Title
Elusive present: Hidden past and future dependency and why we build models

Journal
Physical Review E, 93(2)

ISSN
2470-0045

Authors
Ara, Pooneh M
James, Ryan G
Crutchfield, James P

Publication Date
2016-02-01

DOI
10.1103/physreve.93.022143

Peer reviewed

Elusive present: Hidden past and future dependency and why we build models

Pooneh M. Ara,* Ryan G. James,† and James P. Crutchfield‡
Complexity Sciences Center and Physics Department, University of California, Davis, One Shields Avenue, Davis, California 95616, USA
(Received 2 July 2015; revised manuscript received 30 January 2016; published 29 February 2016)

Modeling a temporal process as if it is Markovian assumes that the present encodes all of a process's history. When this occurs, the present captures all of the dependency between past and future. We recently showed that if one randomly samples in the space of structured processes, this is almost never the case. So, how does the Markov failure come about? That is, how do individual measurements fail to encode the past? and How many are needed to capture dependencies between the past and future? Here, we investigate how much information can be shared between the past and the future but not reflected in the present. We quantify this elusive information, give explicit calculational methods, and outline the consequences, the most important of which is that when the present hides past-future correlation or dependency we must move beyond sequence-based statistics and build state-based models.

DOI: 10.1103/PhysRevE.93.022143

I. INTRODUCTION

Until the turn of the nineteenth century, temporal processes were almost exclusively considered to be independently sampled at each time from the same statistical distribution: each sample uncorrelated with its predecessor. These studies were initiated by Jacob Bernoulli in the 1700s [1] and refined by Simeon Poisson [2] and Pafnuty Chebyshev [3] in the 1800s, leading to the weak law of large numbers and the central limit theorem. These powerful results were the first hints at universal laws in stochastic processes, but they applied only to independent, identically distributed processes—unstructured processes with no temporal correlation, no memory. Moreover, until the turn of the century it was believed that these laws required independence. It fell to Andrei Andreevich Markov (1856–1922) to realize that independence is not necessary. To show this he introduced a new kind of sequence or "chain" of dependent random variables, along with the concepts of transition probabilities, irreducibility, and stationarity [4,5].

Introducing his "complex chains" in 1907, Markov initiated the modern study of structured, interdependent, and correlated processes. Indeed, in the first and now-famous application of complex chains, he analyzed the pair distribution (2-gram) in the 20 000 vowels and consonants in Pushkin's poem *Eugeny Onegin* and the 100 000 letters in Aksakov's novel *The Childhood of Bagrov, the Grandson* [6,7]. Since Markov's time the study of complex chains has developed into one of the most powerful mathematical tools, applied far beyond quantitative linguistics in physics [8], chemistry [9], biology [10], finance [11], and even numerical methods of estimation and optimization [12] and Google's PageRank algorithm [13].

To study correlation in structured processes, we take an information-theoretic view of Markov's concept of complexity—that arising from temporal interdependency between observed symbols that are functions of chain states. That is, individual observation symbols are not themselves the chain states and therefore need not encode all the past. Specifically, we consider stationary, ergodic processes generated by hidden Markov chains: introduced in the midtwentieth century, as a generalization of Markov's complex chains, to model processes generated by communication channels [14]. When are these hidden processes described by finite Markov chains? When are they not Markovian? What is the informational signature in this case? and What are "states" in the first place? Can we discover them from observations of a hidden process?

The following is the first in a series that addresses these questions: which have been answered, which can be answered, and which are open. Here, we concentrate on how the present—a sequence of ℓ consecutive measurements—statistically shields the past from the future, as a function of ℓ. We introduce the elusivity σ_μ^ℓ as a quantitative measure of the present failing to encode the past—a quantitative signature of Markov failure. We show how to calculate it explicitly via a novel construction and then describe and interpret its behavior through examples. The methods introduced also lead to compact expressions and efficient estimation of related measures—ephemeral, bound, and enigmatic informations—whose behaviors we also explore. As an application we use the results to reinterpret the persistent mutual information (PMI) introduced in Ref. [15] as a measure of "emergence" in complex systems. Finally, we note that the sequel [16] is analytical, giving closed-form solutions and proving various properties, including several of those used here.

To address Markov's notion of complex chains, the next section reviews the minimal necessary background: measures of information content and correlation from information theory [17], their application to stochastic processes via computational mechanics [18,19], and a recent analysis of the information content of the single-time-step present within the context of the past and future [20]. This then sets the stage for a thorough analysis of Markovian complexity: generalizing the previous framework so that the present can be an arbitrary duration. This gives rise to our main new result: expressing the elusivity in terms of a process's causal states. Notably, this result draws on the prior introduction of stochastic process models—the so-called bimachines—that are agnostic with respect to the direction of time. We show how this leads to a simple and efficient expression for the elusive information

*mohammadiara@ucdavis.edu
†rgjames@ucdavis.edu
‡Corresponding author: chaos@ucdavis.edu

and its companion measures. Those expressions, in turn, give the basis for further analytical development and for empirical estimation. With the general theory laid out, we make the ideas and methods concrete by calculating the elusivity and related quantities for a number of prototype complex processes, characterizing the variety of their convergence behaviors. Finally, we apply the insights gained to evaluate a proposed information-theoretic measure of emergence. We close by recapping the results and drawing conclusions for future applications.

II. INFORMATION IN COMPLEX PROCESSES

A. Processes

We are interested in a general stochastic process $\mathcal{P}$: the distribution of all of a system's behaviors or realizations $\{\ldots x_{-2},x_{-1},x_0,x_1,\ldots\}$ as specified by their joint probabilities $\Pr(\ldots X_{-2},X_{-1},X_0,X_1,\ldots)$. X_t is a random variable that is the outcome of the measurement at time t, taking the values x_t from a finite set $\mathcal{A}$ of all possible events. We denote a contiguous chain of random variables $X_{0:\ell} = X_0 X_1 \ldots X_{\ell-1}$. Left indices are inclusive; right indices, exclusive. We suppress indices that are infinite. We consider only stationary processes for which $\Pr(X_{t:t+\ell}) = \Pr(X_{0:\ell})$ for all t and ℓ.

Our particular emphasis in the following is that a process $\Pr(X_{:0},X_{0:\ell},X_{\ell:})$ is a communication channel that transfers information from the past $X_{:0} = \ldots X_{-3}X_{-2}X_{-1}$ to the future $X_{\ell:} = X_\ell X_{\ell+1} X_{\ell+2}\ldots$ by storing parts of it in the present $X_{0:\ell} = X_0 X_1 \ldots X_{\ell-1}$ of length ℓ. Of primary concern is whether $X_{:0} \to X_{0:\ell} \to X_{\ell:}$ forms a Markov chain in the sense of Ref. [21]:

$$\begin{aligned}&\Pr(X_{-m:0},X_{0:\ell},X_{\ell:n})\\&\quad= \Pr(X_{-m:0}|X_{0:\ell})\Pr(X_{\ell:n}|X_{0:\ell})\Pr(X_{0:\ell})\end{aligned}$$

for all $m,n \in \mathbb{Z}^+$.

B. Channel information

In analyzing this channel we need to measure a variety of information metrics, each capturing a different aspect of the information being communicated. The simplest is the *Shannon entropy* [21]:

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

Three other information-theoretic measures based on the entropy are employed throughout. First, the *conditional entropy*, measuring the amount of information remaining in a variable X (alphabet $\mathcal{X}$) once the information in a variable Y (alphabet $\mathcal{Y}$) is accounted for:

$$\mathrm{H}[X|Y] = -\sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} \Pr(x,y)\log_2 \Pr(x|y). \tag{2}$$

Second, the deficiency of the conditional entropy relative to the full entropy is known as the *mutual information*, characterizing the information that is contained in both X and Y:

$$\begin{aligned}\mathbf{I}[X:Y] &= \mathrm{H}[X] - \mathrm{H}[X|Y]\\ &= \sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} \Pr(x,y)\log_2\frac{\Pr(x,y)}{\Pr(x)\Pr(y)}.\end{aligned} \tag{3}$$

Finally, we have the *conditional mutual information*, the mutual information between two variables once the information in a third (Z with alphabet $\mathcal{Z}$) has been accounted for:

$$\mathbf{I}[X:Y|Z] = \sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}\\ z\in\mathcal{Z}}} \Pr(x,y,z)\log_2\frac{\Pr(x,y|z)}{\Pr(x|z)\Pr(y|z)}. \tag{4}$$

Perhaps the most naïve way of information-theoretically analyzing a process, capturing the randomness and dependencies in sequences of random variables, is via the *block entropies*:

$$\mathrm{H}(\ell) = \mathrm{H}[X_{0:\ell}]. \tag{5}$$

This quantifies the amount of information in a contiguous block of observations. Its growth with ℓ gives insight into a process's randomness and structure [18,22]:

$$\mathrm{H}(\ell) \approx \mathbf{E} + h_\mu \ell, \quad \ell \gg 1. \tag{6}$$

The asymptotic growth h_μ, here, is a process's rate of information generation, or the Shannon entropy rate:

$$h_\mu = \mathrm{H}[X_0|X_{:0}], \tag{7}$$

where the subscript μ denotes the specific probability measure over bi-infinite strings, defining the process of interest and its sequence probabilities. Finally, the amount of future information predictable from the past is the past-future mutual information or *excess entropy*:

$$\begin{aligned}\mathbf{E} &= \mathbf{I}[X_{:0}:X_{0:}]\\ &= \mathrm{H}[X_{:0}] - \mathrm{H}[X_{:0}|X_{0:}].\end{aligned} \tag{8}$$

The excess entropy naturally arises when considering channels with a length $\ell = 0$ present, where it is effectively the only direct information quantity over the variables $X_{:0}$ and $X_{0:}$. It is well known that if the excess entropy vanishes, then there is no information temporally communicated by the channel [18].

Generically, Eq. (8) is of the form $\infty - \infty$, which is meaningless. In such situations one refers to finite sequences and then takes a limit:

$$\lim_{m,n\to\infty}(\mathrm{H}[X_{-m:0}] - \mathrm{H}[X_{-m:0}|X_{0:n}]).$$

Here, we generally use the informal infinite variables in equations for clarity and simplicity unless the details of the limit are important for the analysis at hand. To be concrete, we write $f(X_{:0})$ to mean $\lim_{m\to\infty} f(X_{-m:0})$ and $f(X_{\ell:})$ to mean $\lim_{n\to\infty} f(X_{\ell:n})$.

C. Information atoms

The foregoing setup views a process as a channel that communicates the past to the future via the present. Our goal, then, is to analyze the channel's properties as a function of the present's length ℓ. The cases of $\ell = 0$ and $\ell = 1$ have been addressed: $\ell = 0$ in Ref. [23] and $\ell = 1$ in Ref. [20]. Our initial development closely mirrors theirs. We borrow notation but must include a superscript to denote the ℓ dependence of the quantities. Broadly, our approach is to decompose the information in a random variable—such as the present—into components (atoms) associated with other random variables.[1] Our immediate concern is that of monitoring the amount of dependency remaining between the past and the future if the present is known. We use the mutual information between the past and the future conditioned on the present to do so—the elusivity that soon becomes our focus:

$$\begin{aligned}\sigma_\mu^\ell &= \mathbf{I}[X_{:0} : X_{\ell:} | X_{0:\ell}] \\ &= \mathrm{H}[X_{:0}|X_{0:\ell}] + \mathrm{H}[X_{\ell:}|X_{0:\ell}] - \mathrm{H}[X_{:0}, X_{\ell:}|X_{0:\ell}]. \end{aligned} \tag{9}$$

Note that $\sigma_\mu^0 = \mathbf{E}$.

Next, extending Ref. [20], we decompose the length ℓ present. When considering only the past, the information in the present separates into two components: $\rho_\mu^\ell = \mathbf{I}[X_{:0} : X_{0:\ell}]$, the information that can be anticipated from the past; and $h_\mu^\ell = \mathrm{H}[X_{0:\ell}|X_{:0}]$, the random component that cannot be anticipated. Naturally, $\mathrm{H}[X_{0:\ell}] = h_\mu^\ell + \rho_\mu^\ell$. Connecting directly to Ref. [20], our ρ_μ^1 is their ρ_μ, and likewise, our h_μ^1 is their h_μ.

If one also accounts for the future's behavior, then the random, unanticipated component h_μ^ℓ breaks into two kinds of information: one part, $b_\mu^\ell = \mathbf{I}[X_{0:\ell} : X_{\ell:}|X_{:0}]$, that, while a degree of randomness, is relevant for predicting the future; and the remaining part, $r_\mu^\ell = \mathrm{H}[X_{0:\ell}|X_{:0}, X_{\ell:}]$, which is ephemeral, existing only fleetingly in the present and then dissipating, leaving no trace on future behavior.

The redundant portion ρ_μ^ℓ of $\mathrm{H}[X_{0:\ell}]$ itself splits into two pieces. The first part, $\mathbf{I}[X_{:0} : X_{0:\ell}|X_{\ell:}]$—also b_μ^ℓ when the process is stationary—is shared between the past and the current observation, but its relevance stops there. The second piece, $q_\mu^\ell = \mathbf{I}[X_{:0} : X_{0:\ell} : X_{\ell:}]$, is anticipated by the past, is present currently, and also plays a role in future behavior. Notably, this informational piece can be negative [20,25].

Due to a duality between set-theoretic and information-theoretic operators, we can graphically represent the relationship between these various pieces of information in a Venn-like display called an *information diagram* [26]; see Fig. 1. In contrast to a Venn diagram, size indicates Shannon entropy rather than set cardinality and overlaps are not set intersection but mutual information. Each area on the diagram represents one or another of Shannon's information measures.

As mentioned above, the past splits $\mathrm{H}[X_{0:\ell}]$, yielding two pieces: h_μ^ℓ, the part outside the past; and ρ_μ^ℓ, the part inside. This partitioning arises naturally when predicting a process [20]. To emphasize, Fig. 2(a) displays this decomposition. If we include the future in the diagram, we obtain a more detailed understanding of how information is transmitted from

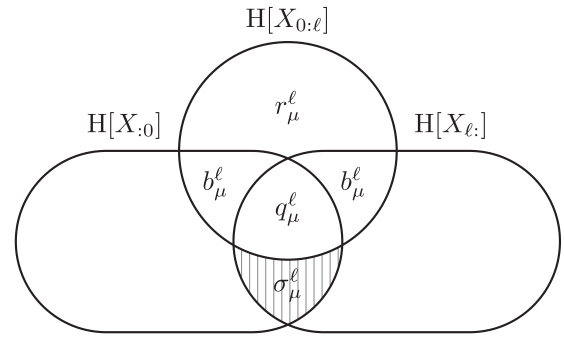

FIG. 1. The process information diagram, which places the present in its temporal context: the past ($X_{:0}$) and the future ($X_{\ell:}$) partition the present ($X_{0:\ell}$) into four components with quantities r_μ^ℓ and q_μ^ℓ and two with b_μ^ℓ. Notably, the component σ_μ^ℓ, quantifying the hidden dependency shared by the past and the future, lies outside of the present and so is not part of it.

the past to the future. The past and the future together divide the present $\mathrm{H}[X_{0:\ell}]$ into four parts, as shown in Fig. 2(b).

The process information diagram makes it rather transparent in which sense r_μ^ℓ is an amount of *ephemeral information*: its information lies outside both the past and the future and so it exists only in the present moment. It has no repercussions for the future and is no consequence of the past. It is the amount of information in the present observation neither communicated to the future nor from the past. With $\ell = 1$, this has been referred to as the residual entropy rate [27], as it is the amount of uncertainty that remains in the present even after accounting for every other variable in the time series. It has also been studied as the erasure information [28] (there H^-), as it is the information irrecoverably erased in a binary erasure channel.

The *bound information* b_μ^ℓ is the amount of spontaneously generated information present now, not explained by the past, but that has consequences for the future. In this sense it hints at being a measure of structural complexity [20,27], though we discuss more direct measures of structure shortly.

Due to stationarity, the mutual information $\mathbf{I}[X_{0:\ell} : X_{\ell:}|X_{:0}]$ between the present $X_{0:\ell}$ and the future $X_{\ell:}$ conditioned on the past $X_{:0}$ is the same as the mutual information $\mathbf{I}[X_{0:\ell} : X_{:0}|X_{\ell:}]$ between the present $X_{0:\ell}$ and the past $X_{:0}$

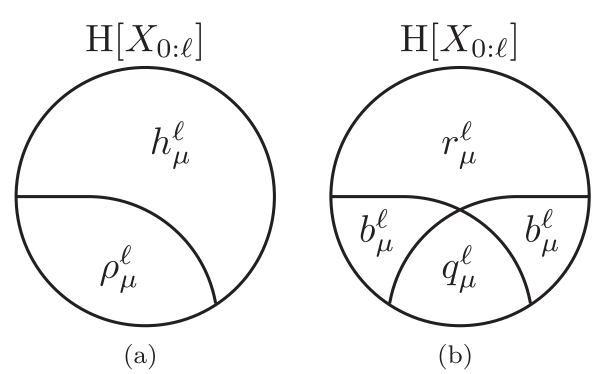

FIG. 2. Alternative decompositions of the present information $\mathrm{H}[X_{0:\ell}]$. (a) Decomposition due to the past, (b) decomposition due to the past and the future.

[1] It is important to note that we are only associating portions of a random variable's information with other random variables. It is generically not possible to actually partition a random variable into several other random variables, each of which has an entropy equal to that of the atom of interest [24].

conditioned on the future $X_{\ell:}$. Therefore they are both of size b_μ^ℓ, as shown in Fig. 1. This lends a symmetry to the process information diagram that need not exist for nonstationary processes.

D. Elusivity

Two components remain in the process information diagram: two that have not been significantly analyzed previously. The first is $q_\mu^\ell = \mathbf{I}[X_{:0} : X_{0:\ell} : X_{\ell:}]$—the three-way mutual information (or *co-information* [25]) shared by the past, present, and future. Notably, unlike Shannon entropies and two-way mutual information, q_μ^ℓ (and co-information in general), can be negative. The other component, $\sigma_\mu^\ell = \mathbf{I}[X_{:0} : X_{\ell:}|X_{0:\ell}]$, the quantity of primary interest here (shaded in Fig. 1), is the information shared between the past and the future that does not exist in the present. Since it measures dependency hidden from the present, we call it the *elusive information,* or *elusivity* for short. Generally, it indicates that a process has hidden structures that are not appropriately captured by the present—that is, by finite random-variable blocks. In this case, and as we discuss at length towards the end, one must build models whose elements, which we call "states" below, represent how a process's internal mechanism is organized.

A process's internal organization somehow must store all the information from the past that is relevant for generating the future behavior. Only when the observed process is Markovian is it sufficient to keep track of just the current observable or block of observables. For the general case of non-Markovian processes, though, information relevant for prediction is spread arbitrarily far back in the process's history and so cannot be captured by the present regardless of its duration. This fact is reflected in the existence of σ_μ^ℓ. When $\sigma_\mu^\ell > 0$ for all ℓ, the description of the process requires determining its internal organization. This is one reason to build a model of the mechanism that generates sequences rather than simply describing a process as a list of sequences.

There are two basic properties that indicate the elusivity's importance. The first is that σ_μ^ℓ decreases monotonically as a function of the present's length ℓ. That is, dependency cannot increase if we interpolate more random variables between the past and the future.

Proposition 1. For $k > 0$, $\sigma_\mu^\ell \geqslant \sigma_\mu^{\ell+k}$.

Proof.

$$\begin{aligned}\sigma_\mu^\ell &= \mathbf{I}[X_{:0} : X_{\ell:}|X_{0:\ell}]\\ &= \mathbf{I}[X_{:0} : (X_{\ell:\ell+k}, X_{\ell+k:})|X_{0:\ell}]\\ &= \mathbf{I}[X_{:0} : X_{\ell:\ell+k}|X_{0:\ell}] + \mathbf{I}[X_{:0} : X_{\ell+k:}|X_{0:\ell}, X_{\ell:\ell+k}]\\ &= \mathbf{I}[X_{:0} : X_{\ell:\ell+k}|X_{0:\ell}] + \sigma_\mu^{\ell+k},\end{aligned}$$

and by the non-negativity of conditional mutual informations [17], $\sigma_\mu^\ell \geqslant \sigma_\mu^{\ell+k}$.

The second property is that σ_μ^ℓ indicates how poorly the present $X_{0:\ell}$ shields the past $X_{:0}$ and future $X_{\ell:}$. When it does, they are conditionally independent, given the present, and σ_μ^ℓ vanishes. Due to this, it can be used to detect a process's *Markov order* R: the smallest R for which $\Pr(X_{0:}|X_{:0}) = \Pr(X_{0:}|X_{-R:0})$.

Proposition 2. $\sigma_\mu^\ell = 0 \Leftrightarrow \ell \geqslant R$.

Proof. By the definition of the Markov order R, a length-R block is a sufficient statistic [29] for $X_{:0}$ about $X_{0:}$:

$$\Pr(X_{0:}|X_{-R:0}) = \Pr(X_{0:}|X_{:0}, X_{-R:0}),$$

and therefore also obeys [17]:

$$\mathbf{I}[X_{:0} : X_{0:}] = \mathbf{I}[X_{-R:0} : X_{0:}].$$

Hence:

$$\begin{aligned}\mathbf{I}[X_{:0} : X_{0:}] &= \mathbf{I}[(X_{:-R}, X_{-R:0}) : X_{0:}]\\ &= \mathbf{I}[X_{-R:0} : X_{0:}] + \mathbf{I}[X_{:-R} : X_{0:}|X_{-R:0}]\end{aligned}$$

implies:

$$\mathbf{I}[X_{:-R} : X_{0:}|X_{-R:0}] = \sigma_\mu^R = 0,$$

due to stationarity.

To calculate σ_μ^ℓ, recall its definition as a conditional mutual information:

$$\begin{aligned}\sigma_\mu^\ell &= \mathbf{I}[X_{:0} : X_{\ell:}|X_{0:\ell}]\\ &= \sum_{\substack{x_{:0} \in X_{:0}\\ x_{0:\ell} \in X_{0:\ell}\\ x_{\ell:} \in X_{\ell:}}} \Pr(x_{:}) \log_2 \frac{\Pr(x_{:0}, x_{\ell:}|x_{0:\ell})}{\Pr(x_{:0}|x_{0:\ell}) \Pr(x_{\ell:}|x_{0:\ell})},\end{aligned}$$

where we have used the notational shorthand for the bi-infinite joint distribution $\Pr(x_{:}) = \Pr(x_{:0}, x_{0:\ell}, x_{\ell:})$.

Note that for an order-R Markov process, if $\ell \geqslant R$ the past and the future are independent over range R [22] and so $\Pr(X_{:0}, X_{\ell:}|X_{0:\ell}) = \Pr(X_{:0}|X_{0:\ell}) \Pr(X_{\ell:}|X_{0:\ell})$. With this, it is clear that σ_μ^ℓ vanishes in such cases. This property has been discussed in prior literature as well [30].

Anticipating the needs of our calculations later, we replace conditional distributions with joint ones, $\Pr(x_{:0}, x_{\ell:}|x_{0:\ell}) = \Pr(x_{:}) / \Pr(x_{0:\ell})$ and $\Pr(x_{:0}|x_{0:\ell}) = \Pr(x_{:0}, x_{0:\ell}) / \Pr(x_{0:\ell})$, obtaining:

$$\sigma_\mu^\ell = \sum_{\substack{x_{:0} \in X_{:0}\\ x_{0:\ell} \in X_{0:\ell}\\ x_{\ell:} \in X_{\ell:}}} \Pr(x_{:}) \log_2 \frac{\Pr(x_{0:\ell}) \Pr(x_{:})}{\Pr(x_{:0}, x_{0:\ell}) \Pr(x_{0:\ell}, x_{\ell:})}. \quad (10)$$

Notably, all the terms needed to compute σ_μ^ℓ are either $\Pr(x_{:0}, x_{0:\ell}, x_{\ell:})$ or marginals thereof. Our next goal, therefore, is to develop the theoretical infrastructure necessary to compute that distribution in closed form.

Similar expressions, which we use later but do not record here, can be developed for the other information measures, h_μ^ℓ, r_μ^ℓ, b_μ^ℓ, and q_μ^ℓ.

III. STRUCTURAL COMPLEXITY

To analytically calculate the elusive information σ_μ^ℓ and related measures we must go beyond the information theory of sequences and introduce computational mechanics, the theory of process structure [19]. The representation it uses for a given process is a form of the hidden Markov model (HMM) [31]: the *ϵ-machine*, which consists of a set $\boldsymbol{\mathcal{S}}$ of causal states, with associated random variable S, and a transition dynamic T. ϵ-Machines satisfy three conditions: irreducibility, unifilarity, and probabilistically distinct states [32]. Irreducibility implies that the associated state-transition graph is strongly connected.

Unifilarity, perhaps the most distinguishing feature, means that for each state $\sigma \in \boldsymbol{\mathcal{S}}$ and each observed symbol x there is at most one outgoing transition from σ labeled $x \in \mathcal{A}$. Critically, unifilarity enables one to directly calculate various process quantities, such as conditional mutual information, using properties of the hidden (causal) states. Notably, many of these quantities cannot be directly calculated using the states of general (nonunifilar) HMMs. Finally, an HMM has probabilistically distinct states when, for every pair of states σ and σ', there exists a word w such that the probability of observing w from each state is distinct: $\Pr(w|\sigma) \neq \Pr(w|\sigma')$. An irreducible, unifilar model with probabilistically distinct states is minimal in the sense that no model with fewer states or transitions generates the process. An HMM satisfying these three properties is an ϵ-machine.

A. Constructing the ϵ-Machine

Given a process, how does one construct its ϵ-machine? First, the set of a process's forward causal states:

$$\boldsymbol{\mathcal{S}}^+ = X_{:}/{\sim}^+_\epsilon, \tag{11}$$

is the partition defined via the causal equivalence relation:

$$x_{:t} \sim^+_\epsilon x'_{:t} \Leftrightarrow \Pr(X_{t:}|X_{:t} = x_{:t}) = \Pr(X_{t:}|X_{:t} = x'_{:t}). \tag{12}$$

That is, each causal state $\sigma^+ \in \boldsymbol{\mathcal{S}}^+$ is an element of the coarsest partition of a process's pasts such that every $x_{:0} \in \sigma^+$ makes the same prediction $\Pr(X_{0:}|\cdot)$. In fact, the causal states are the minimal sufficient statistic of the past to predict the future. We define the reverse causal states:

$$\boldsymbol{\mathcal{S}}^- = X_{:}/{\sim}^-_\epsilon, \tag{13}$$

by similarly partitioning the process's futures:

$$x_{t:} \sim^-_\epsilon x'_{t:} \Leftrightarrow \Pr(X_{:t}|X_{t:} = x_{t:}) = \Pr(X_{:t}|X_{t:} = x'_{t:}). \tag{14}$$

Second, the causal equivalence relation provides a natural unifilar dynamic over the states. For each state σ and next symbol x, either there is a successor state σ' such that the updated past $x_{:t+1} = x_{:t}x \in \sigma'$, for all $x_{:t} \in \sigma$, or $x_{:t+1}$ does not occur. Due to causal-state equivalence, every past within a state collectively either can or cannot be followed by a given symbol. Moreover, since the causal states form a partition of all pasts, there is at most one causal state to which each past can advance.

For a general HMM with states $\rho \in \mathcal{R}$, its symbol-labeled transition matrix $T^{(x)}$ has elements that give the probability of going from state ρ to state ρ' and generating the symbol x:

$$T^{(x)}_{\rho\rho'} \equiv \Pr(X_t = x, \mathcal{R}_{t+1} = \rho' | \mathcal{R}_t = \rho). \tag{15}$$

Furthermore, the internal-state dynamics are governed by the stochastic matrix $T = \sum_x T^{(x)}$. Its unique left eigenvector π, associated with eigenvalue 1, gives the asymptotic state probability $\Pr(\rho)$. By extension, the transition matrix giving the probability of a word $w = x_0 x_1 \ldots x_{\ell-1}$ of length ℓ is the product of transition matrices of each symbol in w:

$$\begin{aligned} T^{(w)} &\equiv \prod_{x_i \in w} T^{(x_i)} \\ &= T^{(x_0)} T^{(x_1)} \cdots T^{(x_{\ell-1})}. \end{aligned} \tag{16}$$

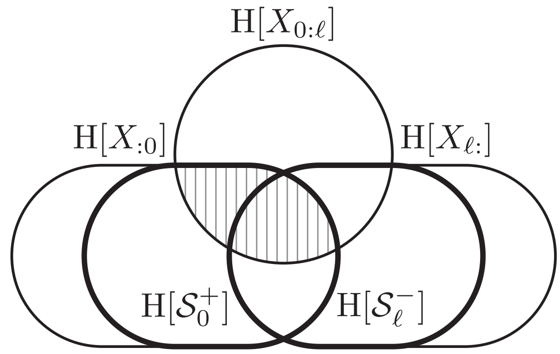

FIG. 3. The mutual information $\mathbf{I}[X_{:0} : X_{0:\ell}]$ between the past and the present (shaded) is equivalent to the mutual information $\mathbf{I}[\mathcal{S}_0^+ : X_{0:\ell}]$ between the forward causal state and the present.

B. Rendering σ_μ^ℓ finitely computable

We can put the forward and reverse causal states to use since they are proxies for a process's semi-infinite pasts and futures, respectively. See, e.g., Fig. 3. In this way, we transform Eq. (9) into a form containing only finite sets of random variables. We calculate directly:

$$\begin{aligned} \sigma_\mu^\ell &= \mathbf{I}[X_{:0} : X_{\ell:}|X_{0:\ell}] \\ &= \mathbf{I}[X_{:0} : (X_{0:\ell}, X_{\ell:})] - \mathbf{I}[X_{:0} : X_{0:\ell}] \\ &= \mathbf{I}[X_{:0} : X_{0:}] - \mathbf{I}[X_{:0} : X_{0:\ell}] \\ &\overset{(a)}{=} \mathbf{I}[\mathcal{S}_0^+ : \mathcal{S}_0^-] - \mathbf{I}[\mathcal{S}_0^+ : X_{0:\ell}] \\ &= \mathbf{I}[\mathcal{S}_0^+ : \mathcal{S}_0^-] - (\mathbf{I}[\mathcal{S}_0^+ : X_{0:\ell} : \mathcal{S}_0^-] + \mathbf{I}[\mathcal{S}_0^+ : X_{0:\ell}|\mathcal{S}_0^-]) \\ &\overset{(b)}{=} \mathbf{I}[\mathcal{S}_0^+ : \mathcal{S}_0^-] - \mathbf{I}[\mathcal{S}_0^+ : X_{0:\ell} : \mathcal{S}_0^-] \\ &= \mathbf{I}[\mathcal{S}_0^+ : \mathcal{S}_0^- | X_{0:\ell}] \\ &= \mathbf{I}[\mathcal{S}_0^+ : \mathcal{S}_0^- : \mathcal{S}_\ell^- | X_{0:\ell}] + \mathbf{I}[\mathcal{S}_0^+ : \mathcal{S}_0^- | X_{0:\ell}, \mathcal{S}_\ell^-] \\ &\overset{(c)}{=} \mathbf{I}[\mathcal{S}_0^+ : \mathcal{S}_0^- : \mathcal{S}_\ell^- | X_{0:\ell}] \\ &= \mathbf{I}[\mathcal{S}_0^+ : \mathcal{S}_\ell^- | X_{0:\ell}] - \mathbf{I}[\mathcal{S}_0^+ : \mathcal{S}_\ell^- | X_{0:\ell}, \mathcal{S}_0^-] \\ &= \mathbf{I}[\mathcal{S}_0^+ : \mathcal{S}_\ell^- | X_{0:\ell}] - (\mathrm{H}[\mathcal{S}_0^+ | X_{0:\ell}, \mathcal{S}_0^-] \\ &\quad - \mathrm{H}[\mathcal{S}_0^+ | \mathcal{S}_0^-, X_{0:\ell}, \mathcal{S}_\ell^-]) \\ &\overset{(d)}{=} \mathbf{I}[\mathcal{S}_0^+ : \mathcal{S}_\ell^- | X_{0:\ell}]. \end{aligned} \tag{17}$$

Above, (a) is true due to Eqs. (11) to (14) and Ref. [33], (b) is true due to Eqs. (13) and (14), (c) is true due to Eq. (14) and unifilarity, and, finally, (d) is true due to both entropy terms' being equal to $\mathrm{H}[\mathcal{S}_0^+|\mathcal{S}_0^-]$ by Eqs. (13) and (14). That is, $\mathcal{S}_0^-$ informationally subsumes both $X_{0:\ell}$ and $\mathcal{S}_\ell^-$ when it comes to $X_{:0}$ and, therefore, also when it comes to $\mathcal{S}_0^+$. All other equalities are basic information identities found in Ref. [21].

In this way, Eq. (17) says that Eq. (10) becomes, in terms of causal states, a new expression for elusivity:

$$\sigma_\mu^\ell = \sum_{\substack{\sigma_0^+ \in \mathcal{S}_0^+ \\ x_{0:\ell} \in X_{0:\ell} \\ \sigma_\ell^- \in \mathcal{S}_\ell^-}} \Pr(\sigma_0^+, x_{0:\ell}, \sigma_\ell^-) \log_2 \frac{\Pr(x_{0:\ell}) \Pr(\sigma_0^+, x_{0:\ell}, \sigma_\ell^-)}{\Pr(\sigma_0^+, x_{0:\ell}) \Pr(x_{0:\ell}, \sigma_\ell^-)}. \tag{18}$$

We transformed the key distribution $\Pr(x_{:0},x_{0:\ell},x_{\ell:})$ over random variables $X_{:0}$ and $X_{\ell:}$ with cardinality of the continuum to $\Pr(\sigma_0^+,x_{0:\ell},\sigma_\ell^-)$ over $\mathcal{S}_0^+$ and $\mathcal{S}_\ell^-$ with typically smaller cardinality. When the causal states are finite or countably infinite, the benefit is substantial. We now turn our attention to computing this joint distribution.

Since the distribution is over both forward and reverse causal states, we must track both simultaneously. The key tool for this is Ref. [34]'s bidirectional machine or *bimachine*. We refer the reader there for details regarding its construction and properties. One feature we need immediately, though, is that bimachine states $\rho_t = (\sigma_t^+,\sigma_t^-)$ are pairs of forward and reverse causal states.

Generally, given an HMM with states $\rho \in \mathcal{R}$, we can construct the distribution of interest if we can find a way to build distributions of the form $\Pr(\rho_i,w,\rho_j)$: the probability of being in state ρ_i, generating the word w, and ending in state ρ_j. The word transition matrix [Eq. (16)] gives exactly this and allows us to build the distribution directly:

$$\Pr(\rho_i,w,\rho_j) = (\pi \circ \mathbf{1}_i)T^{(w)}\mathbf{1}_j^\intercal, \tag{19}$$

where ρ_i and ρ_j are the states of an arbitrary HMM, $a \circ b$ is the Hadamard (elementwise) product of vectors a and b, and $\mathbf{1}_i$ is the row vector with all its elements 0 except for the ith, which is 1.

Applying Eq. (19) to the bimachine, we arrive at the distribution $\Pr((\mathcal{S}_0^+,\mathcal{S}_0^-),\ X_{0:\ell},\ (\mathcal{S}_\ell^+,\mathcal{S}_\ell^-))$, which can be marginalized to $\Pr(\mathcal{S}_0^+,X_{0:\ell},\mathcal{S}_\ell^-)$, the distribution needed to compute Eq. (18). Figure 4 illustrates this distribution for a present of length $\ell = 3$ in the setting of the process's random variable lattice and the forward and reverse causal-state processes.

C. Companion atomic measures

Causal-state expressions for h_μ^ℓ, r_μ^ℓ, b_μ^ℓ, and q_μ^ℓ that we use in the following are:

$$\begin{aligned} h_\mu^\ell &= \mathrm{H}[X_{0:\ell}|\mathcal{S}_0^+], \\ r_\mu^\ell &= \mathrm{H}[X_{0:\ell}|\mathcal{S}_0^+ : \mathcal{S}_\ell^-], \\ b_\mu^\ell &= \mathbf{I}[X_{0:\ell} : \mathcal{S}_0^+|\mathcal{S}_\ell^-], \quad \text{and} \\ q_\mu^\ell &= \mathbf{I}[\mathcal{S}_0^+ : X_{0:\ell} : \mathcal{S}_\ell^-]. \end{aligned}$$

These are derived in ways paralleling that above for σ_μ^ℓ and so we do not give details. They, too, also depend on the joint distribution above in Eq. (19) and its marginals.

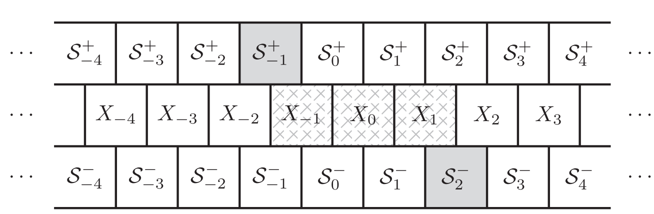

FIG. 4. Random variable lattice illustrating the relationship among forward causal states $\mathcal{S}_t^+$, observed symbols X_t, and reverse causal states $\mathcal{S}_t^-$. Variables in the distribution $\Pr(\mathcal{S}_{-1}^+,X_{-1:2},\mathcal{S}_2^-)$ are highlighted. In particular, the elusivity σ_μ^3 is the mutual information between the two shaded cells ($\mathcal{S}_{-1}^+$ and $\mathcal{S}_2^-$) conditioned on the hatched cells ($X_{-1:2} = X_{-1}X_0X_1$).

IV. EXAMPLES

Let us consider several example processes, to illustrate calculation methods and to examine the behavior of σ_μ^ℓ and companion measures.

A. Golden Mean process

As the first example we analyze the Golden Mean (GM) process, whose ϵ-machines and bimachine state-transition diagrams are given in Fig. 5. The GM process consists of all bi-infinite strings such that no consecutive 1s occur, with probabilities such that either symbol is equally likely following a 0. A stochastic generalization of subshifts of finite type [35], this process can be described by a Markov chain with order $R = 1$. Due to Proposition 2 we expect $\sigma_\mu^1 = 0$. To verify this, we compute each term in Eq. (18) using the edges of

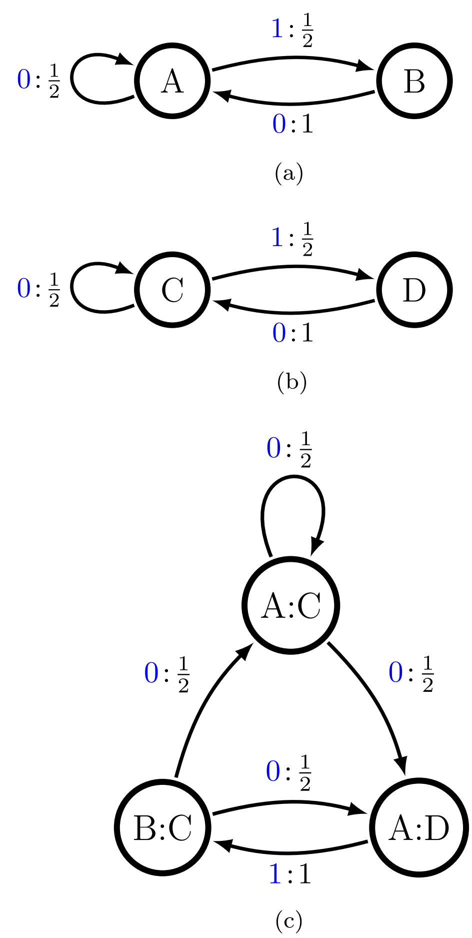

FIG. 5. The several faces of the Golden Mean process. (a) Forward ϵ-machine, (b) reverse ϵ-machine, (c) bimachine. Each representation consists of states (labeled circles) and transitions (arrows) labeled "symbol:probability". The bimachine is effectively the Cartesian product of the forward and reverse machines, though constructed here for forward-time generation, and therefore is generically nonunifilar. For example, state B:C means that the machine is in the superposition of forward state B and reverse state C. Going forward (rightward in Fig. 4) we know that state B must output a 0 and transition to state A. Being in reverse state C we must have transitioned there on a 0 coming from either state C or state D. Thus, we have transitions from B:C to both A:C and A:D on a 0.

the bimachine [Fig. 5(c)] and the invariant state distribution $\pi = (1/3, 1/3, 1/3)$:

$$\frac{\Pr(0)\Pr(A,0,C)}{\Pr(A,0)\Pr(0,C)} = \frac{2/3 \cdot 1/6}{1/3 \cdot 1/3} = 1,$$
$$\frac{\Pr(0)\Pr(A,0,D)}{\Pr(A,0)\Pr(0,D)} = \frac{2/3 \cdot 1/6}{1/3 \cdot 1/3} = 1,$$
$$\frac{\Pr(0)\Pr(B,0,C)}{\Pr(B,0)\Pr(0,C)} = \frac{2/3 \cdot 1/6}{1/3 \cdot 1/3} = 1,$$
$$\frac{\Pr(0)\Pr(B,0,D)}{\Pr(B,0)\Pr(0,D)} = \frac{2/3 \cdot 1/6}{1/3 \cdot 1/3} = 1, \quad \text{and}$$
$$\frac{\Pr(1)\Pr(A,1,C)}{\Pr(A,1)\Pr(1,C)} = \frac{1/3 \cdot 1/3}{1/3 \cdot 1/3} = 1.$$

We see that the argument of each $\log_2$ in Eq. (18) is 1, confirming that $\sigma_\mu^1 = 0$.

B. Information measures versus present length

We now investigate the behavior of σ_μ^ℓ and its companions q_μ^ℓ, b_μ^ℓ, and r_μ^ℓ for several example processes: the aforementioned GM, the Even, the Noisy Period 3 (NP3), and the Noisy Random Phase-Slip (NRPS) processes. The ϵ-machines for the latter are shown in Fig. 6. Each exhibits different convergence behaviors with ℓ for the differing measures; see the graphs in Fig. 7. We now characterize each of them.

We first consider σ_μ^ℓ, shown in the upper-left panel in Fig. 7. While for each process σ_μ^ℓ vanishes with increasing ℓ, the convergence behaviors differ. The GM process is identically 0 at all lengths due to its order 1 Markov nature, just noted. The NRPS process, with a Markov order of $R = 5$, has nonzero σ_μ^ℓ for $\ell < 5$ and zero $\sigma_\mu^\ell = 0$ beyond. Finally, both the Even and the Nemo processes are infinite-order Markov and so their σ_μ^ℓ never exactly vanishes, though they converge exponentially rapidly. (See Figs. 8 and 9.) The next section (Sec. IV C) analyzes exponential convergence in more detail.

Next, consider q_μ^ℓ and b_μ^ℓ, which are closely associated, as it turns out. To see why, first examine the large-ℓ limit:

$$\lim_{\ell\to\infty} \mathrm{I}[X_{0:\ell} : X_{\ell:}] = \mathbf{E}$$
$$= q_\mu^\infty + b_\mu^\infty .$$

Furthermore, we can characterize q_μ^∞ as remembering the "phase" of a process:

$$q_\mu^\infty = \log_2 \mathrm{per}(\epsilon\text{-machine}),$$

where per(ϵ-machine) is the *graph period* of the ϵ-machine; the greatest common divisor of all the graph's cycles. The GM, Even, and NRPS processes each contain state self-loops and so their period is 1. For these processes, $b_\mu^\infty = \mathbf{E}$. All cycles in the NP3 process are of length 3, and so its period is 3. And, in this instance, $\mathbf{E} = \log_2 3$ and so $b_\mu^\infty = 0$.

We see these behaviors play out in the upper-right and lower-left panels of Fig. 7. At the upper right, q_μ^ℓ converges to 0 for the GM, Even, and NRPS processes, while the NP3 process limits to the logarithm of its period, $\log_2 3 = \mathbf{E}$. At the lower left, the b_μ^ℓ curves for the GM, Even, and NRPS processes limit to their respective excess entropies, while NP3 has $b_\mu^\ell = 0$ for all ℓ.

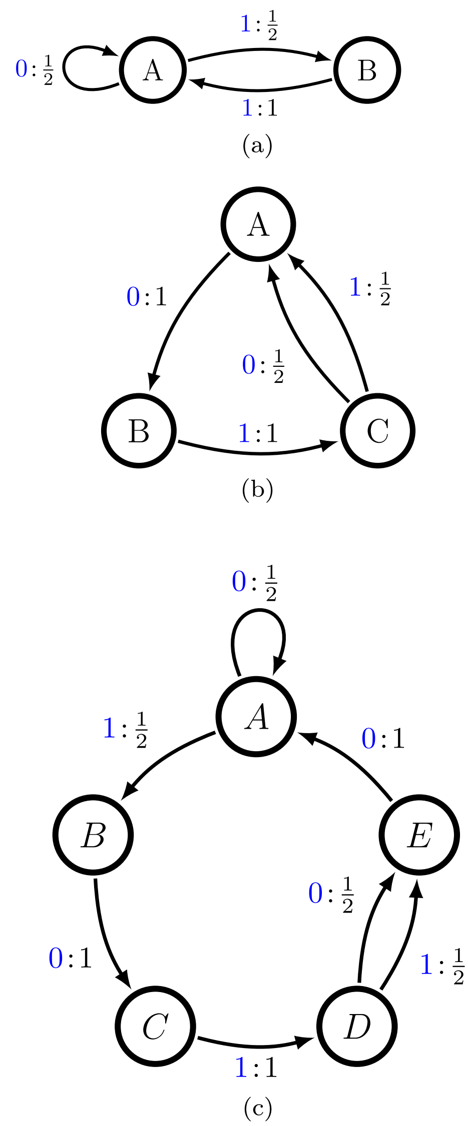

FIG. 6. ϵ-Machines for the example processes: (a) The Even process, (b) the Noisy Period Three (NP3) process, (c) the Noisy Random Phase-slip (NRPS) process.

Finally, the ephemeral information r_μ^ℓ, plotted in the lower-right panel in Fig. 7, also depends on the ϵ-machine's period. For $\ell \gg 1$:

$$\mathrm{H}(\ell) = \mathbf{E} + \ell h_\mu$$
$$= q_\mu^\ell + 2b_\mu^\ell + r_\mu^\ell$$
$$\implies r_\mu^\ell = \ell h_\mu - b_\mu^\ell,$$

where the subextensive part $-b_\mu^\ell$ ranges from 0 (processes whose ϵ-machine is a "noisy" cycle (like the NP3 process) to $-\mathbf{E}$ (processes with a graph period of 1).

C. Exponential convergence of σ_μ^ℓ

One way to classify processes is according to whether or not an observer can determine the causal state a process is in from finite or infinite sequence measurements. If so, then the process is synchronizable. All of the previous examples are

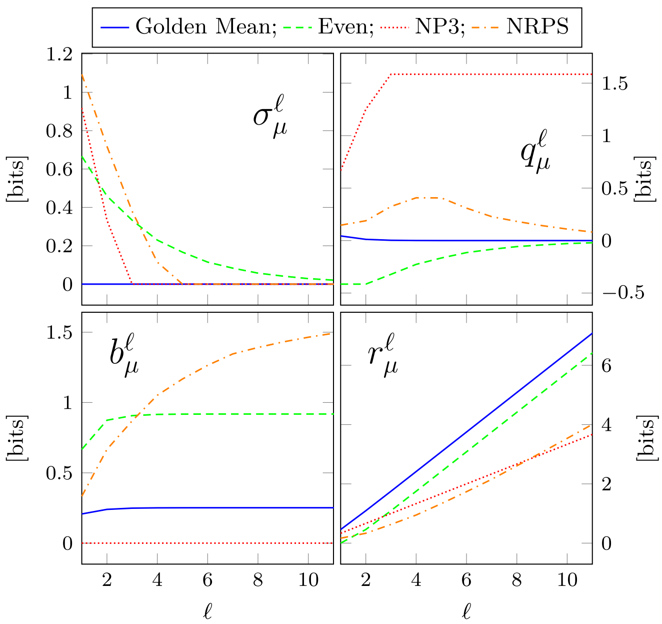

FIG. 7. Information measures as a function of the present's length ℓ. Since the examples are stationary, finite processes, both σ_μ^ℓ and q_μ^ℓ converge to 0 with increasing ℓ and b_μ^ℓ converges to a constant value of $\mathbf{E}$. And so, the growth of $\mathrm{H}(\ell)$ is entirely captured in r_μ^ℓ, and it grows linearly with ℓ.

synchronizable. References [36] and [37] proved that for any synchronizable process described by a finite-state HMM, there exist constants $K > 0$ and $0 < \alpha < 1$ such that:

$$h_\mu(\ell) - h_\mu \leqslant K\alpha^\ell \quad \text{for all} \quad \ell \in \mathbb{N}, \tag{20}$$

where:

$$h_\mu(\ell) = \mathrm{H}(\ell) - \mathrm{H}(\ell - 1).$$

Note that $h_\mu(1) - h_\mu = \rho_\mu^1$. One well-known identity [18] is that the sum of the $h_\mu(\ell)$ terms is the excess entropy:

$$\mathbf{E} = \sum_{k=1}^{\infty}(h_\mu(k) - h_\mu) \tag{21}$$

$$= \rho_\mu^1 + \sum_{k=2}^{\infty}(h_\mu(k) - h_\mu). \tag{22}$$

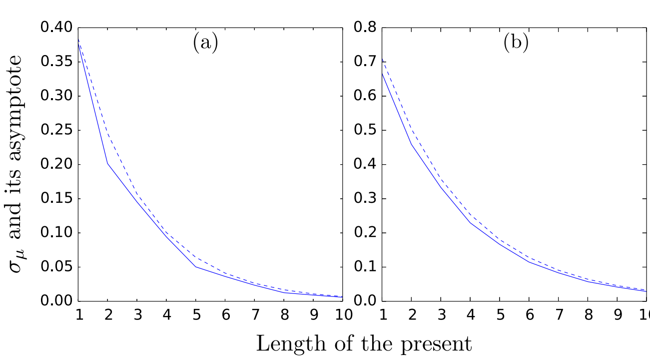

FIG. 8. σ_μ^ℓ (solid line) and its asymptote (dashed line) for (a) the Nemo process with $K = 0.6$ and $\alpha = 0.64$ and (b) the even process with $K = 1$ and $\alpha = 0.71$.

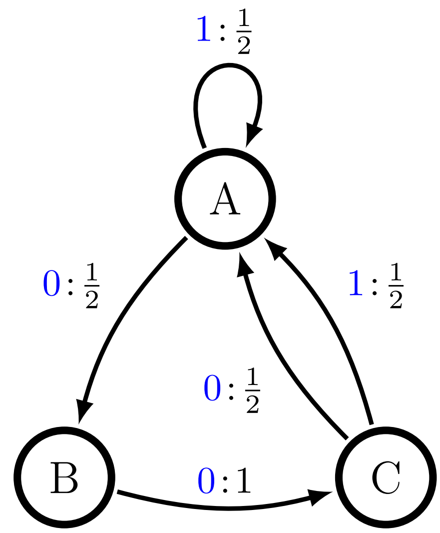

FIG. 9. The Nemo process.

This provides a new identity [16]:

$$\sigma_\mu^1 = \sum_{k=2}^{\infty}(h_\mu(k) - h_\mu), \tag{23}$$

which can be generalized to:

$$\sigma_\mu^\ell = \sum_{k=\ell+1}^{\infty}(h_\mu(k) - h_\mu). \tag{24}$$

Applying the bound from Eq. (20) to each term, we find:

$$\sum_{k=\ell+1}^{\infty}(h_\mu(k) - h_\mu) \leqslant \sum_{k=\ell+1}^{\infty} K\alpha^k.$$

The right-hand side, being a convergent geometric series, yields:

$$\sigma_\mu^\ell \leqslant \frac{K\alpha^{\ell+1}}{1-\alpha}$$

or simply:

$$\sigma_\mu^\ell \leqslant K'\alpha^\ell. \tag{25}$$

We now drop the prime, simplifying the notation. Thus, the elusive information vanishes exponentially rapidly for synchronizable processes.

Figure 8 compares σ_μ^ℓ with its best-fit exponential bound for two processes: the Nemo process, shown in Fig. 9, and the Even process. For each, the solid line represents σ_μ^ℓ and the dashed line is the fit. Estimated values for the Nemo process are $K = 0.6$ and $\alpha = 0.64$. The fit parameters for the Even process are $K = 1.0$ and $\alpha = 0.71$. They were estimated in accordance with the conditions stated for Eq. (20). The fits validate the convergence in Eq. (25).

V. MEASURES OF EMERGENCE?

The elusive information σ_μ^ℓ is conditioned on a present of length ℓ. What if we do not condition it, simply ignoring the present? It becomes the *persistent mutual information* [15,38]:

$$\mathrm{PMI}(\ell) = \mathbf{I}[X_{:0} : X_{\ell:}]. \tag{26}$$

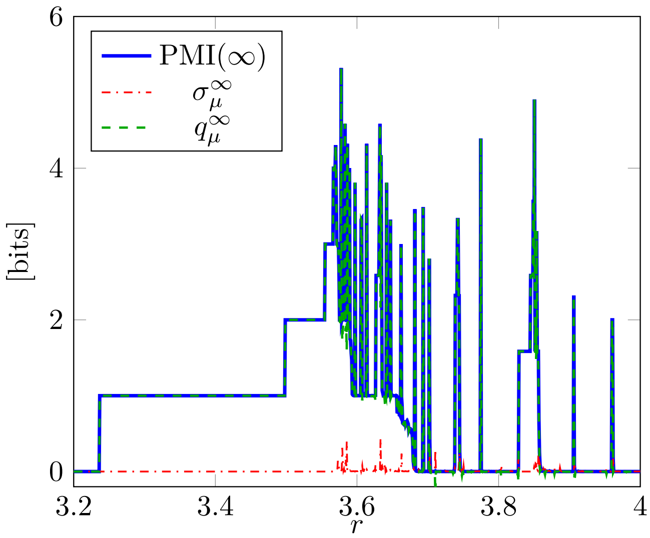

FIG. 10. Persistent mutual information PMI(∞), elusivity σ_μ^∞, and multivariate mutual information q_μ^∞ of the Logistic Map symbolic dynamics as a function of the map control parameter r. Recall that the symbolic dynamics does not see period doubling until r is above the appearance of the associated superstable periodic orbit. This discrepancy in the appearance of periodicity as a function of r does not occur when the map is chaotic. (Compare Fig. 1 in Ref. [15].)

Notably, PMI(∞) was offered up as a measure of "emergence" in general complex systems. Our preceding analysis, though, gives a more nuanced view of this interpretation, especially when emergence is considered in light of structural criteria introduced previously [39,40]. Our framework reveals that PMI(ℓ) is not an atomic measure; rather it consists of two now-familiar components:

$$\mathrm{PMI}(\ell) = q_\mu^\ell + \sigma_\mu^\ell. \tag{27}$$

Which component is most important? Are both? Which is associated with emergence? Are both?

Section IV C showed that synchronizable processes have $\sigma_\mu^\ell \to 0$. So, for this broad class at least, PMI(∞) $= q_\mu^\infty$. Based on extensive process surveys that we do not report on here, we conjecture that $\sigma_\mu^\infty = 0$ holds even more generally. And so, it appears that PMI(∞) generally is dominated by the multivariate mutual information q_μ^∞. Moreover, recalling the analysis of q_μ^ℓ for the Noisy Period 3 process shown in the upper-right panel in Fig. 7, it appears that PMI(∞) is only sensitive to state-phase preservation, giving $\log_2 p$ where p is the period of the ϵ-machine graph.

As a test of our conjecture that the elusive information vanishes and that PMI(∞) is dominated by q_μ^∞, we applied our information-measure estimation methods to the symbolic dynamics generated by the Logistic Map of the unit interval as a function of its control parameter r. Figure 10 plots the results. Indeed, the elusive information does vanish. Thus, we conclude that PMI(∞) is a property of q_μ^∞.

In addition, our simulation results reproduced those in the PMI(∞) analysis of the Logistic Map in Ref. [15], though their estimation method for PMI(∞) differs markedly. Here, we calculate via the Logistic Map symbolic dynamics; there, joint distributions over the continuous unit-interval domain were used. Both investigations lead to the conclusion that PMI(∞) is equal to (the logarithm of) the number of chaotic "bands" cyclically permuted or the period of the periodic orbit at a given parameter value. In short, PMI(∞) is a measure of nonmixing dynamics.

The status of the excess entropy as a measure of emergence has been criticized in the past due to its sensitivity to periodic processes. Here, we showed in Sec. IV B that this sensitivity to periodic behaviors can be naturally factored out. However, it is exactly this "undesirable" periodic component that PMI(∞) captures. Furthermore, if $\mathrm{I}[(X_{:0}, X_{0:\ell}) : X_{\ell:}]$ is viewed through the lens of the *partial information decomposition* [41]—decomposing the mutual information into nonnegative redundant, unique, and synergistic components—we get further insight, if somewhat dire for the PMI. In that framework, $X_{:0}$ and $X_{0:\ell}$ are considered "inputs" to a function with $X_{\ell:}$ as the "output". Here, σ_μ^ℓ is equal to the information uniquely transmitted from $X_{:0}$ to $X_{\ell:}$ plus what is synergistically provided by both $X_{:0}$ and $X_{0:\ell}$. Similarly, b_μ^ℓ is the information uniquely provided by $X_{0:\ell}$ plus what is synergistically provided by both $X_{:0}$ and $X_{0:\ell}$. Finally, q_μ^ℓ is equal to what is redundantly provided by both $X_{:0}$ and $X_{0:\ell}$, minus what those two synergistically provide. Since $\sigma_\mu^\infty = 0$, the synergistic effects must vanish. This means that q_μ^∞ (and so PMI(∞)) is exactly the information *redundantly* provided by both $X_{:0}$ and $X_{0:\ell}$. While emergence is in many ways still under active debate, it is clear that it is not redundancy.

Given the restricted form of structure (periodicity) to which it is sensitive, PMI(∞) cannot be taken as a general measure for detecting the emergence of organization in complex systems. No matter; though a quarter of a century old, the statistical complexity [42]—a direct measure of the structural organization and stored information—continues to fill the role of detecting emergent organization quite well. Moreover, computational mechanics' ϵ-machines directly show what the emergent organization is.

VI. CONCLUSION

We first defined the elusive information and developed a closed-form analytic expression to calculate it from a process's hidden Markov model. The sequel [16] shows how to use spectral methods [43] to develop alternative closed-form expressions for the elusive information and its companions, giving exact expressions and a direct understanding of the origin of their convergence behaviors.

Investigating how the present shields the past and future is essentially a study of what Markov order means for structured processes. It gives much insight into the endeavor of model building and even into general concerns about the emergence of organization in complex systems. First, this study of Markovian complexity gives a common ground on which to contrast structural inference and emergence, showing that we should not conflate these two distinct questions. Second, and perhaps most constructively, though, it sheds light on the challenges of inference for complex systems. In particular, when $\sigma_\mu^\ell > 0$ sequence statistics are inadequate for modeling and so we must employ state-based models to properly, finitely represent a process's internal organization.

The results show rather directly how present observables typically do not contain all of the information that correlates the past and the future. One consequence is that instantaneous measurements are not enough. This means, exactly, that Markov chain models of complex physical systems are fundamentally inadequate, though eminently helpful and simplifying when they are appropriate representations. When predicting the behavior and structure of complex systems, the larger consequence is that we must build state-based models and not use mere lookup tables or sequence histograms. That said, states are little more than a conditional coarse-graining of sequences into subsets that are more compactly predictive than sequences alone. And this implies, in turn, that monitoring *only* prediction performance is inadequate. We must also monitor model complexity, not as an antidote to overfitting, but as a fundamental goal for both predicting and understanding hidden mechanisms. This, we believe, most fully respects Markov's contribution to the sciences.

ACKNOWLEDGMENTS

This material is based upon work supported by, or in part by, the U.S. Army Research Laboratory and the U.S. Army Research Office under Contracts No. W911NF-13-1-0390 and No. W911NF-13-1-0340.

[1] J. Bernoulli, *Ars Conjectandi, Opus Posthumum, Accedit Tractatus de Seriebus infinitis, et Epistola Gallice Scripta de Ludo Pilae recticularis*, Basileae, 1713, Chapters 1–4 Translated into English by B. Sung, *Ars Conjectandi*, Technical Report No. 2 (Department of Statistics, Harvard University, Cambridge, MA, 1966).

[2] S.-D. Poisson, *Recherches sur la Probabilité des Jugements en Matiere Criminelle et en Metière Civile, Préédées des Regles Généales du Calcul des Probabilitiés. Imprimeur–Libraire pour les Mathematiques* (Bachelier, Paris, 1837).

[3] P. L. Chebyshev, Des valuers moyennes, J. Math. Pure Appl. **12**, 177 (1867).

[4] A. A. Markov, Issledovanie zamechatel'nogo sluchaya zavisimyh ispytanij, Izvest. Akad. Nauk **93**, 61 (1907); Translated into French: Recherches sur un cas remarquable d'epreuves dependantes, Acta Math. Stockholm **33**, 87 (1910).

[5] A. A. Markov, Rasprostranenie predel'nyh teorem ischisleniya veroyatnostej na summu velichin svyazannyh v cep', *Zapiski Akademii Nauk po Fiziko-matematicheskomu otdeleniyu*, *3* (1908). Translated into German: Ausdehnung der Satze uber die Grenzwerte in der Wahrscheinlichkeitsrechnung auf eine Summe verketteter Grossen, in *Wahrscheinlichkeitsrechnung*, edited by A. A. Markoff, Translated by H. Liebmann (B. G. Teuber, Leipzig, 1912), pp. 272–298. Translated into English: Extension of the limit theorems of probability theory to a sum of variables connected in a chain, in *Dynamic Probabilities Systems*, edited by R. A. Howard, translated by S. Petelin (Wiley, New York, 1971), Vol. 1, pp. 552–576.

[6] A. A. Markov, Primer statisticheskogo issledovaniya nad tekstom "Evgeniya Onegina," illyustriruyuschij svyaz' ispytanij v cep', Izv. Akad. Nauk, SPb **93**, 153 (1913).

[7] A. A. Markov, Ischislenie veroyatnostej (SPB, Devon, UK, 1900), 2-e izd. (SPB, Devon, UK, 1913). Translated into German: *Wahrscheinlichkeits-Rechnung* (Teubner, Leipzig, 1912), 3-e izd. (SPB, Devon, UK, 1913), 4-e izd (Moskva, Moscow, 1924).

[8] O. Penrose, *Foundations of Statistical Mechanics; A Deductive Treatment* (Pergamon Press, Oxford, UK, 1970).

[9] K. A. Dill and S. Bromberg, *Molecular Driving Forces: Statistical Thermodynamics in Chemistry and Biology* (Garland Science, New York, 2011).

[10] D. L. Hartl and A. G. Clark, *Principles of Population Genetics*, 4th ed. (Sinauer Associates, New York, 2006).

[11] L. Calvert and A. Fisher, How to forecast long-run volatility: Regime-switching and the estimation of multifractal processes, J. Finan. Econmetr. **2**, 49 (2004).

[12] J. S. Liu, *Monte Carlo Strategies in Scientific Computing, Springer Series in Statistics* (Springer, New York, 2008).

[13] L. Page, S. Brin, R. Motwani, and T. Winograd, The PageRank citation ranking: Bringing order to the web, Technical report (Stanford Information Laboratory, Palo Alto, CA, 1999).

[14] D. Blackwell, The entropy of functions of finite-state Markov chains, in *First Prague Conference on Information Theory, Liblice, 1956* (Czechoslovak Academy of Sciences, Prague, 1957), Vol. 28, pp. 13–20.

[15] R. Ball, M. Diakonova, and R. S. Mackay, Quantifying emergence in terms of persistent mutual information, Adv. Complex Syst. **13**, 327 (2010).

[16] P. M. Ara, P. M. Riechers, and J. P. Crutchfield, The convergence to Markov order in hidden processes (unpublished) (2015).

[17] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley-Interscience, New York, 1991).

[18] J. P. Crutchfield and D. P. Feldman, Regularities unseen, randomness observed: Levels of entropy convergence, Chaos **13**, 25 (2003).

[19] J. P. Crutchfield, Between order and chaos, Nat. Phys. **8**, 17 (2012).

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

[21] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. (Wiley-Interscience, New York, 2006).

[22] R. G. James, J. R. Mahoney, C. J. Ellison, and J. P. Crutchfield, Many roads to synchrony: Natural time scales and their algorithms, Phys. Rev. E **89**, 042135 (2014).

[23] J. P. Crutchfield, C. J. Ellison, J. R. Mahoney, and R. G. James, Synchronization and control in intrinsic and designed computation: An information-theoretic analysis of competing models of stochastic computation, Chaos **20**, 037105 (2010).

[24] P. Gacs and J. Korner, Common information is much less than mutual information, Probl. Contrib. Inf. Theor. **2**, 149 (1973).

[25] A. J. Bell, The co-information lattice, in *Proceedings of the Fifth International Workshop on Independent Component Analysis and Blind Signal Separation, ICA 2003*, edited by S. Makino, S. Amari, A. Cichocki, and N. Murata (Springer, New York, 2003).

[26] R. W. Yeung, *Information Theory and Network Coding* (Springer, New York, 2008).

[27] S. A. Abdallah and M. D. Plumbley, A measure of statistical complexity based on predictive information with application to finite spin systems, Phys. Lett. A **376**, 275 (2012).

[28] S. Verdú and T. Weissman, The information lost in erasures, IEEE Trans. Info. Theor. **54**, 5030 (2008).

[29] G. Casella and R. L. Berger, *Statistical Inference, Vol. 2* (Duxbury, Pacific Grove, CA, 2002).

[30] P. Gmeiner, Properties of Persistent Mutual Information and Emergence, arXiv:1210.5058 [math-ph].

[31] L. R. Rabiner, A tutorial on hidden Markov models and selected applications, IEEE Proc. **77**, 257 (1989).

[32] N. F. Travers and J. P. Crutchfield, Equivalence of history and generator ϵ-machines, arXiv:1111.4500 [math.PR] (2011).

[33] J. P. Crutchfield and C. J. Ellison, The Past and the Future in the Present, arXiv:1012.0356 [nlin.CD] (2010).

[34] 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**, 1005 (2009).

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

[36] N. Travers and J. P. Crutchfield, Exact synchronization for finite-state sources, J. Stat. Phys. **145**, 1181 (2011).

[37] N. Travers and J. P. Crutchfield, Asymptotic synchronization for finite-state sources, J. Stat. Phys. **145**, 1202 (2011).

[38] M. Diakonova and R. S. Mackay, Mathematical examples of space-time phases, Int. J. Bifurcat. Chaos **21**, 2297 (2011).

[39] J. P. Crutchfield, Is anything ever new? Considering emergence, in *Complexity: Metaphors, Models, and Reality*. *Vol. XIX*. *Santa Fe Institute Studies in the Sciences of Complexity*, edited by G. Cowan, D. Pines, and D. Melzner (Addison-Wesley, Redwood, CA, 1994), pp. 479–497 [Reprinted in *Emergence: Contemporary Readings in Philosophy and Science*, edited by M. A. Bedau and P. Humphreys (Bradford Book, MIT Press, Cambridge, MA, 2008), pp. 269–286.

[40] J. P. Crutchfield, The calculi of emergence: Computation, dynamics, and induction, Physica D **75**, 11 (1994).

[41] P. L. Williams and R. D. Beer, Nonnegative decomposition of multivariate information, arXiv:1004.2515 (2010).

[42] J. P. Crutchfield and K. Young, Inferring Statistical Complexity, Phys. Rev. Lett. **63**, 105 (1989).

[43] J. P. Crutchfield, P. Riechers, and C. J. Ellison, Exact complexity: Spectral decomposition of intrinsic computation, Phys. Lett. A **380**, 998 (2015).