= \sum_{\ell=1}^{L} \big(\Delta H[X_0^\ell, \mathcal{R}_\ell] - h_\mu \big) \\ &= H[X_0^L, \mathcal{R}_L] - H[X_0^0, \mathcal{R}_0] - Lh_\mu \\ &= H[X_0^L, \mathcal{R}_L] - H[X_L^0, \mathcal{R}_L] - Lh_\mu \\ &= H[X_0^L|\mathcal{R}_L] - Lh_\mu \ . \end{aligned} \tag{32}$$

Similarly, if the presentation is counifilar, then it follows

that $H[X_0^L|\mathcal{R}_L] = 0$ and $\mathcal{J}_1(L) = 0$. So, counifilarity is a sufficient condition for $\mathcal{J}_1 = 0$, but it is not a necessary condition.

B. Asymptotics

Theorems 1 and 2 tell us $H[\mathcal{S}_0, X_0^L]$ and $H[X_0^L, \mathcal{S}_L]$ are convex functions in L and that the slope limits to the entropy rate. This means that each curve converges to a linear asymptote, cf. Eq. (5):

$$H[\mathcal{R}_0, X_0^L] \propto Y_{\text{SBE}} + h_\mu L \tag{33}$$

$$H[X_0^L, \mathcal{R}_L] \propto Y_{\text{BSE}} + h_\mu L \ , \tag{34}$$

where Y_{SBE} and Y_{BSE} are constants independent of L. The pictures that one should have in mind for the growth of these entropies are those of Figs. 1, 5, 8, 11, and 14, which we will discuss in due course.

In fact, we will take this behavior as the definition of the following linear asymptotes:

$$\begin{aligned} Y_{\text{SBE}} &\equiv \lim_{L\to\infty} \Big(H[\mathcal{R}_0, X_0^L] - h_\mu L \Big) \\ &= \lim_{L\to\infty} \Big(H[\mathcal{R}_0] + H[X_0^L|\mathcal{R}_0] - h_\mu L \Big) \\ &= H[\mathcal{R}_0] + \mathcal{K}_1 \end{aligned} \tag{35}$$

and

$$\begin{aligned} Y_{\text{BSE}} &\equiv \lim_{L\to\infty} \Big(H[X_0^L, \mathcal{R}_L] - h_\mu L \Big) \\ &= \lim_{L\to\infty} \Big(H[\mathcal{R}_L] + H[X_0^L|\mathcal{R}_L] - h_\mu L \Big) \\ &= H[\mathcal{R}_0] + \mathcal{J}_1 \ . \end{aligned} \tag{36}$$

These tell us that $\mathcal{K}_1$ and $\mathcal{J}_1$ are not the sublinear parts of the state-block and block-state entropies. This is in contrast to the corresponding result for the block entropies:

$$Y_{\text{BE}} \equiv \lim_{L\to\infty} \Big(H[X_0^L] - h_\mu L \Big) \tag{37}$$

$$= \lim_{L\to\infty} \Big(H[X_0^0] + H[X_0^L] - h_\mu L \Big) \tag{38}$$

$$= H[X_0^0] + \mathcal{I}_1 \ . \tag{39}$$

The term $H[X_0^0]$ was dropped in the earlier partial sum formulation—i.e., Eq. (10)—since it corresponds to no measurement being made and so is zero. It is reintroduced above, though, to complete the formal parallel to the state-block and block-state entropy cases.

The result for block entropy is that the offset of the linear asymptote was equal to the $\mathcal{I}_1$, the excess entropy. However, the argument just given clearly establishes that, in fact, one should think of the first derivatives as offsets from the initial value of their corresponding curves.

Finally, recall that $\mathcal{K}_1$ and $\mathcal{J}_1$ are not greater than zero, so Y_{SBE} and Y_{BSE} are less than or equal to the presentation state entropy $H[\mathcal{R}_0]$.

V. SYNCHRONIZATION

A. Duality of Synchronization and Control

Synchronization is a question about how an observer comes to know a process's (typically hidden) current internal state through observations. (Recall the picture introduced in Ref. [1].) As such, it requires a notion of state, either the process's causal state (using the ϵ-machine) or the state of some other presentation. In either case we monitor the observer's uncertainty over the states $\boldsymbol{\mathcal{R}}$ after having seen a series of measurements $w = x_0 x_2 \ldots x_{L-1}$ using the conditional state entropy $H[\mathcal{R}|w]$. When this vanishes, the observer is synchronized and we call w a *synchronizing word*.

During synchronization, the observer updates her answer to the question, "Which presentation states can be reached by sequence w?" When there is a unique answer, the observer is synchronized. If the eventual answer, though, is only a proper subset of presentation states, then $0 < H[\mathcal{R}|w] \leq H[\mathcal{R}]$ and the observer can be said to be partially synchronized.

A formal treatment of synchronization appears in Refs. [36, 37]; here we define a related quantity.

Definition 4. *A presentation is* weakly asymptotically synchronizing *if and only if* $\lim_{L\to\infty} H[\mathcal{R}_L|X_0^L] = 0$.

While some processes can have synchronizing words, others have *synchronizing blocks* where every word of a finite length R is a synchronizing word. Such processes are called *Markov processes*. The smallest such R is the *Markov order* [24, 38]. It turns out that the ϵ-machine presentation for a Markov process is *exactly synchronizing* [36].

If a process admits a presentation that is only weakly asymptotically synchronizing, though, then an observer will be in various conditions of state uncertainty until the limit $L \to \infty$. Finitary ϵ-machines, as it turns out, are always at least weakly asymptotically synchronizing and the state uncertainty vanishes exponentially fast [37]: $\Pr\left(H[\mathcal{S}_0|X_0^L] > 0\right) \propto e^{-L}$.

The controllability properties of a process and its models are analogous. However, now there is a designer that has built an implementation of a process. And, starting from an unknown condition, the designer wishes to prepare the process in a particular state or set of states by imposing a sequence of inputs. Phrased this way, one sees

that the implementation is, in effect, a presentation and the control sequence is none other than a synchronizing word. Due to this duality, we only discuss synchronization in the bulk of our development, returning at the end to briefly draw out interpretations of the results for controllability.

B. Synchronizing to the ϵ-Machine

We noted that the ϵ-machine directly gives two important information-theoretic properties—the entropy rate (h_μ) and the statistical complexity (C_μ)—and one (the excess entropy $\mathbf{E}$) indirectly. The difference between C_μ and $\mathbf{E}$ was introduced as the *crypticity* [22, 23]

$$\chi = C_\mu - \mathbf{E} \tag{40}$$

to describe how much of the internal state information (C_μ) is not locally present in observed sequences ($\mathbf{E}$).

Synchronization, as we discussed, is a property of the recurrent portion of the ϵ-machine and since it is unifilar, if one knows its current state and follows transitions according to the word being considered, then one will always know the ϵ-machine's final state. However, it is also useful to consider the scenario when one does not know the ϵ-machine's current state. Given no other information, the best estimate for the current state is to draw from the stationary state distribution $\Pr(\mathcal{S})$. Then, as each symbol is observed, one updates this *belief* distribution and estimates the next state from this updated distribution.

As noted above, $H[\mathcal{S}_L|X_0^L]$ converges to zero exponentially fast for all ϵ-machines with a finite number of recurrent causal states. At each L before that point, there is an uncertainty in the causal state of the ϵ-machine. If we add up the uncertainty at each length, then we have the *synchronization information*:

$$\mathbf{S} \equiv \sum_{L=0}^{\infty} H[\mathcal{S}_L|X_0^L] \tag{41}$$

$$= \sum_{L=0}^{\infty} \Big(H[X_0^L, \mathcal{S}_L] - H[X_0^L] \Big) \ . \tag{42}$$

Importantly, the second line shows that synchronization information can be visualized as the sum of all differences between the block-state and the block entropy curves.

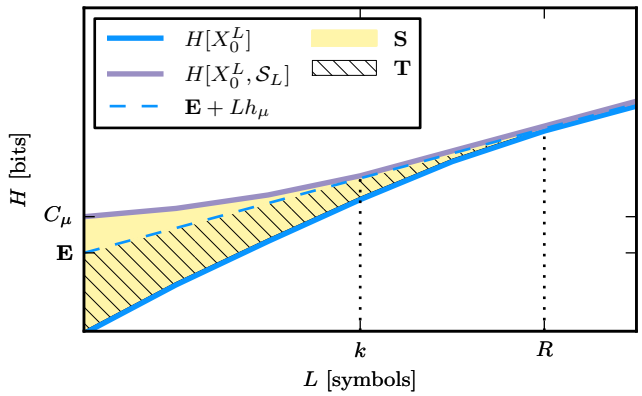

FIG. 1. Block entropy and block-state entropy growth for a generic finitary stationary process: It is easily seen that the synchronization information upper bounds the transient information, $\mathbf{T} \leq \mathbf{S}$, as $\mathbf{T}$ is a component of $\mathbf{S}$. The Markov order R and cryptic order k are also shown in their proper relationship $k \leq R$: R indicates where the block entropy meets the $\mathbf{E} + h_\mu L$ asymptote and k, where the block-state entropy meets the same asymptote.

Moreover, starting from Eq. (42) we find:

$$\mathbf{S} = \sum_{L=0}^{\infty} \Big(H[X_0^L, \mathcal{S}_L] - (\mathbf{E} + Lh_\mu) \Big) - \sum_{L=0}^{\infty} \Big(H[X_0^L] - (\mathbf{E} + Lh_\mu) \Big) \tag{43}$$

$$= \mathcal{J}_0 - \mathcal{I}_0 \ . \tag{44}$$

We know that $\mathbf{T} = -\mathcal{I}_0$. When we identify $\mathcal{J}_0$ with a separate, nonnegative information quantity we conclude immediately that $\mathbf{S} \geq \mathbf{T}$. This relationship is shown graphically in Fig. 1.

The *cryptic order* k, as defined in Ref. [24], can be interpreted as the length at which the block-state curve has converged to its asymptote: $\mathbf{E} + h_\mu L$. Surprisingly, this is not the length at which an ϵ-machine can be considered synchronized, which is given by the Markov order R. Given its definition as the smallest value L for which $H[\mathcal{S}_L|\overrightarrow{X}_0] = 0$, we see that the cryptic order can be interpreted as a measure of how far back in time the state sequence can be retrodicted using the infinite future.

For example, the Even Process consists of all bi-infinite sequences that contain even-length stretches of 1s separated by at least a single 0; see Ref. [12]. This process cannot be considered synchronized at any finite length because all the thus-far seen symbols may be 1s, and so one does not know if the latest symbol is a 1 at an even- or odd-valued location. In contrast, once a 0 has been seen, we know instantly the evenness and oddness of each preceding 1, making the cryptic order $k = 0$. Since the

cryptic order $k = 0$ for the Even Process, one concludes that $\mathcal{J}_0$ does not contribute to $\mathbf{S}$ and $\mathbf{T} = \mathbf{S}$.

The two pieces—$\mathcal{J}_0$ and $\mathcal{I}_0$—comprising $\mathbf{S}$ are both finite due to the exponentially fast convergence of the two block-entropy curves [37]. This shows that $\mathbf{S}$ consists of distinct information contributions drawn from different process features. Referring to Fig. 1, the lower piece, the transient information $\mathbf{T}$, is information recorded due to an over-estimation of the entropy rate h_μ at block lengths L less than the Markov order R. This over-estimation is due, in effect, to L being shorter than the longest correlations in the data. In a complementary way, the upper portion $\mathcal{J}_0$ can be viewed as the amount of state information that cannot be retrodicted, even given the infinite future.

The relative roles of the contributions to synchronization information can be clearly seen for one-dimensional range-R spin systems. Reference [12] claimed that for spin chains:

$$\mathbf{S} = \mathbf{T} + \frac{1}{2}R(R+1)h_\mu \ , \tag{45}$$

where R is the coupling range (Markov order) of the spin chain. This can be established rather directly, and understood for the first time, using the geometric convergence picture just introduced for $\mathbf{S}$. First, Ref. [35] showed that for a spin chain $H[X_0^L, \mathcal{S}_L]$ is flat (zero slope) for $0 \leq L \leq R$, after which it converges to its asymptote. Second, combining these, we have:

$$\begin{aligned}
\mathcal{J}_0 &= \sum_{L=0}^{R} H[X_0^L, \mathcal{S}_L] - (\mathbf{E} + Lh_\mu) && (46)\\
&= \sum_{L=0}^{R} (\mathbf{E} + Rh_\mu) - (\mathbf{E} + Lh_\mu) && (47)\\
&= \sum_{L=0}^{R} (R - L)h_\mu && (48)\\
&= \frac{1}{2}R(R+1)h_\mu \ . && (49)
\end{aligned}$$

So, the amount of state information that cannot be retrodicted is quadratic in Markov order.

Finally, $H[X_0^L]$ and $H[X_0^L, \mathcal{S}_L]$ give lower and upper bounds on $\mathbf{E}$, respectively: the first monotonically approaches $\mathbf{E} + Lh_\mu$ from below and the second monotonically approaches it from above. This way, given an ϵ-machine, it is straightforward to compute $\mathbf{E}$ with any accuracy required from the block entropies, which themselves can be efficiently estimated from the ϵ-machine. Similarly, since $H[X_0^L]$ over-estimates the entropy rate while approaching from above and $H[X_0^L, \mathcal{S}_L]$ underestimates the entropy rate while approaching from below, one obtains an analogous pair of bounds on h_μ. This block-state technique for bounding the entropy rate, however, holds for any type of presentation of the process. (Cf. Ref. [21, Sec. 4.5].)

VI. PRESENTATION QUANTIFIERS

The development and results have focused, so far, on ϵ-machines and their information-theoretic properties. Due to the ϵ-machine's uniqueness, these were also properties of the corresponding processes themselves. Now, we relax the defining characteristics of ϵ-machines to consider generic presentations. Naturally, this destroys our ability to directly identify presentation properties with those of the process represented. A process's entropy rate (h_μ) and excess entropy ($\mathbf{E}$) remain unchanged, however, since they are defined solely through its observables $\Pr(\overleftarrow{X}, \overrightarrow{X})$. Widening our purview to generic presentations leads us to briefly introduce several new properties that capture information processing in presentations. Perhaps more distinctly, this also leads us to quantify the kinds of information in a presentation that are *not* characteristics of the process it represents. Section VII then provides more detailed expositions on their meaning and example processes to illustrate them.

A. Crypticity

The statistical complexity C_μ is the amount of information a process stores to generate future behavior. The crypticity χ is that part of C_μ not transmitted to the future: $\chi = C_\mu - \mathbf{E}$. Roughly, it can be thought of as the irreducible overhead that arises from the process's causal structure. Reference [22] defined crypticity for ϵ-machines as $\chi = H[\mathcal{S}_0|\overrightarrow{X}_0]$. Now, we generalize this to define crypticity for generic presentations.

Definition 5. *The* presentation crypticity $\chi(\mathcal{R})$ *is the amount of state information shared with the past that is not transmitted to the future:*

$$\chi \equiv I[\overleftarrow{X}_0; \mathcal{R}_0|\overrightarrow{X}_0] \ . \tag{50}$$

When the presentation states are causal states, this quantity reduces to the original definition—the process's crypticity. Furthermore, the crypticity is the difference between the presentation state entropy and the y-intercept of block-state entropy curve, Eq. (34).

Theorem 3. *The presentation crypticity* $\chi(\mathcal{R})$ *is the difference between the presentation state entropy* $H[\mathcal{R}_0]$*:*

$$\chi = -\mathcal{J}_1 \ . \tag{51}$$

Proof. Starting with the length-L approximation of the

crypticity, we work our way to the L^{th} partial sum of $-\mathcal{J}_1$ via a straightforward calculation:

$$\begin{aligned}
&I[X_{-L}^L; \mathcal{R}_0 | X_0^L] \\
&\quad = H[X_{-L}^L | X_0^L] - H[X_{-L}^L | \mathcal{R}_0, X_0^L] && (52)\\
&\quad = H[X_{-L}^L | X_0^L] - H[X_{-L}^L | \mathcal{R}_0] && (53)\\
&\quad = Lh_\mu - H[X_{-L}^L | \mathcal{R}_0] + H[X_{-L}^L | X_0^L] - Lh_\mu && (54)\\
&\quad = -\mathcal{J}_1(L) + H[X_{-L}^L | X_0^L] - Lh_\mu && (55)\\
&\quad = -\mathcal{J}_1(L) + H[X_0^L | X_{-L}^L] - Lh_\mu && (56)\\
&\quad = -\mathcal{J}_1(L) + \sum_{j=0}^{L-1} H[X_j | X_0^j, X_{-L}^L] - Lh_\mu && (57)\\
&\quad = -\mathcal{J}_1(L) + \sum_{j=L}^{2L-1} H[X_j | X_0^{L+j}] - Lh_\mu \ . && (58)
\end{aligned}$$

Equation (53) follows because the states (in any hidden Markov model) shield the past from the future: the future is a function of the state. Equation (55) follows from the definition of $\mathcal{J}_1$, and Eq. (56) from stationarity. Equation (57) follows from the chain rule for block entropies [21], and Eq. (58) from using stationarity again.

Finally, we take the large-L limit. By definition, we have $\mathcal{J}_1(L) \to \mathcal{J}_1$. The remaining difference converges to zero due to a result in Ref. [37] that the conditional block entropies converge to the entropy rate faster than linearly in L. □

B. Oracular Information

We now introduce a sibling of crypticity—the *oracular information*.

Definition 6. *The* oracular information *is the amount of state information shared with the future that is not derived from the past:*

$$\zeta \equiv I[\mathcal{R}_0; \overrightarrow{X}_0 | \overleftarrow{X}_0] \ . \tag{59}$$

This new quantity is always zero for the ϵ-machine and nonzero only for nonunifilar presentations. We have the following characterization.

Theorem 4. *The oracular information is the difference between the presentation state entropy $H[\mathcal{R}_0]$ and the y-intercept of the state-block entropy curve, Eq.* (33)*:*

$$\zeta = -\mathcal{K}_1 \ . \tag{60}$$

Proof. The proof proceeds almost identically to the corresponding result for crypticity. Namely,

$$\begin{aligned}
I[\mathcal{R}_0; X_0^L | X_{-L}^L] = &- \mathcal{K}_1(L) \\
&+ \sum_{j=L}^{2L-1} H[X_j | X_0^j] - Lh_\mu \ . && (61)
\end{aligned}$$

Then, taking the large-L limit proves the result. □

C. Gauge Information

When moving away from the optimal representation afforded by a process's ϵ-machine, it is possible to encounter presentations containing state information that is not justified by a process's bi-infinite set of observables. We call this *gauge information* to draw a parallel with the descriptional degrees of freedom that gauge theory addresses in physical systems [39].

Definition 7. *The* gauge information *is the uncertainty in the presentation states given the entire past and future:*

$$\varphi \equiv H[\mathcal{R}_0 | \overleftarrow{X}_0, \overrightarrow{X}_0] \ . \tag{62}$$

That is, to the extent there is uncertainty in the states, even after the past and the future are known, the presentation contains state uncertainty above and beyond the process. Thus, there are components of the model that are not determined by the process; rather they are the result of a choice of presentation.

Intuitively, gauge information can be related to the total state entropy, crypticity, oracular information, and excess entropy. Later, we will discuss information diagrams as a useful visualization tool, but for now, we simply point out that one can "visually" verify the following theorem from Figure 13.

Theorem 5. *Gauge information is the difference between the state entropy and the sum of the crypticity, oracular information, and excess entropy:*

$$\varphi = H[\mathcal{R}] - (\chi + \zeta + \mathbf{E}) \ . \tag{63}$$

Proof. Since we are working with hidden Markov models, the future and past are conditionally independent given the current state. Thus, $\mathbf{E} \equiv I[\overleftarrow{X}; \overrightarrow{X}] = I[\overleftarrow{X}; \mathcal{R}; \overrightarrow{X}]$. Now, the proof proceeds as a simple verification:

$$\begin{aligned}
\chi(L) + \zeta(L) + \mathbf{E}(L) &= I[\overleftarrow{X}^L; \mathcal{R} | \overrightarrow{X}^L] \\
&\quad + I[\mathcal{R}; \overrightarrow{X}^L | \overleftarrow{X}^L] \\
&\quad + I[\overleftarrow{X}^L; \mathcal{R}; \overrightarrow{X}^L] \\
&= H[\mathcal{R}] - H[\mathcal{R} | \overleftarrow{X}^L, \overrightarrow{X}^L] \ .
\end{aligned}$$

So, finite-length approximations to the gauge information

can be written as:

$$H[\mathcal{R}] - \big(\chi(L) + \zeta(L) + \mathbf{E}(L)\big) = H[\mathcal{R}|\overleftarrow{X}^L, \overrightarrow{X}^L] \ .$$

Taking the limit, we achieve our desired result. □

D. Synchronization Information

As we noted, it is always possible to asymptotically synchronize to an ϵ-machine with a finite number of recurrent causal states. For some processes, synchronization can happen in finite time. While in others, it can only happen in the limit as the observation window tends to infinity. In either case, it is always true that $H[\mathcal{S}_\infty|\overrightarrow{X}] = 0$.

When we generalize to presentations that differ from ϵ-machines, it is no longer true that one always synchronizes to the presentation states. In such cases, there is irreducible state uncertainty, even after observing an infinite number of symbols. This kind of state uncertainty cannot be reduced by past observations alone. Due to this, the synchronization information, as previously defined, diverges.

Definition 8. *The* presentation synchronization information *is the total uncertainty in the presentation states:*

$$\mathbf{S} \equiv \sum_{L=0}^{\infty} H[\mathcal{R}_L|X_0^L] \ . \tag{64}$$

We will show in Sec. VI H that this can be understood in terms of the gauge and oracular informations.

E. Cryptic Order

The cryptic order was defined in Ref. [24] as the minimum length k for which $H[\mathcal{S}_k|\overrightarrow{X}_0] = 0$. Reference [38] shows that the cryptic order is a topological property of the *irreducible sofic shift* [32] describing the support of the ϵ-machine. However, we can understand the cryptic order geometrically as the length k_χ at which the block-state entropy $H[X_0^L, \mathcal{S}_L]$ reaches its asymptote; see Eq. (33). It turns out that this concept generalizes directly to generic presentations.

Definition 9. *The* presentation cryptic order *is the length k at which the block-state entropy curve reaches its asymptote:*

$$k_\chi \equiv \min\big\{L : H[X_0^L, \mathcal{R}_L] = H[\mathcal{R}_0] - \chi + h_\mu L\big\} \ . \tag{65}$$

One would like to understand the cryptic order in terms of an explicit limit, as done for ϵ-machines, where cryptic order is the minimum k for which $H[\mathcal{S}_k|\overrightarrow{X}_0] = 0$. The obvious complication for presentations, in general, is that one might never synchronize to a particular state. However, it turns out that one can understand the presentation cryptic order in terms of one's uncertainty in the distribution over *distributions of states*—that is, the uncertainty in the distribution over *mixed states* [23, 40]. Specifically, we frame the generalized cryptic order in terms of synchronizing to distributions over presentation states. We outline the approach briefly; a detailed exposition will appear elsewhere [38].

As measurements are made, an observer's uncertainty in the state of the presentation varies. However, the pattern of variation becomes regular as more observations are made. The cryptic order, then, is understood as the number of distributions over presentation states that one cannot know with certainty from time $t = 0$ given the entire future. Said differently, the cryptic order is the time at which an observer becomes absolutely certain about the uncertainty in the presentation states.

F. Oracular Order

The oracular order definition parallels those of the cryptic and the Markov orders.

Definition 10. *The* oracular order *is the length k_ζ at which the state-block entropy curve reaches its asymptote:*

$$k_\zeta \equiv \min\big\{L : H[\mathcal{R}_0, X_0^L] = H[\mathcal{R}_0] - \zeta + h_\mu L\big\} \ . \tag{66}$$

It always vanishes for ϵ-machines. So, this new length scale is a property of the presentation only and not of the process generated by the presentation.

G. Gauge Order

The gauge order definition also parallels those of the cryptic, Markov, and oracular orders.

Definition 11. *The* gauge order *is the length k_φ at which $H[\mathcal{R}_0|X_{-L}^L X_0^L]$ reaches its asymptote.*

$$k_\varphi \equiv \min\{L : H[\mathcal{R}_0|X_{-L}^L, X_0^L] = \varphi\} \ . \tag{67}$$

Geometrically, we visualize the gauge order as the length at which the difference between two curves—$H[X_{-L}^L, \mathcal{R}_0, X_0^L]$ and $H[X_{-L}^L, X_0^L]$—becomes fixed to their asymptotic difference.

Theorem 6. *The* gauge order *is the maximum of the Markov, cryptic, and oracular orders:*

$$k_\varphi = \max\{R, k_\chi, k_\zeta\} \ . \tag{68}$$

Proof. The gauge information can be understood as the left-over state information after the excess entropy, cryp-

ticity, and oracular information [41] have been extracted:

$$\varphi = H[\mathcal{R}_0] - \mathbf{E} - \chi - \zeta \ . \tag{69}$$

Thus, as soon as the observer reaches each of the Markov, cryptic, and oracular orders, the remaining state information exactly equals the gauge information. □

The Markov, cryptic, and oracular orders are similar in that they refer to minimum word lengths one must examine to remove corresponding uncertainties. Unlike them, the gauge order does *not* indicate a scale at which an amount of information is contained. Rather, it is more the opposite. The gauge order, as it is not removable given the entire past or future, is defined in the negative. It is the minimum word length one must consider to remove *all* of the above uncertainties (except gauge), thus leaving only the gauge information as unknown. In short, it is the length scale beyond which there is no point attempting to extract more state information (even with an oracle). This is so precisely because the remainder is the gauge information and, therefore, not correlated with the process language. It corresponds to what one calls a gauge freedom in physics.

H. Synchronization Order

As mentioned in Sec. V B, the length at which an observer has synchronized to an ϵ-machine is always R, the Markov order. Recall, any order-R Markov process has $I[\overrightarrow{X}_R; \overleftarrow{X}_0 | X_0^R] = 0$. Synchronization to the ϵ-machine requires that $H[\mathcal{S}_L | X_0^L] = 0$, and it is straightforward to see that this holds for $L = R$. As we generalize to non-ϵ-machine presentations, though, we must look beyond Markov order to address the fact that one might only synchronize to distributions over presentation states.

Definition 12. *The* presentation synchronization order *is the length* $k_{\mathbf{S}}$ *at which* $H[\mathcal{R}_L | X_0^L]$ *reaches its asymptote:*

$$k_{\mathbf{S}} \equiv \min\{L : H[\mathcal{R}_L | X_0^L] = \varphi + \zeta\} \ . \tag{70}$$

The motivation for this definition is that the asymptote is simply the difference of the asymptotes for the block-state and block entropy curves. That is, the synchronization order is also thought of as the length at which the state uncertainty equals its irreducible state uncertainty: $\varphi + \zeta = H[\mathcal{R}_0 | \overleftarrow{X}_0]$.

Now, we show that the synchronization order must occur at either the presentation cryptic order or the Markov order.

Theorem 7. *The presentation synchronization order is the maximum of the Markov and presentation cryptic orders:*

$$k_{\mathbf{S}} = \max\{R, k_\chi\} \ . \tag{71}$$

Proof. When both the block-state and block entropy curves have reached their asymptotes the observer will have extracted $\mathbf{E} + \chi$ bits of state information. This leaves $H[\mathcal{R}_0] - \mathbf{E} - \chi = \varphi + \zeta$ bits. This is exactly the irreducible state uncertainty—that which cannot be learned from the observables. □

Note that for ϵ-machines: $\mathbf{E} + \chi = C_\mu$. So, when an observer has extracted all that can be learned about the process from the past observables, the observer has learned everything about the causal states.

When the synchronization order is finite, $H[\mathcal{R}_L | X_0^L]$ is fixed at the presentation's irreducible state uncertainty for all $L > k_{\mathbf{S}}$. Then, it can be helpful to view the presentation synchronization information as consisting of two contributions:

$$\mathbf{S} = \sum_{L=0}^{k_{\mathbf{S}}-1} H[\mathcal{R}_L | X_0^L] + \sum_{L=k_{\mathbf{S}}}^{\infty} (\varphi + \zeta) \ . \tag{72}$$

When the synchronization order is not finite, it can be useful to interpret the synchronization information in a slightly different manner:

$$\mathbf{S} = \mathcal{I}_0 + \mathcal{J}_0 + \lim_{L \to \infty} (\varphi + \zeta) L \ . \tag{73}$$

I. Synchronization Time

Reference [13] defined the *synchronization time* τ of a periodic process to be the average time needed to synchronize to the states. Let $w = w_0 \cdots w_{p-1}$ be a cyclic permutation of the word that is repeated by a periodic process having period p. It follows that

$$\Pr(X_0^p = w) = \frac{1}{p} \ , \tag{74}$$

since any cyclic permutation is just as likely as another. Now, while each permutation has the same probability, it is not true that each permutation is equally informative in terms of synchronization. For example, consider the process that repeats the word 00011, indefinitely. If an observer saw 01, then the observer would be synchronized. In contrast, the observer would not be synchronized if 00 had been observed instead. Reference [13] defined τ_w as the synchronization time of the cyclic permutations of w. Then,

$$\tau = \sum_w \tau_w \Pr(X_0^p = w) \ . \tag{75}$$

Since $h_\mu = 0$ for all periodic processes,

$$\Pr(X_0^p = w) = \Pr(X_0^{\tau_w} = w_0 \cdots w_{\tau_w - 1}) \ . \tag{76}$$

Thus, we can rewrite τ suggestively as:

$$\tau = \sum_w \tau_w \Pr(X_0^{\tau_w} = w_0 \cdots w_{\tau_w - 1}) \ . \tag{77}$$

Then, instead of summing over all cyclic permutations of w, we can just sum over the set $\mathcal{L}_{\text{sync}}$ of all minimal synchronizing words. (A word is a *minimal synchronizing word* if no prefix of the word is also synchronizing.) Now, we can extend τ to all finitary processes, not just periodic ones.

Definition 13. *The* process synchronization time *is the average time required to synchronize to the ϵ-machine's recurrent causal states:*

$$\tau \equiv \sum_{w \in \mathcal{L}_{\text{sync}}} |w| \Pr\left(X_0^{|w|} = w\right) \ . \tag{78}$$

Note that any order-R Markov process has $\tau \leq R$. The synchronization time gives an intuition for how long it takes to synchronize to a stochastic process.

As an example, recall the Even Process [12]. It has the property that there are arbitrarily long minimal synchronizing words. For example, 1^k0 is always a minimal synchronizing word, for any k. Despite this fact, the synchronization time of the Even Process is $\tau = 10/3$. After repeatedly observing sequences four symbols in length, on average an observer will be synchronized to the states of the ϵ-machine.

When considering more general presentations it is not always the case that one can synchronize to the states, as τ can be infinite. Just as with the cryptic order, however, one can synchronize to distributions over the presentation states. This motivates the presentation synchronization time.

Definition 14. *The* presentation synchronization time *is the average time required to synchronize to a recurrent distribution over presentation states.*

We provide an intuitive definition here, leaving a more detailed discussion, where notation is properly developed, for a sequel.

VII. CLASSIFYING PRESENTATIONS

The ϵ-machine is frequently the preferred presentation of a process, especially when one is interested in understanding fundamental properties of the process itself. However, one might be interested in the properties of particular presentations of a process, and it would be helpful if there was an analogous theory similar to that for ϵ-machines.

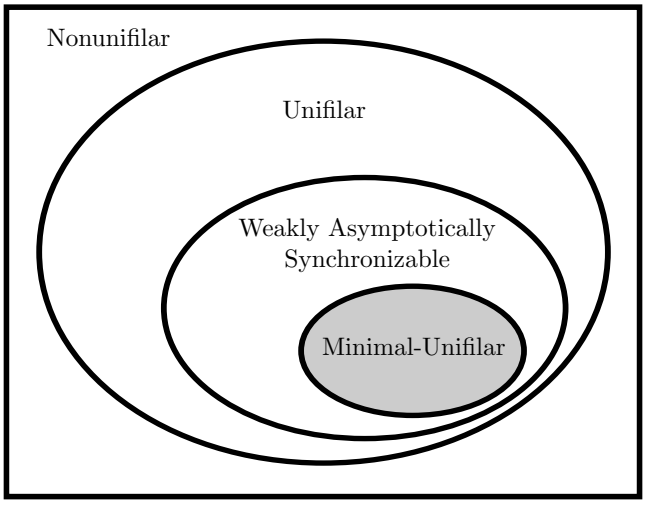

FIG. 2. The hierarchy of presentations for a finitary process. The gray region represents that portion to which the ϵ-machine belongs.

To develop this, we establish a classification of a process's presentations. The classes are defined in terms of whether a presentation is nonunifilar, unifilar, weakly asymptotically synchronizable, and minimal unifilar. The result is shown in Fig. 2, which shows that the presentation classes form a nested hierarchy.

The most general type of presentation is nonunifilar, where we allow the possibility that $H[\mathcal{R}_1|\mathcal{R}_0, X_0] > 0$. Then, unifilar presentations are the subset of nonunifilar presentations for which this quantity is exactly zero. In the unifilar class, there can be redundant states—states from which the future looks exactly the same and also states which have the exact same histories mapping to them. When we move to weakly asymptotically synchronizable presentations, all redundant states are removed and the remaining states must induce a partition on the set of histories that is a refinement of the causal state partition; cf. Ref. [29, Lemma 7]. Finally, minimal unifilar presentations are the ϵ-machines, whose partition of the pasts is the coarsest one possible.

In this light, one might conclude that ϵ-machines are an overly restricted set of presentations. They are indeed a restricted set, but it is a restriction with purpose: The ϵ-machine is the unique minimal prescient presentation within the set of a process's presentations. Moreover, all of a process's properties can be determined from its ϵ-machine. These facts allow one to purposefully conflate properties of the ϵ-machine with process's properties.

We will use a *information diagram* (I-diagram) [3] to analyze what happens as one relaxes the defining properties of the ϵ-machine presentation's random variables.

With the ϵ-machine, we have the past $\overleftarrow{X}$, the causal states $\mathcal{S}$, and the future $\overrightarrow{X}$. As we move away from the ϵ-machine's causal states, we must consider in addition the rival states $\mathcal{R}$.

In total, there are four random variables to consider. The full range of their possible information-theoretic relationships appears in the information diagram (I-diagram) of Fig. 3. However, Appendix C shows that 7 of the 15 atoms (elemental components of the multivariate information-measure sigma algebra) vanish. This allows us to simplify other atoms dramatically. For example, the atom:

$$
\begin{aligned}
I[\overleftarrow{X}; \mathcal{S}; \mathcal{R}; \overrightarrow{X}] & \quad (79)\\
= I[\overleftarrow{X}; \mathcal{S}; \overrightarrow{X}] - I[\overleftarrow{X}; \mathcal{S}; \overrightarrow{X} | \mathcal{R}] & \quad (80)\\
= I[\overleftarrow{X}; \mathcal{S}; \overrightarrow{X}] & \quad (81)\\
= I[\overleftarrow{X}; \overrightarrow{X}] - I[\overleftarrow{X}; \overrightarrow{X} | \mathcal{S}] & \quad (82)\\
= I[\overleftarrow{X}; \overrightarrow{X}] - (I[\overleftarrow{X}; \mathcal{R}; \overrightarrow{X} | \mathcal{S}] + I[\overleftarrow{X}; \overrightarrow{X} | \mathcal{S}, \mathcal{R}]) & \quad (83)\\
= I[\overleftarrow{X}; \overrightarrow{X}] \ , & \quad (84)
\end{aligned}
$$

where we made use of the atoms that vanish. Thus, the four-way mutual information simply reduces to the mutual information between the past and the future—the excess entropy:

$$
I[\overleftarrow{X}; \mathcal{S}; \mathcal{R}; \overrightarrow{X}] = \mathbf{E} \ . \quad (85)
$$

Similar calculations reduce the other information measures in Fig. 3 correspondingly. We now consider these reductions in turn.

A. Case: Minimal Unifilar Presentation

The set of minimal unifilar presentations corresponds exactly to the ϵ-machines, up to state relabeling. The states in these presentations, the causal states, induce a partition of the infinite pasts via the function $\epsilon(\overleftarrow{x})$.

The information diagram and entropy growth plot are particularly simple, as seen in Fig. 4 and Fig. 5. This simplicity derives from the efficient predictive role the causal states play. Referring to the I-diagram, $H[\mathcal{S}|\overleftarrow{X}] = 0$ because of determinism of the $\epsilon(\overleftarrow{x})$ map, Eq. (18). Next, causal states, as well as all other states we consider, are prescient states and so $I[\overleftarrow{X}; \overrightarrow{X} | \mathcal{R}] = 0$. These straightforward requirements entirely determine the form of the ϵ-machine I-diagram in Fig. 4. As we step through the space of presentation classes, we will see these relationships become more complex.

There are three quantities that require attention in this figure. First, the state entropy $H[\mathcal{R}]$ is equal to C_μ—the statistical complexity. This particular state information is considered privileged as it is the state information associated with the ϵ-machine and so the process. The excess entropy $\mathbf{E}$ is the mutual information between the past and future and is also exactly that information which the (causal) states contain about the future. Lastly, the crypticity χ is the amount of information "overhead" required for prediction using the ϵ-machine. Generally, this overhead is associated with the presentation as well as the process *itself*, due to the uniqueness of the ϵ-machine presentation. It is the irreducible memory associated with the process. At any time, the process itself or a predictive model must keep track of C_μ bits of state information, while only $\mathbf{E}$ bits of this information are correlated with the future.

The entropy growth plot, Fig. 5, is also simplified by using causal states. In terms of our newly defined integrals: $\mathcal{K}_n = 0$ for all n and $\mathcal{J}_1 = H[\mathcal{S}] - \mathcal{I}_1 = \chi$.

A simple example that illustrates all of these points is provided by the Golden Mean Process and its ϵ-machine; see Fig. 6. When the probability p is chosen to be $\frac{1}{2}$, the values of our information measures are $C_\mu = \log_2(3) - \frac{2}{3} = 0.9183$ bits, $\chi = \frac{2}{3}$ bits, and $\mathbf{E} = C_\mu - \chi = 0.2516$ bits. As we explore alternate presentations, we will return to this process as a common thread for explanation and intuition.

B. Case: Weakly Asymptotically Synchronizable Presentations

Let's relax the minimality constraint leaving the ϵ-machines for presentations that are nonminimal unifilar and weakly asymptotically synchronizable. Again, the states correspond to a partition of the infinite pasts, but since they are prescient and not minimal unifilar, the partition must be a refinement of the causal-state partition [29].

The effect of this is benign as seen in both the I-diagram (Fig. 7) and the entropy growth plot (Fig. 8). In Fig. 7, weakly asymptotically synchronizability ensures that $H[\mathcal{R}|\overleftarrow{X}] = 0$. Demanding prescient states determines the form of the I-diagram. Figure 7 indicates that $H[\mathcal{R}] > H[\mathcal{S}]$. This is a consequence of $\boldsymbol{\mathcal{R}}$ being a nontrivial refinement of $\boldsymbol{\mathcal{S}}$.

Examining the entropy growth plot, the increased state information is reflected in the values of the block-state and state-block entropy curves at $L = 0$. Additionally, it is interesting to note what happens to the cryptic order. We generalized the definition of cryptic order to be that length where the block-state entropy reaches its asymptote. Since block-state entropy is nondecreasing, this suggests that it might be forced to reach its asymptote at a larger value of L than the cryptic order for the

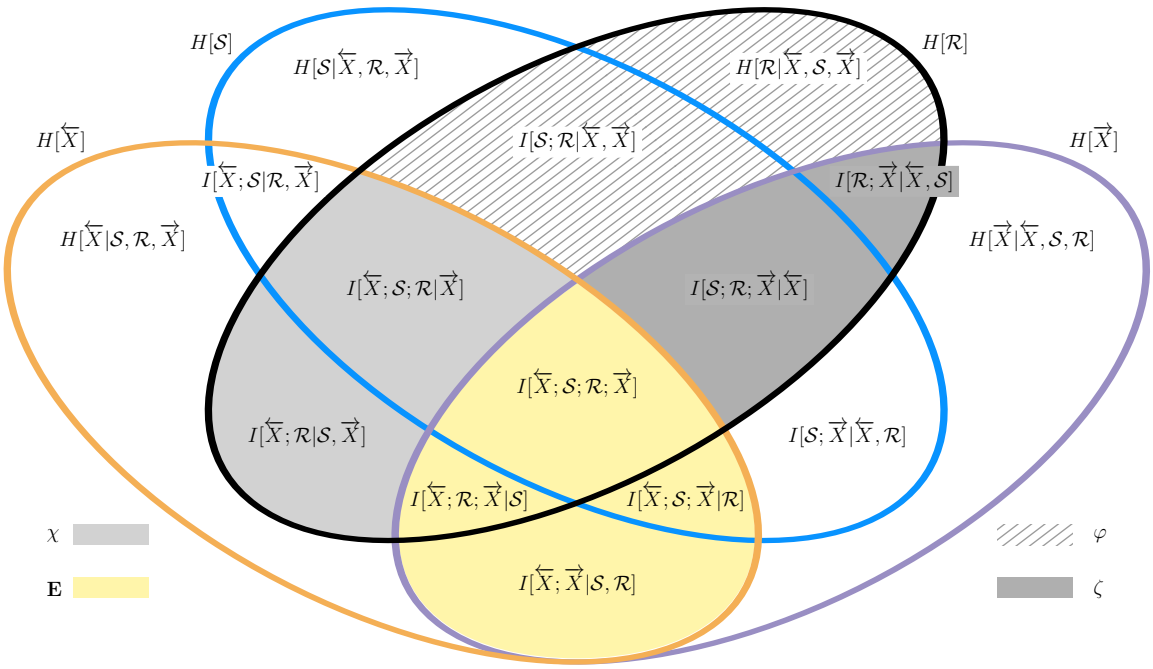

FIG. 3. The general four-variable information diagram involving $\overleftarrow{X}$, $\mathcal{S}$, $\mathcal{R}$, and $\overrightarrow{X}$. The shaded light gray is the generalized crypticity χ. The yellow is the excess entropy $\mathbf{E}$. The dark gray is the oracular information ζ. The hatched area is the gauge information φ. Note that this is only a schematic diagram of the interrelationships. In particular, potentially infinite quantities—such as, $H[\overleftarrow{X}]$ and $H[\overrightarrow{X}]$—are depicted with finite areas.

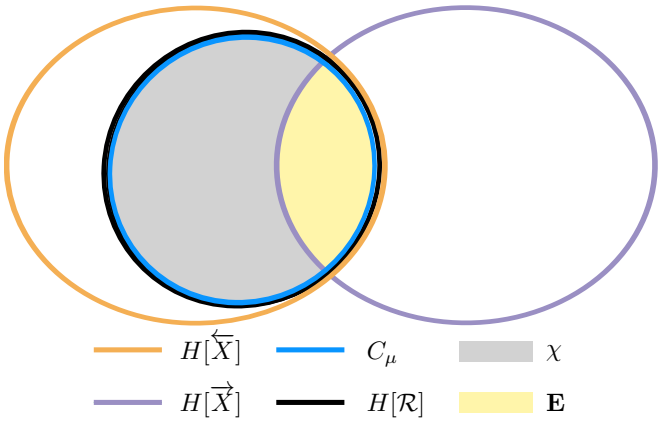

FIG. 4. The information diagram for an ϵ-machine. The states of the presentation are causal states and induce a partition on the past. The entropy over the states, $H[\mathcal{R}_0] = H[\mathcal{S}_0]$, defines the statistical complexity C_μ. The process crypticity is the difference of the statistical complexity and the excess entropy $\mathbf{E}$.

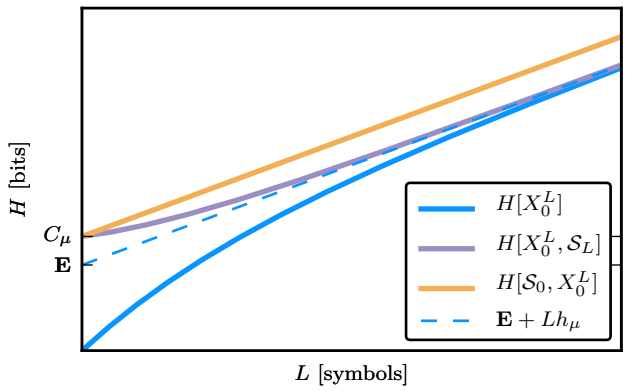

FIG. 5. Entropy growth for a generic ϵ-machine. $H[X_0^L]$ and $H[X_0^L, \mathcal{S}_L]$ both converge to the same asymptote (dashed line). $H[\mathcal{S}_0, X_0^L]$ is linear.

ϵ-machine presentation. We can see that this is in fact true by expanding $H[X_0^L, \mathcal{S}_L, \mathcal{R}_L]$ in two ways. Note that this joint entropy term combines variables from two *different* presentations. In the first expansion,

$$\begin{aligned} H[X_0^L, \mathcal{S}_L, \mathcal{R}_L] &= H[\mathcal{S}_L | X_0^L, \mathcal{R}_L] + H[X_0^L, \mathcal{R}_L] \\ &= H[X_0^L, \mathcal{R}_L] \end{aligned}$$

The conditional entropy is zero since the rival states $\mathcal{R}$ are a refinement of the causal states $\mathcal{S}$. In the second

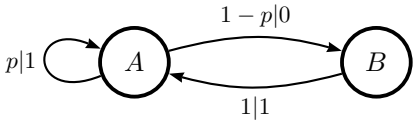

FIG. 6. The ϵ-machine presentation of the Golden Mean Process.

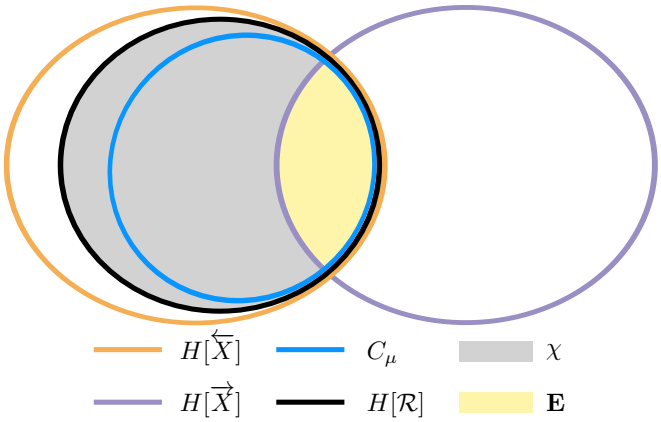

FIG. 7. The information diagram for a presentation that is weakly asymptotically synchronizable, but not necessarily minimal unifilar. The states still induce a partition on the infinite past. The presentation crypticity $\chi(\mathcal{R})$ is the difference of the state entropy $H[\mathcal{R}] \geq C_\mu$ and the excess entropy $\mathbf{E}$.

expansion,

$$H[X_0^L, \mathcal{S}_L, \mathcal{R}_L] = H[\mathcal{R}_L | X_0^L, \mathcal{S}_L] + H[X_0^L, \mathcal{S}_L]$$

Then, we combine these expansions to obtain:

$$\begin{aligned} H[X_0^L, \mathcal{R}_L] &= H[\mathcal{R}_L | X_0^L, \mathcal{S}_L] + H[X_0^L, \mathcal{S}_L] \\ &\geq H[X_0^L, \mathcal{S}_L] \ . \end{aligned} \tag{86}$$

This shows that the block-state curve for the nonminimal presentation lies above or on the curve for the ϵ-machine presentation. Since block and block-state entropies share an asymptote—$\mathbf{E} + Lh_\mu$—the nonminimal unifilar block-state entropy will reach its asymptote at a value greater than or equal to the process's cryptic order. More care will be required in the subsequent cases, as the relations among entropy growth functions are more complicated.

To illustrate these class characteristics, consider the following three-state presentation of the Golden Mean Process in Fig. 9. The original causal-state partition, $\{A = *1, B = *0\}$, has become refined. (Here, $*$ denotes any allowed history.) We now have $\{A = *11, B = *0, C = *01\}$. It is straightforward to verify that $H[\mathcal{R}] = \log_2(3) = 1.585$ bits. Excess entropy is unchanged as it is a feature of the process language and not the presentation. As illustrated in Fig. 7, the crypticity grows commensurately with $H[\mathcal{R}]$.

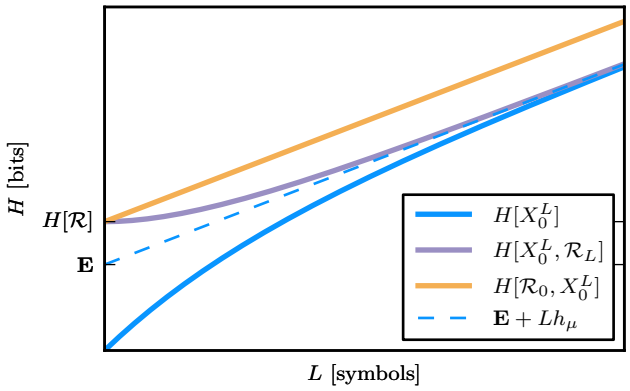

FIG. 8. Entropy growth for a weakly asymptotically synchronizing presentation. $H[X_0^L]$ and $H[X_0^L, \mathcal{S}_L]$ both converge to the same asymptote (dashed line). $H[\mathcal{S}_0, X_0^L]$ is linear. $H[\mathcal{R}]$ is larger than C_μ.

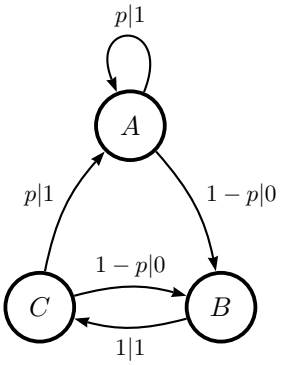

FIG. 9. A weakly asymptotically synchronizable and nonminimal unifilar presentation of the Golden Mean Process: observing a 0 synchronizes the observer to state B.

We have shown that for weakly asymptotically synchronizable presentations the presentation cryptic order generally will be larger than the cryptic order. It is interesting to note that it is also possible for the presentation cryptic order to surpass even the Markov order. Our three-state example (Fig. 9) is 2-cryptic while the Markov order remains $R = 1$ as it also depends only on the process language.

Since the Markov order R bounds the domain of the $\mathcal{I}_0$ integral and the presentation cryptic order k bounds the domain of the $\mathcal{J}_0$ integral, the domain of the synchronization information is bounded by $\max\{R, k\}$.

C. Case: Unifilar Presentations

Removing the requirement that a presentation be weakly asymptotically synchronizable, we no longer operate with (recurrent) states that correspond to a partition of the infinite past, but rather to a covering of the set

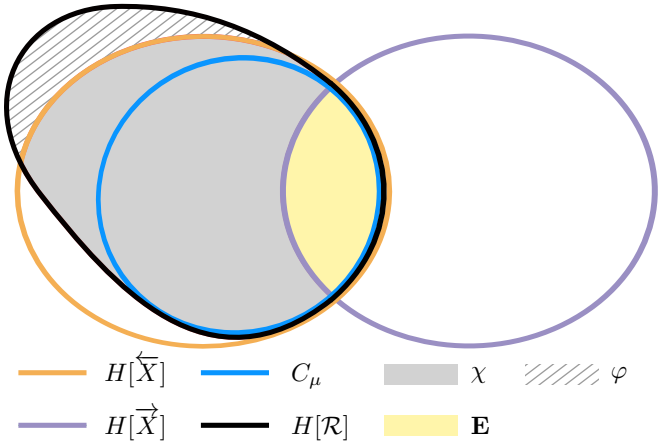

FIG. 10. The information diagram for a presentation that is not weakly asymptotically synchronizable, but still unifilar. The states are prescient, but no longer induce a partition on the infinite past. Furthermore, the states contain information that the past does not contain. The presentation crypticity is the difference of the state entropy $H[\mathcal{R}_0] \geq C_\mu$ and the excess entropy $\mathbf{E} = I[\overleftarrow{X}_0; \overrightarrow{X}_0]$.

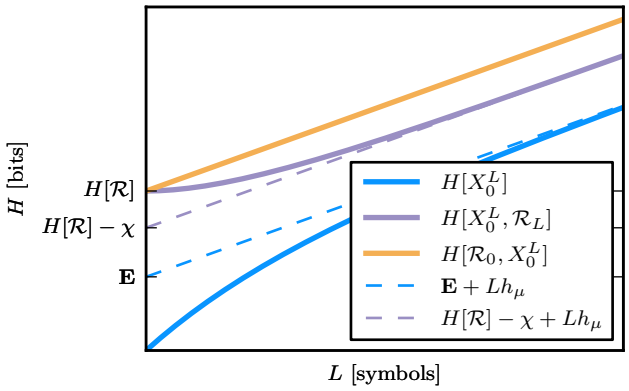

FIG. 11. Entropy growth for a not weakly asymptotically synchronizable, but unifilar presentation. $H[X_0^L]$ and $H[X_0^L, \mathcal{S}_L]$ both converge to different asymptotes (lower and upper dashed lines, respectively). $H[\mathcal{S}_0, X_0^L]$ is linear. $H[\mathcal{R}]$ is larger than C_μ.

of infinite pasts. That is, $\eta(\overleftarrow{x})$ can be multivalued, although for each $\rho \in \boldsymbol{\mathcal{R}}$, $\eta^{-1}(\rho)$ is a set of pasts that is a subset of some causal state's set of pasts.

Every allowable infinite history induces at least one state in the presentation—this is the definition of an allowable infinite history. Additionally, any presentation that is not weakly asymptotically synchronizable must have a (positive measure) set of histories where each history induces more than one state.

Consider a unifilar presentation and an infinite history that induces only one state. Due to unifilarity, we can use this history to construct an infinite set of histories that are also synchronizing. We conjecture that this set of histories must have zero measure and, even stronger, that for finite-state unifilar presentations with a single recurrent component, there are no synchronizing histories.

This inability to synchronize, a product of the nontrivial covering, is represented as the information measure φ in Fig. 10. This information is not captured by the causal states. In fact, it is not even captured by the past (or the future). It also is not necessary for making predictions with the same power as the ϵ-machine. Like $\chi(\mathcal{R})$, φ is unnecessary for prediction. However, unlike $\chi(\mathcal{R})$, φ does not capture any structural property of the process. Instead, it represents degrees of freedom entirely decoupled (informationally) from the process language and prediction. For this reason, we called it the *gauge information*.

The entropy growth plot of Fig. 11 has a new and significant feature representing the change in class. The asymptotes of the block entropy and block-state entropy become nondegenerate. This has the effect of making the synchronization information diverge. Although this fact follows immediately from the definition of weakly asymptotically synchronizable, it is instructive to see its geometric representation.

Since, from this point forward, synchronization information is always infinite, we find it necessary to re-express what synchronization information means. It can be denoted, recall Eq. (73), as the sum of a finite piece and the limit of a linear (in L) piece: $\mathbf{S} = \mathcal{I}_0 + \mathcal{J}_0 + \lim_{L\to\infty} L\varphi$. This rate of increase of the linear piece is exactly the gauge information.

It is also interesting to note that when this information is obtained—that is, a constraint is imposed upon the descriptional degrees of freedom—unifilarity maintains synchronization as more data is produced. In this sense, acquiring gauge information is a "one-time" cost.

The Golden Mean Process presentation in Fig. 12 illustrates all of the features described above. It is straightforward to see that this presentation is not weakly asymptotically synchronizing. Any history, finite or infinite, has exactly two states that it induces. This degeneracy is never broken, due to unifilarity. Rephrasing, the gauge information value, $\varphi = 1$ bit, derives from the fact that each infinite history induces one of two states with equal likelihood. This relies on the fact that there is no oracular information contribution—$\zeta = 0$ bits since the presentation is unifilar—to disentangle from the gauge information.

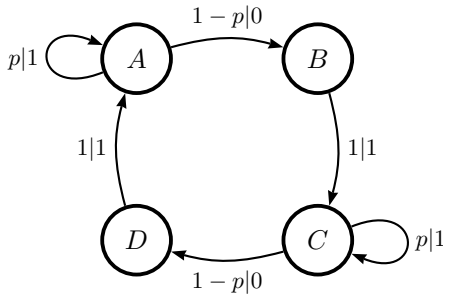

FIG. 12. A unifilar, but not weakly asymptotically synchronizing, presentation of the Golden Mean Process.

D. Case: Nonunifilar Presentations

Finally, we remove the requirement of unifilarity and examine the much larger, complementary space of nonunifilar presentations. Only one nonunifilar state must be present to change the class of the whole presentation. This ease of breaking unifilarity is why nonunifilar presentations form a much larger class.

Examining the I-diagram in Fig. 13, we notice one new feature: the oracular information $\zeta = I[\mathcal{R}; \overrightarrow{X} | \overleftarrow{X}] = I[\mathcal{R}; \overrightarrow{X} | \mathcal{S}] \neq 0$. The oracular information is a curious quantity and so deserves careful interpretation. It is the degree to which the presentation state reduces uncertainty in the future beyond that for which the past can account. One might think of this feature as "superprescience". Not only is the information from the past being maximally utilized for prediction, but some additional information is also injected. We make several remarks about this.

It is well known that a process's nonunifilar presentations may be smaller than the corresponding ϵ-machine. This fact is sometimes cited [27] as providing evidence that the smaller nonunifilar presentation is the more "natural" one [42]. While it is true that the state information $H[\mathcal{R}]$ can be smaller than C_μ, and in fact often is, the I-diagram makes plain the fact that oracular information must be introduced to determine $\mathcal{R}$ and, thus, make a super-prescient prediction. For this reason, unless one is transparent about allowing for oracular information, it is not appropriate to make a judgment about naturalness of nonunifilar presentations.

Given that we do not have the luxury of access to an oracle, we might like to know how these presentations perform without this information. The nonoracular part of $I[\mathcal{R}; \overrightarrow{X}]$ is simply **E**. That is, without the oracular information, we predict just as we would with any other prescient presentation. However, the predictions are made using *distributions over states* rather than individual states. (The former are the mixed states of Ref. [23].) More importantly, as we continue to make predictions, the state distribution evolves through a series of distributions. These distributions are in 1-to-1 correspondence with the causal states of the ϵ-machine. And so, for a nonoracular user of a nonunifilar presentation to communicate her history-induced state to another requires the transmission of C_μ bits. The statistical complexity is inescapable as the proper information storage of the process.

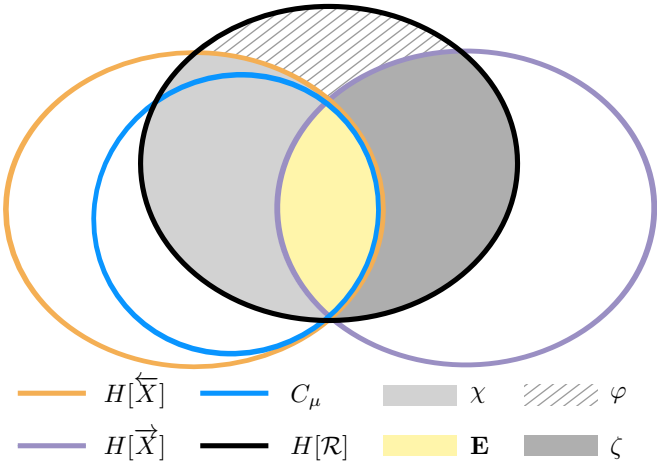

FIG. 13. The information diagram for a presentation that is not unifilar. The states are super-prescient, do not induce a partition on the past, and have information not contained in the past. The presentation crypticity is the difference of the state entropy $H[\mathcal{R}]$ and the excess entropy **E**. Note, the state entropy can also be smaller than the statistical complexity C_μ.

When discussing unifilar presentations for which $H[\mathcal{R}_L | X_0^L] \neq 0$ at any finite L or even in the limit, we indicated that the gauge information was a "one-time" cost. Now, we ask the same question of the two informations—gauge and oracular—that are not products of the past. Since we no longer have unifilarity, state uncertainty is dynamically reintroduced as synchronization is lost. That is, nonunifilar presentations are allowed to *locally* resynchronize following the introduction of state uncertainty. The net result is that over time synchronization is repeatedly lost and reacquired.

The entropy growth plot of Fig. 14 makes one last adjustment to acknowledge the change in class. For the first time, the state-block entropy is nonlinear. It approaches its asymptote from above and, moreover, the asymptote is independent of the block-state asymptote. The projection back onto the y-axis mirrors our final and most general I-diagram of Fig. 13.

A nonunifilar presentation of the Golden Mean Process is shown in Fig. 15. All of the above-mentioned quantities are nonzero for this presentation: For $p = 1/2$, the crypticity $\chi(\mathcal{R}) = 1/3$ bits, the gauge information $\varphi = 1$ bit, and the oracular information $\zeta = 1/3$ bits. The value

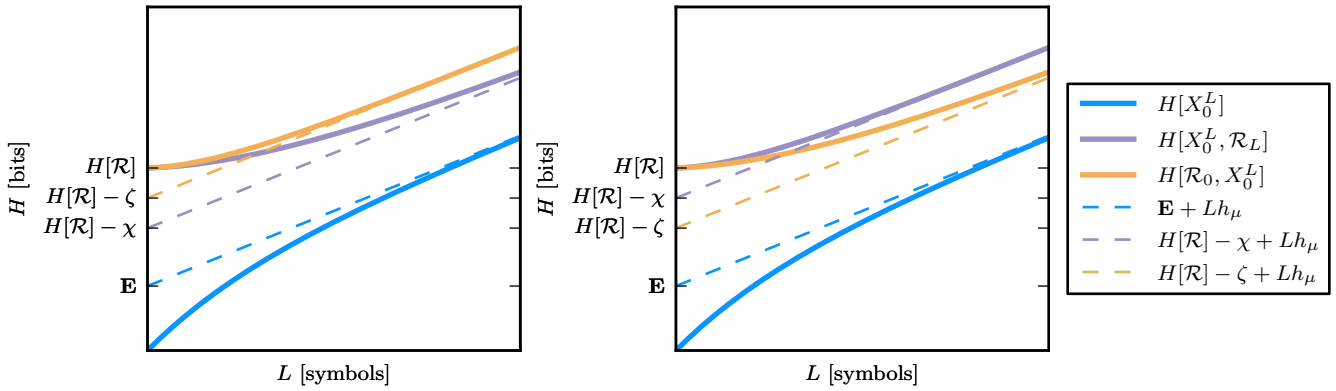

FIG. 14. Entropy growth for a nonunifilar presentation. Left: $H[X_0^L]$ and $H[X_0^L, \mathcal{S}_L]$ both converge to different asymptotes; $H[\mathcal{S}_0, X_0^L]$ is not linear and $H[\mathcal{R}]$ is larger than C_μ. Right: The same as on the left, but illustrating that χ is independent of ζ.

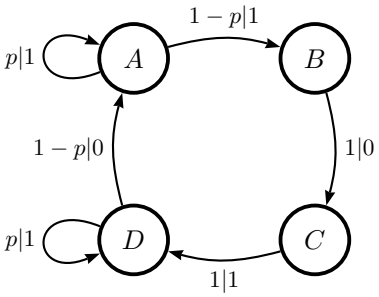

FIG. 15. A nonunifilar presentation of the Golden Mean Process.

of the gauge information (1 bit) is easy to understand. It indicates that the nonunifilar presentation is two copies of a unifilar presentation of the Golden Mean Process sutured together. All of history space is covered twice and the choice of which component of the cover is visited is a fair coin flip. The crypticity and oracular information (crypticity's time-reversed analog) are the same, due to the nonunifilar presentation respecting the time-reverse symmetry of the Golden Mean Process [23].

VIII. CONCLUSIONS

Our development started out discussing synchronization and control. The tools required to address these—the block-state and state-block entropies—quickly led to a substantially enlarged view of the space of competing models, the rival presentations, and a new collection of information measures that reflect their subtleties and differences.

As milestones along the way, we gave example presentations of the well-known Golden Mean Process that went from the ϵ-machine to a nonminimal nonsynchronizing nonunifilar presentation. Table I summarizes the quantitative results. It gives the entropy rate h_μ, statistical complexity C_μ, excess entropy $\mathbf{E}$, and the crypticity χ for the process itself. Immediately following, it compares the analogous measures for the range of presentations considered. In addition, the gauge information φ and the oracular information ζ, being properties of presentations, are added. Careful study of the table shows how the measures track the presentations' structural changes.

A few comments are in order about the tools the development required. The first were the block-state and state-block entropies, as noted. Analyzing their word-length convergence properties was the backbone of the approach—one directly paralleling the previously introduced entropy convergence hierarchy [12]. Another important tool was the I-diagram. While it is not necessary in establishing final results, it is immensely helpful in organizing one's thinking and in managing the complications of multivariate information measures. Methodologically speaking, the principal subject was the four-variable—past, future, causal state, and presentation state—I-diagram with its sigma algebra of 15 atoms. Thus, the methodology of the development turned on just two tools—block entropy convergence and presentation information measures.

As for the concrete results, we showed that there are two mechanisms operating in processes that are hard to synchronize to, as measured by the synchronization information which consists of two corresponding independent contributions. The first is the transient information which reflects entropy-rate overestimates that occur at small block lengths. The second, new here, reflects the state information that is not retrodictable using the future. With these two contributions laid out, the general connection between synchronization and transient information, previously introduced in Ref. [12], became clear. We also pointed out that the synchronization information

Information Measures for Alternative Presentations

Process	h_μ	C_μ	$\mathbf{E}$	χ
Golden Mean	$2/3$	$\log_2(3) - 2/3$	$\log_2(3) - 4/3$	$2/3$

Presentation	$H[X\|\mathcal{R}]$	$H[\mathcal{R}]$	$I[\mathcal{R};\overrightarrow{X}]$	$\chi(\mathcal{R})$	φ	ζ
ϵ-Machine	h_μ	C_μ	$\mathbf{E}$	χ	0	0
Synchronizable	h_μ	$\log_2(3)$	$\mathbf{E}$	$4/3$	0	0
Unifilar	h_μ	$\log_2(3) + 1/3$	$\mathbf{E}$	$5/3$	1	0
Nonunifilar	$1/3$	$\log_2(3) + 1/3$	$\log_2(3) - 1$	$1/3$	1	$1/3$

TABLE I. Comparison of information measures for presentations of the Golden Mean Process with transition parameter $p = 1/2$.

for nonsynchronizing presentations can diverge. This, in turn, called for a generalized definition of synchronization appropriate to all presentations.

We also generalized the process crypticity, beyond the domain of ϵ-machine optimal presentations, to describe the amount of presentation state information that is shared with the past but not transmitted to the future. A sibling of the crypticity, we introduced a new information measure for generic presentations—the oracular information—that is the amount of state information shared with the future, but not derivable from the past.

Finally, to account for "components", either explicitly or implicitly included in a presentation, that are not justified by the process statistics, we introduced the gauge information, intentionally drawing a parallel to the concept of gauge degrees of freedom familiar from physics.

One immediate result was that the information measures allowed us to delineate the hierarchy of a process's presentations. The hierarchy goes from the unique, minimal unifilar, optimal predictor (ϵ-machine) to nonminimal unifilar, weakly asymptotically synchronizing presentations to nonsynchronizing, unifilar presentations. We showed these are nested classes. Stepping outside to the nonunifilar presentations leaves one in a markedly larger class for which all of the information measures play a necessary role.

We trust that the presentation hierarchy makes the singular role of the ϵ-machine transparent. First, the ϵ-machine's minimality and uniqueness are those of the corresponding process. This cannot be said for alternative presentations. Second, there is a wide range of properties that can be efficiently calculated, when alternative presentations may preclude this. One cannot calculate a process's stored information (C_μ) or information production rate (h_μ) from, for example, nonunifilar presentations. The latter must be converted, either directly or indirectly, to the process's ϵ-machine to calculate them.

Nonetheless, as discussed at some length in Ref. [27], in varying circumstances—limited material, inference, or compute-time resources; ready access to sources of ideal randomness; noisy implementation substrates; and the like—the ϵ-machine may not be how an observer should model a process. A minimal nonunifilar presentation, that is necessarily more stochastic internally than the ϵ-machine [29], may be preferred due to it having a smaller set of states.

Recalling the duality of synchronization and control, we close by noting that essentially all of the results here apply to the setting in which an agent attempts to steer a process into desired states. The efficiency with which the control signals achieve this is reflected in the analogue of block entropy convergence. The very possibility of control has its counterparts in an implementation hierarchy that mirrors the presentation hierarchy, but with controllability instead of synchronizability.

Appendix A: Notation Change for Total Predictability

The definition for $\mathcal{I}_n$ in Eq. (14)—the total predictability—represents a minor change in notation from Ref. [12]. (We refer to the latter as RURO, abbreviating its title.) There, the minimum L was usually n except for $n = 2$, when the minimum L value was $L = 1$ instead. One reason for the change in definition is that $\mathcal{I}_2$ now does not depend on any assumption (prior) for symbol entropy rate and depends *only* on asymptotic properties of the process.

To make this explicit, note that the original definition of total predictability contained a boundary term:

$$\mathbf{G}_{\mathrm{RURO}} = \Delta^2 H(1) + \sum_{L=2} \Delta^2 H(L) \ , \tag{A1}$$

where

$$\Delta^2 H(1) = h_\mu(1) - h_\mu(0) = H(1) - \log_2 |\mathcal{A}| \ . \quad \text{(A2)}$$

The logarithm term characterized the entropy rate estimate before any probabilities are considered. In the modified definition of total predictability, we drop the boundary term, giving:

$$\mathbf{G} \equiv \mathcal{I}_2 = \sum_{L=2} \Delta^2 H(L) \ . \quad \text{(A3)}$$

The two quantities are related by:

$$\begin{aligned} \mathbf{G}_{\text{RURO}} &= \mathbf{G} + \Delta^2 H(1) && \text{(A4)} \\ &= \mathbf{G} + H(1) - \log_2 |\mathcal{A}| \ . && \text{(A5)} \end{aligned}$$

This affects relationships involving $\mathbf{G}$. Previously, for example,

$$\mathbf{G}_{\text{RURO}} = -\mathbf{R} \leq 0 \ , \quad \text{(A6)}$$

where $\mathbf{R}$ is the total redundancy. Now,

$$\begin{aligned} \mathbf{G} &= -\mathbf{R} - \Delta^2 H(1) && \text{(A7)} \\ &= \log_2 |\mathcal{A}| - H(1) - \mathbf{R} \ . && \text{(A8)} \end{aligned}$$

Appendix B: State-Block Entropy Rate Estimate

In this section, we prove Thm. 1, which states that $H[X_L|\mathcal{R}_0, X_0^L]$ converges monotonically (nondecreasing) to the entropy rate.

Proof. *First, we show that difference in the $H[\mathcal{R}_0, X_0^L]$ forms a nondecreasing sequence:*

$$\begin{aligned} & H[X_{L-1}|\mathcal{R}_0, X_0^{L-1}] && \text{(B1)} \\ &= H[X_L|\mathcal{R}_1, X_1^{L-1}] && \text{(B2)} \\ &= H[X_L|\mathcal{R}_0, X_0, \mathcal{R}_1, X_1^{L-1}] && \text{(B3)} \\ &\leq H[X_L|\mathcal{R}_0, X_0, X_1^{L-1}] && \text{(B4)} \\ &= H[X_L|\mathcal{R}_0, X_0^L] \ . && \text{(B5)} \end{aligned}$$

Next, we show this sequence is bounded and, thus, has a limit. For all $k \geq 0$, we have:

$$\begin{aligned} & H[X_L|\mathcal{R}_0, X_0^L] && \text{(B6)} \\ &= H[X_L|X_{-k}^k, \mathcal{R}_0, X_0^L] && \text{(B7)} \\ &\leq H[X_L|X_{-k}^{L+k}] && \text{(B8)} \\ &= H[X_{L+k}|X_0^{L+k}] \ . && \text{(B9)} \end{aligned}$$

Since this holds for all k, it also holds in the limit as k tends to infinity, which is the definition of the entropy rate. Thus, $H[X_L|\mathcal{R}_0, X_0^L]$ is a nondecreasing sequence and bounded above by h_μ.

Finally, we show that this bounded sequence converges to h_μ. To do this, we will show that the difference

$$H[X_L|X_0^L] - H[X_L|\mathcal{R}_0, X_0^L] = I[\mathcal{R}_0; X_L|X_0^L]$$

converges to zero. Then, since the first term (differences in the block entropies) is known to converge to the entropy rate, the claim will be proved. We have:

$$\begin{aligned} H[\mathcal{R}_0] &\geq \lim_{L\to\infty} I[\mathcal{R}_0; X_0^L, X_L] && \text{(B10)} \\ &= \lim_{L\to\infty} \sum_{k=0}^{L} I[\mathcal{R}_0; X_k|X_0^k] \ . && \text{(B11)} \end{aligned}$$

Since the sum is finite, the terms must tend to zero. □

Appendix C: Reducing the Presentation I-Diagram

Proving that the various multivariate information measures vanish makes use of a few facts about states:

- $H[\mathcal{S}|\overleftarrow{X}] = 0$.
- $H[\overleftarrow{X}; \overrightarrow{X}|\mathcal{S}] = 0$.
- $I[\overleftarrow{X}; \overrightarrow{X}|\mathcal{R}] = H[\overrightarrow{X}|\mathcal{R}] - H[\overrightarrow{X}|\mathcal{R}, \overleftarrow{X}] = 0$.

The last one follows from limiting ourselves to states that actually generate the process. Thus, additional conditioning on the past cannot influence the future, as the current state alone determines the future.

The following atoms vanish:

- $H[\mathcal{S}|\overleftarrow{X}, \mathcal{R}, \overrightarrow{X}]$:

$$H[\mathcal{S}|\overleftarrow{X}, \mathcal{R}, \overrightarrow{X}] \leq H[\mathcal{S}|\overleftarrow{X}] = 0 \ .$$

- $I[\mathcal{S}; \mathcal{R}|\overleftarrow{X}, \overrightarrow{X}]$:

$$\begin{aligned} I[\mathcal{S}; \mathcal{R}|\overleftarrow{X}, \overrightarrow{X}] &= H[\mathcal{S}|\overleftarrow{X}, \overrightarrow{X}] - H[\mathcal{S}|\overleftarrow{X}, \mathcal{R}, \overrightarrow{X}] \\ &= H[\mathcal{S}|\overleftarrow{X}, \overrightarrow{X}] - 0 \\ &\leq H[\mathcal{S}|\overleftarrow{X}] \\ &= 0 \ . \end{aligned}$$

- $I[\mathcal{S}; \mathcal{R}; \overrightarrow{X}|\overleftarrow{X}]$:

$$\begin{aligned} I[\mathcal{S}; \mathcal{R}; \overrightarrow{X}|\overleftarrow{X}] &= I[\mathcal{S}; \mathcal{R}|\overleftarrow{X}] - I[\mathcal{S}; \mathcal{R}|\overleftarrow{X}, \overrightarrow{X}] \\ &= I[\mathcal{S}; \mathcal{R}|\overleftarrow{X}] - 0 \\ &= H[\mathcal{S}|\overleftarrow{X}] - H[\mathcal{S}|\mathcal{R}, \overleftarrow{X}] \\ &= 0 - H[\mathcal{S}|\mathcal{R}, \overleftarrow{X}] \ . \end{aligned}$$

Finally, note that

$$\begin{aligned} |H[\mathcal{S}|\mathcal{R}, \overleftarrow{X}]| &\leq |H[\mathcal{S}|\overleftarrow{X}]| \\ &= 0 \ . \end{aligned}$$

- $I[\mathcal{S}; \overrightarrow{X}|\overleftarrow{X}, \mathcal{R}]$:

$$\begin{aligned} I[\mathcal{S}; \overrightarrow{X}|\overleftarrow{X}, \mathcal{R}] &= H[\mathcal{S}|\overleftarrow{X}, \mathcal{R}] - H[\mathcal{S}|\overleftarrow{X}, \overrightarrow{X}, \mathcal{R}] \\ &= H[\mathcal{S}|\overleftarrow{X}, \mathcal{R}] - 0 \\ &\leq H[\mathcal{S}|\overleftarrow{X}] \\ &= 0 \ . \end{aligned}$$

- $I[\overleftarrow{X}; \overrightarrow{X}|\mathcal{S}, \mathcal{R}]$:

$$\begin{aligned} I[\overleftarrow{X}; \overrightarrow{X}|\mathcal{S}, \mathcal{R}] &= H[\overrightarrow{X}|\mathcal{S}, \mathcal{R}] - H[\overrightarrow{X}|\mathcal{S}, \mathcal{R}, \overleftarrow{X}] \\ &= H[\overrightarrow{X}|\mathcal{S}, \mathcal{R}] - H[\overrightarrow{X}|\mathcal{S}, \mathcal{R}] \\ &= 0 \ . \end{aligned}$$

- $I[\overleftarrow{X}; \mathcal{R}; \overrightarrow{X}|\mathcal{S}]$:

$$\begin{aligned} I[\overleftarrow{X}; \mathcal{R}; \overrightarrow{X}|\mathcal{S}] &= I[\overleftarrow{X}; \overrightarrow{X}|\mathcal{S}] - I[\overleftarrow{X}; \overrightarrow{X}|\mathcal{S}, \mathcal{R}] \\ &= 0 \ . \end{aligned}$$

- $I[\overleftarrow{X}; \mathcal{S}; \overrightarrow{X}|\mathcal{R}]$:

$$\begin{aligned} I[\overleftarrow{X}; \mathcal{S}; \overrightarrow{X}|\mathcal{R}] &= I[\overleftarrow{X}; \overrightarrow{X}|\mathcal{R}] - I[\overleftarrow{X}; \overrightarrow{X}|\mathcal{S}, \mathcal{R}] \\ &= 0 \ . \end{aligned}$$

The first four vanish due to the causal states being a function of the past. The last three vanish since any presentation that generates the process captures all the information shared between past and future.

ACKNOWLEDGMENTS

This work was partially supported by the DARPA Physical Intelligence Program. The authors thank Dave Feldman, Nick Travers, and Luke Grecki for helpful comments on the manuscript.

[1] J. P. Crutchfield and D. P. Feldman. Synchronizing to the environment: Information theoretic limits on agent learning. *Adv. in Complex Systems*, 4(2):251–264, 2001.

[2] S. Wiggins. *Chaotic Transport in Dynamical Systems*. Springer, New York, 1992.

[3] R. W. Yeung. A new outlook on Shannon's information measures. *IEEE Trans. Info. Th.*, 37(3):466–474, 1991.

[4] J. Klamka. *Controllability of Dynamical Systems*. Springer, New York, 1991.

[5] R. J. Elliott, L. Aggoun, and J. B. Moore. *Hidden Markov models: Estimation and control*. Springer, New York, 1994.

[6] B. R. Andrievskii and A. L. Fradkov. Control of chaos: Methods and Applications. I. Methods. *Automation and Control*, 64(5):673–713, 2004.

[7] B. R. Andrievskii and A. L. Fradkov. Control of chaos: Methods and Applications. II. Applications. *Automation and Control*, 65(4):505–533, 2004.

[8] J. M. Gonzalez-Miranda. *Synchronization and Control of Chaos: An Introduction for Scientists and Engineers*. World Scientific, Singapore, 2004.

[9] A. Pikovsky and J. Kurths M. Rosenblum. *Synchronization: A Universal Concept in Nonlinear Sciences*. Cambridge Nonlinear Science Series. Cambridge University Press, New York, 2001.

[10] N. Jonoska. Sofic shifts with synchronizing presentations. *Theo. Comp. Sci.*, 158:81–115, 1996.

[11] S. Strogatz. *Sync: The Emerging Science of Spontaneous Order*. Hyperion, New York, 2003.

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

[13] D. P. Feldman and J. P. Crutchfield. Synchronizing to periodicity: The transient information and synchronization time of periodic sequences. *Advances in Complex Systems*, 7(3-4):329–355, 2004.

[14] H. Marko. The bidirectional communication theory: A generalization of information theory. *IEEE Trans. Comm.*, COM-21(12):1345–135, 1973.

[15] X. Feng, K. A. Loparo, , and Y. Fang. Optimal state estimation for stochastic systems: An information theoretic approach. *IEEE Trans. Auto. Control*, 42(6):771–785, 1997.

[16] N. U. Ahmed. *Linear and Nonlinear Filtering for Engineers and Scientists*. World Scientific Publishers, Singapore, 1998.

[17] N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw. Geometry from a time series. *Phys. Rev. Let.*, 45:712, 1980.

[18] F. Takens. Detecting strange attractors in fluid turbulence. In D. A. Rand and L. S. Young, editors, *Symposium on Dynamical Systems and Turbulence*, volume 898, page 366, Berlin, 1981. Springer-Verlag.

[19] E. Ott, B.R. Hunt, I. Szunyogh aand A.V. Zimin, E. J. Kostelich, M. Corazza, E. Kalnay, D. J. Patil, and J. A.

Yorke. Estimating the state of large spatio-temporally chaotic systems. *Physics Letters A*, 330:365–370, 2004.

[20] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley-Interscience, New York, 2005.

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

[22] 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.

[23] 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.

[24] J. R. Mahoney, C. J. Ellison, and J. P. Crutchfield. Information accessibility and cryptic processes. *J. Phys. A: Math. Theo.*, 42:362002, 2009.

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

[26] A process's causal states consist of both transient and recurrent states. To simplify the presentation, we henceforth refer *only* to recurrent causal states.

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

[28] J. P. Crutchfield and C. R. Shalizi. Thermodynamic depth of causal states: Objective complexity via minimal representations. *Phys. Rev. E*, 59(1):275–283, 1999.

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

[30] In the theory of computation, unifilar is referred to as "deterministic" [43].

[31] Specifically, each transition matrix $T^{(x)}$ has, at most, one nonzero component in each row.

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

[33] S. Still, J. P. Crutchfield, and C. J. Ellison. Optimal causal inference: Estimating store information and approximating causal architecture. *CHAOS*, page in press, 2010.

[34] A stochastic mapping, known as a mixed state, is discussed in [23].

[35] J. R. Mahoney, C. J. Ellison, Ryan G. James, and J. P. Crutchfield. *in preparation*, 2010. arxiv.org:10XX.XXX [cond-mat].

[36] N. Travers and J. P. Crutchfield. Exact synchronization for finite-state sources. 2010. SFI Working Paper 10-08-XXX; arxiv.org:10XX.XXXX [XXXX].

[37] N. Travers and J. P. Crutchfield. Asymptotically synchronizing to finite-state sources. 2010. SFI Working Paper 10-09-XXX; arxiv.org:10XX.XXXX [XXXX].

[38] Ryan G. James, J. R. Mahoney, C. J. Ellison, and J. P. Crutchfield. *in preparation*, 2010. arxiv.org:10XX.XXX [cond-mat].

[39] P. H. Frampton. *Gauge Field Theories*. Wiley-VGH Verlag, Weinheim, 2008.

[40] 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.

[41] Oracular information cannot be extracted from the past observables. This point will be discussed further in Sec. VII.

[42] Similar observations appeared recently; for example, see Ref. [44]. In a sequel we compare this to the earlier results of Refs. [27, 40].

[43] J. E. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley, Reading, 1979.

[44] W. Loehr and N. Ay. Non-sufficient memories that are sufficient for prediction. In J. Zhou, editor, *Complex Sciences 2009*, volume 4 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pages 265–276. Springer, New York, 2009.