Santa Fe Institute Working Paper 14-12-046
arxiv.org:1412.2859 [cond-mat.stat-mech]

Predictive Rate-Distortion for Infinite-Order Markov Processes

Sarah E. Marzen[1,*] and James P. Crutchfield[2,†]

[1]Department of Physics,
Redwood Center for Theoretical Neuroscience
University of California at Berkeley, Berkeley, CA 94720-5800
[2]Complexity Sciences Center and Department of Physics
University of California at Davis, One Shields Avenue, Davis, CA 95616-5720

(Dated: March 19, 2016)

Predictive rate-distortion analysis suffers from the curse of dimensionality: clustering arbitrarily long pasts to retain information about arbitrarily long futures requires resources that typically grow exponentially with length. The challenge is compounded for infinite-order Markov processes, since conditioning on finite sequences cannot capture all of their past dependencies. Spectral arguments confirm a popular intuition: algorithms that cluster finite-length sequences fail dramatically when the underlying process has long-range temporal correlations and can fail even for processes generated by finite-memory hidden Markov models. We circumvent the curse of dimensionality in rate-distortion analysis of finite- and infinite-order processes by casting predictive rate-distortion objective functions in terms of the forward- and reverse-time causal states of computational mechanics. Examples demonstrate that the resulting algorithms yield substantial improvements.

Keywords: optimal causal filtering, computational mechanics, epsilon-machine, causal states, predictive rate-distortion, information bottleneck

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

I. INTRODUCTION

Biological organisms and engineered devices are often required to predict the future of their environment either for survival or performance. Absent side information about the environment that is inherited or hardwired, their only guide to the future is the past. One strategy for adapting to environmental challenges, then, is to memorize as much of the past as possible—a strategy that ultimately fails, even for simple stochastic environments, due to the exponential growth in required resources—the curse of dimensionality.

One way to circumvent resource limitations is to identify minimal sufficient statistics of prediction, or the forward-time causal states $\mathcal{S}^+$. Storing these states costs on average $C_\mu^+ = \mathrm{H}[\mathcal{S}^+]$ bits of Shannon information, a quantity more popularly known as the statistical complexity [1–3]. However, for most processes [4, 5], statistical complexity is infinite and so storing the causal states themselves exceeds the capacity of any learning strategy.

As such, one asks for approximate, lossy features that predict the future as well as possible given resource constraints. Shannon introduced rate-distortion theory to analyze such trade-offs [6, 7]. When applied to prediction, rate-distortion theory provides a principled framework for calculating the function delineating achievable from unachievable predictive distortion for a given amount of memory. In practice, one typically compresses finite-length pasts to retain information about finite-length futures [8, 9]. This can yield reasonable estimates of predictive rate-distortion functions at sufficient lengths, but how long is long enough?

We introduce a new theory and algorithm for calculating predictive rate-distortion functions and lossy predictive features when given a model of a process. The heart of this is a new theorem that identifies lossy predictive features as lossy causal states, an extension of a previous result identifying lossless predictive features as causal states [8, 9]. The theorem allows us to calculate lossy predictive features and predictive rate-distortion functions directly from bidirectional models, without ever having to calculate trajectory probabilities—effectively leveraging the mechanistic information supplied by the model to obtain only the needed information about the process' statistics. This ameliorates, and sometimes eliminates, the aforementioned curse of dimensionality.

Most research in this area is primarily focused on new techniques for building predictive models from data, suggesting the question: why build an optimal approximate predictive model when a maximally predictive model is known? We envision at least two applications. Accurate calculation of lossy predictive features has already found utility in testing the predictive capabilities of biological sensory systems [10]. As such, the results presented here expand the range of stimuli for which an organism's predictive capabilities can be tested. And, more broadly,

* smarzen@berkeley.edu
† chaos@ucdavis.edu

this or similar work might aid computation of optimally coarse-grained dynamical models, which can be useful when one wants to interpret the results of large-scale simulations.

The usefulness of the algorithms presented here naturally depend on the quality of the model with which one starts. The examples shown here suggest that when one's model is accurate, and when the underlying process has relatively long-range temporal correlations, the new predictive rate-distortion algorithm given here can substantially outperform existing algorithms.

Section II reviews minimal maximally predictive models and predictive rate-distortion theory. Section III describes fundamental limitations to current predictive rate-distortion algorithms. Section IV introduces a new theorem that reformulates predictive rate-distortion objectives in terms of minimal sufficient statistics of prediction and retrodiction. Section V then describes a new class of algorithms for computing lossy causal states based on this theorem, given a model of a process, and illustrates its performance on several simple infinite-order Markov processes. Section VI summarizes outstanding issues, desirable extensions, and future applications.

II. BACKGROUND

When an information source's entropy rate falls below a channel's capacity, Shannon's Second Coding Theorem says that there exists an encoding of the source messages such that the information can be transmitted error-free, even over a noisy channel.

What happens, though, when the source rate is above this error-free regime? This is what Shannon solved by introducing rate-distortion theory [6, 7]. Our view is that, for natural systems, the above-capacity regime is disproportionately more common and important than the original error-free coding with which Shannon and followers started. This viewpoint may be particularly important for understanding biological sensory systems; e.g., as studied in Refs. [10–12]. Summarizing sensory information not only helps reduce demands on memory, but also the computational complexity of downstream perceptual processing, cognition, and acting. For instance, much effort has focused on determining memory and the ability to reproduce a given time series [13], but that memory may only be important to the extent that it affects the ability to predict the future; e.g., see Refs. [4, 10, 14, 15].

We are interested, therefore, as others have been, in identifying lossy predictive features.

First, we review the calculus of minimal maximally predictive models. These, finally, lead us to describe what we mean by lossy causal states. The following assumes familiarity with information theory at the level of Ref. [16].

A. Processes and Their Causal States

When predicting a system the main object is the *process* $\mathcal{P}$ it generates: the list 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 $\Pr(\ldots X_{-2}, X_{-1}, X_0, X_1, \ldots)$. We denote a contiguous chain of random variables as $X_{0:\ell} = X_0 X_1 \cdots X_{\ell-1}$. Left indices are inclusive; right, exclusive. We suppress indices that are infinite. In this setting, the *present* $X_{t:t+\ell}$ is the length-ℓ chain beginning at t, the *past* is the chain $X_{:t} = \ldots X_{t-2} X_{t-1}$ leading up the present, and the *future* is the chain following the present $X_{t+\ell:} = X_{t+\ell+1} X_{t+\ell+2} \cdots$. When being more expository, we use arrow notation; for example, for the past $\overleftarrow{X} = X_{:0}$ and future $\overrightarrow{X} = X_{0:}$. We refer on occasion to the space $\overleftarrow{\mathbf{X}}$ of all pasts. Finally, we assume a process is ergodic and stationary—$\Pr(X_{0:\ell}) = \Pr(X_{t:\ell+t})$ for all $t \in \mathbb{Z}$—and the measurement symbols x_t range over a finite alphabet: $x \in \mathcal{A}$. We make no assumption that the symbols represent the system's states—they are at best an indirect reflection of an internal Markov mechanism. That is, the process a system generates is a *hidden Markov process* [17].

Forward-time causal states $\mathcal{S}^+$ are minimal sufficient statistics for predicting a process's future [1, 2]. This follows from their definition as sets of pasts grouped by the equivalence relation $\sim^+$:

$$
\begin{aligned}
&x_{:0} \sim^+ x'_{:0} \\
&\qquad \Leftrightarrow \Pr(X_{0:}|X_{:0} = x_{:0}) = \Pr(X_{0:}|X_{:0} = x'_{:0}) ~. \quad (1)
\end{aligned}
$$

As a shorthand, we denote a cluster of pasts so defined, a *causal state*, as $\sigma^+ \in \mathcal{S}^+$. We implement Eq. (1) via the *causal state map*: $\sigma^+ = \epsilon^+(\overleftarrow{x})$. Through it, each state σ^+ inherits a probability $\pi(\sigma^+)$ from the process's probability over pasts $\Pr(X_{:0})$. The forward-time *statistical complexity* is defined as the Shannon entropy of the probability distribution over forward-time causal states [1]:

$$
C_\mu^+ = \mathrm{H}[\mathcal{S}^+] ~. \quad (2)
$$

A generative model—the process's ϵ-*machine*—is built out of the causal states by endowing the state set with a transition dynamic:

$$
T^x_{\sigma\sigma'} = \Pr(\mathcal{S}^+_{t+1} = \sigma', X_t = x | \mathcal{S}^+_t = \sigma) ~,
$$

matrices that give the probability of generating the next symbol x_t and ending in the next state σ_{t+1}, if starting in state σ_t. (Since output symbols are generated during transitions there is, in effect, a half time-step difference in index. We suppress notating this.) For a discrete-time, discrete-alphabet process, the ϵ-machine is its minimal unifilar Hidden Markov Model (HMM) [1, 2]. (For general background on HMMs see Refs. [18–20]. For a mathematical development of ϵ-machines see Ref. [21].) Note that the causal-state set of a process generated by even a finite HMM can be finite, countable, or uncountable. *Minimality* can be defined by either the smallest number of causal states or the smallest statistical complexity C_μ [2]. *Unifilarity* is a constraint on the transition matrices such that the next state σ_{t+1} is determined by knowing the current state σ_t and the next symbol x_t.

A similar equivalence relation $\sim^-$ can be applied to find minimal sufficient statistics for retrodiction [22]. Futures are grouped together if they have equivalent conditional probability distributions over pasts:

$$\begin{aligned} & x_{0:} \sim^- x'_{0:} \\ & \quad \Leftrightarrow \Pr(X_{:0}|X_{0:} = x_{0:}) = \Pr(X_{:0}|X_{0:} = x'_{0:}) \ . \end{aligned} \tag{3}$$

A cluster of futures—a *reverse-time causal state*—defined by $\sim^-$ is denoted $\sigma^- \in \mathcal{S}^-$. Again, each σ^- inherits a probability $\pi(\sigma^-)$ from the probability over futures $\Pr(X_{0:})$. And, the *reverse-time statistical complexity* is the Shannon entropy of the probability distribution over reverse-time causal states:

$$C_\mu^- = \mathrm{H}[\mathcal{S}^-] \ . \tag{4}$$

In general, the forward- and reverse-time statistical complexities are not equal [22, 23]. That is, different amounts of information must be stored from the past (future) to predict (retrodict). Their difference $\Xi = C_\mu^+ - C_\mu^-$ is a process's *causal irreversibility* and it reflects this statistical asymmetry.

The amount of information in the future that is *predictable* from the past is the past-future mutual information or *excess entropy*:

$$\mathbf{E} = \mathrm{I}[\overleftarrow{X}, \overrightarrow{X}] \ .$$

The forward- and reverse-time causal states play a key role in prediction. First, one must track the causal states in order to predict the $\mathbf{E}$ bits of future information that are predictable. Second, they *shield* the past and future from one another. That is:

$$\begin{aligned} & \Pr(\overleftarrow{X}, \overrightarrow{X}|\mathcal{S}^+) = \Pr(\overleftarrow{X}|\mathcal{S}^+) \Pr(\overrightarrow{X}|\mathcal{S}^+) \text{ and} \\ & \Pr(\overleftarrow{X}, \overrightarrow{X}|\mathcal{S}^-) = \Pr(\overleftarrow{X}|\mathcal{S}^-) \Pr(\overrightarrow{X}|\mathcal{S}^-) \ , \end{aligned}$$

even though $\mathcal{S}^+$ and $\mathcal{S}^-$ are functions of $\overleftarrow{X}$ and $\overrightarrow{X}$, respectively. Thus, the excess entropy vanishes if one conditions on the causal states: $\mathrm{I}[\overleftarrow{X}; \overrightarrow{X}|\mathcal{S}^+] = 0$.

B. Lossy Predictive Features

Lossy predictive features are naturally defined via *predictive rate-distortion* or its information-theoretic instantiations [8, 9, 24]. Interested readers can refer to Refs. [6, 7, 25] or Ref. [16, Ch. 10] for more detailed expositions of rate-distortion theory. The admittedly brief presentation here is adapted to serve our focus on prediction.

The basic setting of rate-distortion theory requires specifying two items: an information source to encode and a distortion measure d that quantifies the quality of an encoding. The focus on prediction means that the information source is a process's past $\overleftarrow{X}$ with realizations $\overleftarrow{x}$ and the relevant variable is its future $\overrightarrow{X}$. That is, we enforce the Markov chain $\mathcal{R} \to \overleftarrow{X} \to \overrightarrow{X}$ when looking for states $\mathcal{R}$ coarse-grained at a level determined by d.

Our distortion measures have the form:

$$d(\overleftarrow{x}, r) = d\big(\Pr(\overrightarrow{X}|\overleftarrow{X} = \overleftarrow{x}), \Pr(\overrightarrow{X}|\mathcal{R} = r)\big) \ .$$

This form is atypical for distortions and, technically, an extension of traditional rate-distortion theory. More typical distortions would include, for example, a normalized Hamming distance between a given $\overleftarrow{x}$ and the estimated past from the codeword $r \in \mathcal{R}$. However, these "predictive distortions" are well adapted to the applications described earlier. The minimal code rate R at expected distortion D is given by the predictive rate-distortion function:

$$R(D) = \min_{\langle d(\overleftarrow{x}, r)\rangle_{\overleftarrow{X}, \mathcal{R}} \leq D} \mathrm{I}[\mathcal{R}; \overleftarrow{X}] \ . \tag{5}$$

Determining the optimal lossy predictive features $\Pr(\mathcal{R}|\overleftarrow{X})$ that achieve these limits, as well as the predictive rate-distortion function, is the goal of predictive rate distortion theory.

Among predictive distortions, predictive informational distortions of the form

$$d(\overleftarrow{x}, r) = D_{KL}[\Pr(\overrightarrow{X}|\overleftarrow{X} = \overleftarrow{x}) || \Pr(\overrightarrow{X}|\mathcal{R} = r)] \tag{6}$$

are of special interest, as they have been well studied by others [8–10, 24, 26] and also satisfy several reasonable criteria for how one might choose a good distortion measure [27]. The expected value of a predictive information distortion is $I[\overleftarrow{X}; \overrightarrow{X}|\mathcal{R}] = \mathbf{E} - I[\mathcal{R}; \overrightarrow{X}]$, so that minimizing predictive information distortion is equivalent to maximizing $I[\mathcal{R}; \overrightarrow{X}]$. We often find it useful to define the predictive information function:

$$R(I_0) = \min_{\mathrm{I}[\mathcal{R};\overleftarrow{X}] \geq I_0} \mathrm{I}[\mathcal{R};\overleftarrow{X}] \ , \qquad (7)$$

which is related to the corresponding predictive rate-distortion function in a straightforward manner [28]. As in the literature, we refer to this as the *predictive information bottleneck* (PIB). Previous results established that the zero-distortion predictive features are a process's causal states and so the maximal $R(I_0) = C_\mu^+$ [8, 9] and this code rate occurs at an $I_0 = \mathbf{E}$ [22, 29].

The choice of method name can lead to confusion since the *recursive information bottleneck* (RIB) introduced in Ref. [30] is an information bottleneck approach to predictive inference that does not take the form of Eq. (7). However, RIB is a departure from the original IB framework since its objective function explicitly infers lossy machines rather than lossy statistics [31].

III. CURSE OF DIMENSIONALITY IN PREDICTIVE RATE-DISTORTION

Let's consider the performance of any predictive information bottleneck algorithm that clusters pasts of length M to retain information about futures of length N. When finite-block algorithms work, in the lossless limit they find features that capture $\mathrm{I}[X_{-M:0}; X_{0:N}] = \mathbf{E}(M,N)$ of the total predictable information $I[\overleftarrow{X};\overrightarrow{X}] = \mathbf{E}$ at a coding cost of $C_\mu^+(M,N)$. As $M, N \to \infty$, they should recover the forward-time causal states giving predictability $\mathbf{E}$ and coding cost C_μ^+. Increasing M and N come with an associated computational cost, though: storing the joint probability distribution $\Pr(X_{-M:0}, X_{0:N})$ of past and future finite-length trajectories requires storing $|\mathcal{A}|^{M+N}$ probabilities.

More to the point, applying these algorithms at small distortions requires storing and manipulating a matrix of dimension $|\mathcal{A}|^M \times |\mathcal{A}|^N$. This leads to obvious practical limitations—an instantiation of the *curse of dimensionality for prediction.* For example, current computing is limited to matrices of size $10^5 \times 10^5$ or less, thereby restricting rate-distortion analyses to $M, N \leq \log_{|\mathcal{A}|} 10^5$. (This is an overestimate, since the sparseness of the sequence distribution is determined by a process's topological entropy rate.) And so, even for a binary process, when $|\mathcal{A}| = 2$, one is practically limited to $M, N \leq 16$. Notably, $M, N \leq 5$ are more often used in practice [8, 9, 32–34]. Finally, note that these estimates do not account for the computational costs of managing numerical inaccuracies when measuring or manipulating the vanishingly small sequence probabilities that occur at large M and N.

These constraints compete against achieving good approximations of the information rate-distortion function: we require that $\mathbf{E} - \mathbf{E}(M,N)$ be small. Otherwise, approximate information functions provide a rather weak lower bound on the true information function for larger code rates. This has been noted before in other contexts, when approximating non-Gaussian distributions as Gaussians leads to significant underestimates of information functions [35]. This calls for an independent calibration for convergence. We address this by calculating $\mathbf{E} - \mathbf{E}(M,N)$ in terms of the transition matrix W of a process' mixed-state presentation. When W is diagonalizable with eigenvalues $\{\lambda_i\}$, Ref. [36] provides the closed-form expression:

$$\mathbf{E} - \mathbf{E}(M,N) = \sum_{i:\lambda_i \neq 1} \frac{\lambda_i^M + \lambda_i^{N+1} - \lambda_i^{M+N+1}}{1-\lambda_i} \langle \delta_\pi | W_{\lambda_i} | H(W^{\mathcal{A}}) \rangle , \qquad (8)$$

where $\langle \delta_\pi | W_{\lambda_i} | H(W^{\mathcal{A}}) \rangle$ is a dot product between the eigenvector $\langle \delta_\pi | W_{\lambda_i}$ corresponding to eigenvalue λ_i and a vector $H(W^{\mathcal{A}})$ of transition uncertainties out of each mixed state [37]. Here, π is the stationary state distribution, $\langle \delta_\pi |$ is the probability vector over mixed states with full weight on the mixed state corresponding to the stationary state distribution, and W_{λ_i} is the projection operator associated with λ_i. When W's spectral gap $\gamma = 1 - \max_{i:\lambda_i \neq 1} |\lambda_i|$ is small, then $\mathbf{E}(M,N)$ necessarily asymptotes more slowly to $\mathbf{E}$. When γ is small, then (loosely speaking) we need $M, N \sim \log_{1-\gamma}(\epsilon/\gamma)$ in order to achieve a small error $\epsilon \sim \mathbf{E} - \mathbf{E}(M,N) \ll 1$ for the predictive information function.

Figure 1(bottom) shows $\mathbf{E}(M,N)$ as a function of M and N for the Even Process, whose ϵ-machine is displayed in the top panel. The process's spectral gap $\gamma \approx 0.3$ bits and, correspondingly, we see $\mathbf{E}(M,N)/\mathbf{E}$ asymptotes slowly to 1. For example, capturing 90% of the total predictable information requires $M, N \geq 8$. (The figure caption contains more detail on allowed (M,N) pairs.) This, in turn, translates to requiring very good estimates of the probabilities of $\approx 10^4$ length-16 sequences. In Fig. 3 of Ref. [9], by way of contrast, Even Process information functions were calculated using $M = 3$ and $N = 2$. As a consequence, the estimates there captured only 27% of the full $\mathbf{E}$.

The Even Process is generated by a simple two-state HMM, so it is notable that computing its information function (done shortly in Sec. V) is at all challenging. Then again, the Even Process is an infinite-order Markov process [39].

The difficulty can easily become extreme. Altering the Even Process's lone stochastic transition probability can

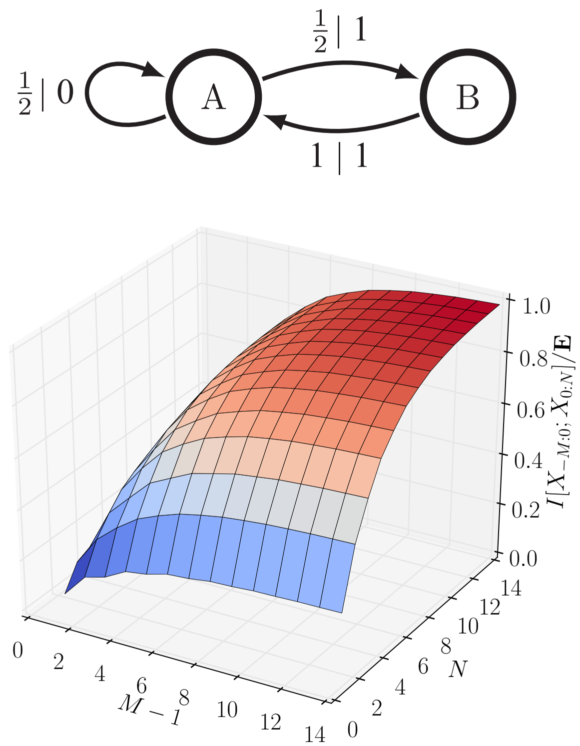


FIG. 1. Curse of dimensionality when predicting the Even Process: (top) The ϵ-machine, its minimal unifilar HMM. Edge labels $p|x$ denote generating symbol $x \in \mathcal{A}$ while taking the transition with probability p. (bottom) $\mathbf{E}(M, N)/\mathbf{E}$ as a function of N and M calculated exactly using Eq. (8) and the values of $\{\lambda_i\}$, $\langle \delta_\pi | W_{\lambda_i} | H(W^{\mathcal{A}}) \rangle$ from the Supplementary Materials for Ref. [38]. The Even Process's total predictable information $\mathbf{E} \approx 0.9183$ bits. Capturing 90% of $\mathbf{E}$ requires: $M = 6,\ N \geq 13$ or $N = 6,\ M \geq 13$; $M = 7,\ N \geq 9$ or $N = 7,\ M \geq 9$; and $M \geq 8, N \geq 8$.

increase its temporal correlations such that correctly calculating its information function requires massive compute resources. Thus, the curse of dimensionality is a critical concern even for finite-C_μ processes generated by finite HMMs.

As we move away from such simple prototype processes and towards real data sets, the attendant inaccuracies generally worsen. Many natural processes in physics, biology, neuroscience, finance, and quantitative social science are highly non-Markovian with slowly asymptoting or divergent $\mathbf{E}$ [40]. This implies rather small spectral gaps if the process has a countable infinity of causal states—e.g., as in Ref. [41]—or a distribution of eigenvalues heavily weighted near $\lambda = 0$, if the process has an uncountable infinity of causal states. In short, complex processes [4, 14] are those for which sequence-based algorithms are most likely to fail.

IV. RECASTING PREDICTIVE RATE DISTORTION THEORY

Circumventing the curse of dimensionality in predictive rate-distortion, even given an accurate model of the process, requires an alternative approach to predictive rate distortion that leverages the structural information about a process captured by that model. The results now turn to describe exactly how this structural information can be exploited. Lemma 1 equates lossy predictive features to lossy forward-time causal states. Theorem 1 shows that, for many predictive distortion measures, reverse-time causal states can replace semi-infinite futures. A corollary is that the predictive information bottleneck—compression of semi-infinite pasts to retain information about semi-infinite futures—can be recast as compression of forward-time causal states to retain information about reverse-time causal states. The joint probability distribution of forward- and reverse-time causal states may seem somewhat elusive, but previous work has shown that this joint probability distribution can be obtained given the process' model [23].

The theory builds on a simple observation: any predictive codebook can be recast as a codebook over forward-time causal states. Though the old and new codebooks have equivalent predictive distortions, the new codebook is either equivalent to or "smaller" than the old codebook. This observation is made precise by the following remark.

Remark. *Given any codebook* $\Pr(\mathcal{R}|\overleftarrow{X})$*, construct a new codebook by setting* $\Pr(\mathcal{R}|\overleftarrow{X} = \overleftarrow{x})$ *to be* $\Pr(\mathcal{R}|\mathcal{S}^+ = \epsilon^+(\overleftarrow{x}))$*. This new codebook has equivalent predictive distortion, since predictive distortion depends only on* $\Pr(r, \sigma^+)$*:*

$$
\begin{aligned}
&\mathbb{E}[d(\overleftarrow{x}, r)] \\
&\quad = \sum_{\overleftarrow{x}, r} \Pr(\overleftarrow{x}, r) d(\Pr(\overrightarrow{X}|\overleftarrow{X} = \overleftarrow{x}), \Pr(\overrightarrow{X}|\mathcal{R} = r)) \\
&\quad = \sum_{\overleftarrow{x}, r} \Pr(\overleftarrow{x}, r) d(\Pr(\overrightarrow{X}|\mathcal{S}^+ = \epsilon^+(\overleftarrow{x})), \Pr(\overrightarrow{X}|\mathcal{R} = r)) \\
&\quad = \sum_{\sigma^+, r} \Pr(\sigma^+, r) d(\Pr(\overrightarrow{X}|\mathcal{S}^+ = \sigma^+), \Pr(\overrightarrow{X}|\mathcal{R} = r)) \ .
\end{aligned}
$$

More importantly, this new codebook has equal or smaller rate, since:

$$
I[\mathcal{R}; \overleftarrow{X}] = I[\mathcal{R}; \mathcal{S}^+] + I[\mathcal{R}; \overleftarrow{X}|\mathcal{S}^+] \geq I[\mathcal{R}; \mathcal{S}^+] \ , \qquad (9)
$$

with equality when $\overleftarrow{X} \to \mathcal{S}^+ \to \mathcal{R}$*; as is true for the new, but not necessarily for the old, codebook.*

After the procedure implied by the remark, we can decrease not just the rate, but the number of predictive features by clustering together r and r' with equivalent future morphs $\Pr(\overrightarrow{X}|\cdot)$. (In a sense, two predictive features with equivalent future morphs are just copies of the same object.) Then, the number of predictive features never exceeds the number of causal states, and the entropy $H[\mathcal{R}]$ never exceeds the statistical complexity. While potentially useful—some models have rate $I[\mathcal{R};\overleftarrow{X}]$ equivalent to the statistical complexity, despite their nonminimality, effectively by copying one or more causal states—this second operation is unnecessary for the statements below.

To start, inspired by the previous finding that PIB recovers the forward-time causal states in the lossless limit [8, 9], we argue that compressing either the past $\overleftarrow{X}$ or forward-time causal states $\mathcal{S}^+$ should yield the same *lossy* predictive features. In other words, lossy predictive features are lossy causal states, and vice versa.

Lemma 1. *Compressing the past $\overleftarrow{X}$ to minimize expected predictive distortion is equivalent to compressing the forward-time causal states $\mathcal{S}^+$ to minimize expected predictive distortion.*

Proof. *A codebook that optimally compresses the past to achieve at most a distortion of $\mathbb{E}[d] \leq D$ minimizes rate $I[\mathcal{R};\overleftarrow{X}]$, while a codebook that optimally compresses forward-time causal states to achieve at most a distortion of $\mathbb{E}[d] \leq D$ minimizes a rate $I[\mathcal{R};\mathcal{S}^+]$. (See Eq. (5) and accompanying text.) Clearly, a codebook that optimally compresses forward-time causal states to minimize expected predictive distortion also optimally compresses pasts to minimize expected predictive distortion, since for such a codebook, the objective functions are equivalent: $I[\mathcal{R};\overleftarrow{X}] = I[\mathcal{R};\mathcal{S}^+]$. In the other direction, suppose that some codebook optimally compresses the past to minimize expected predictive distortion, in that it has the smallest possible rate $I[\overleftarrow{X};\mathcal{R}]$ given distortion $\mathbb{E}[d] \leq D$. From the above remark, this codebook can be conceptualized as a codebook over forward-time causal states and has rate $I[\mathcal{R};\overleftarrow{X}] = I[\mathcal{R};\mathcal{S}^+]$. Hence, the corresponding codebook over forward-time causal states also optimally compresses forward-time causal states to minimize expected predictive distortion, since for such a codebook, the objective functions are again equivalent.*

This lemma already provides a form of dimensionality reduction: semi-infinite pasts are replaced with the (potentially finite) forward-time causal states. Interestingly, in a nonprediction setting, Ref. [42] states Lemma 1 in their Eq. (2.2) without conditions on the distortion measure. However, a distortion measure that is not of the form $d(\overleftarrow{x},r) = d(\Pr(\overrightarrow{X}|\overleftarrow{X}=\overleftarrow{x}),\Pr(\overrightarrow{X}|\mathcal{R}=r))$ can still look like a predictive distortion measure, but actually incorporate potentially unnecessary information about the past, e.g. by penalizing the difference between an estimated and true future trajectories. In those situations, Lemma 1 may not apply, depending on the particular future trajectory estimator. In other situations, further dimensionality reduction is possible depending on the predictive distortion; e.g., as in Ref. [43].

When the distortion measure takes a particular special form, then we can simplify the objective function further. Our inspiration comes from Refs. [22, 29, 44] which showed that the mutual information between past and future is identical to the mutual information between forward and reverse-time causal states: $\mathrm{I}[\overleftarrow{X};\overrightarrow{X}] = \mathrm{I}[\mathcal{S}^+;\mathcal{S}^-]$. In other words, forward-time causal states $\mathcal{S}^+$ are the only features needed to predict the future as well as possible, and reverse-time causal states $\mathcal{S}^-$ are features one *can* predict about the future.

Theorem 1. *Compressing the past $\overleftarrow{X}$ to minimize expected distortion of the future $\overrightarrow{X}$ is equivalent to compressing the forward-time causal states $\mathcal{S}^+$ to minimize expected distortion of reverse-time causal states $\mathcal{S}^-$, if the predictive distortion measure is an f-divergence.*

Proof. *If $d(\cdot,\cdot)$ is an f-divergence, then it takes the form:*

$$d(\overleftarrow{x},r) = \sum_{\overrightarrow{x}} \Pr(\overrightarrow{X}=\overrightarrow{x}|\overleftarrow{X}=\overleftarrow{x}) f\left(\frac{\Pr(\overrightarrow{X}=\overrightarrow{x}|\overleftarrow{X}=\overleftarrow{x})}{\Pr(\overrightarrow{X}=\overrightarrow{x}|\mathcal{R}=r)}\right),$$

for some f [45]. Reverse-time causal states $\mathcal{S}^-$ are functions of the future $\overrightarrow{X}$ that shield the future from the past and the representation: $\mathcal{R} \to \overleftarrow{X} \to \mathcal{S}^- \to \overrightarrow{X}$. And, so:

$$\begin{aligned}&\Pr(\overrightarrow{X}=\overrightarrow{x}|\overleftarrow{X}=\overleftarrow{x}) \\ &\quad= \Pr(\overrightarrow{X}=\overrightarrow{x}|\mathcal{S}^-=\epsilon^-(\overrightarrow{x}))\Pr(\mathcal{S}^-=\epsilon^-(\overrightarrow{x})|\overleftarrow{X}=\overleftarrow{x})\end{aligned}$$

and:

$$\begin{aligned}&\Pr(\overrightarrow{X}=\overrightarrow{x}|\mathcal{R}=r) \\ &\quad= \Pr(\overrightarrow{X}=\overrightarrow{x}|\mathcal{S}^-=\epsilon^-(\overrightarrow{x}))\Pr(\mathcal{S}^-=\epsilon^-(\overrightarrow{x})|\mathcal{R}=r)~.\end{aligned}$$

In this way, predictive distortions that are f-divergences can also be expressed as:

$$\begin{aligned}&d(\overleftarrow{x},r) \\ &\quad= \sum_{\overrightarrow{x}} \Pr(\overrightarrow{X}=\overrightarrow{x}|\mathcal{S}^-=\epsilon^-(\overrightarrow{x}))\Pr(\mathcal{S}^-=\epsilon^-(\overrightarrow{x})|\overleftarrow{X}=\overleftarrow{x}) \\ &\qquad\times f\left(\frac{\Pr(\mathcal{S}^-=\epsilon^-(\overrightarrow{x})|\overleftarrow{X}=\overleftarrow{x})}{\Pr(\mathcal{S}^-=\epsilon^-(\overrightarrow{x})|\mathcal{R}=r)}\right) \\ &\quad= \sum_{\sigma^-} \Pr(\mathcal{S}^-=\sigma^-|\overleftarrow{X}=\overleftarrow{x}) f\left(\frac{\Pr(\mathcal{S}^-=\sigma^-|\overleftarrow{X}=\overleftarrow{x})}{\Pr(\mathcal{S}^-=\sigma^-|\mathcal{R}=r)}\right).\end{aligned}$$

Given this fact and Lemma 1, we recover the theorem's statement.

Distortion measures that are not f-divergences, such as mean squared-error distortion measures, implicitly emphasize predicting one reverse-time causal state over another. The Kullback-Leibler divergence given in Eq. (6), though, is an example of an f-divergence. It follows that informational predictive distortions treat all reverse-time causal states equally. Corollary 1 then follows as a particular application of Thm. 1. It recasts the predictive information bottleneck in terms of forward- and reverse-time causal states.

Corollary 1. *Compressing the past $\overleftarrow{X}$ to retain information about the future $\overrightarrow{X}$ is equivalent to compressing $\mathcal{S}^+$ to retain information about $\mathcal{S}^-$.*

Naturally, there is an equivalent version for the time-reversed setting in which past and future are swapped and the causal state sets are swapped. Also, any forward- and reverse-time prescient statistics can be used in place of $\mathcal{S}^+$ and $\mathcal{S}^-$ in any of the statements above. (Prescient statistics are essentially refinements of causal states [2].)

These proofs follow almost directly from the definitions of forward- and reverse-time causal states. Variations or portions of Lemma 1, Thm. 1, and Coro. 1 are, hopefully, intuitive. That said, to the best of our knowledge, they are also new.

Throughout, we cavalierly manipulated semi-infinite pasts and futures and their conditional and joint probability distributions—e.g., $\Pr(\overrightarrow{X}|\overleftarrow{X})$. This is mathematically suspect, since then many sums should be measure-theoretic integrals, our codebooks seemingly have an uncountable infinity of codewords, many probabilities vanish, and our distortion measures apparently divide 0 by 0. So, a more formal treatment would instead: (i) consider a series of objective functions that compress finite-length pasts to retain information about finite-length futures for a large number of lengths, giving finite codebooks and finite sequence probabilities at each length; (ii) trivially adapt the proofs of Lemma 1, Thm. 1 and Coro. 1 for these objective functions with finite-time causal states; and (iii) take the limit as those lengths go to infinity; e.g., as in Ref. [44]. As long as the finite-time forward- and reverse-time causal states limit to their infinite-length counterparts, which seems to be the case for ergodic stationary processes but not for nonergodic processes, one recovers Lemma 1, Thm. 1 and Coro. 1. We leave the task of an expanded measure-theoretic development to those with greater mathematical fortitude.

These statements nominally reduce the numerically intractable problem of clustering in the infinite-dimensional sequence space $(\overleftarrow{X},\overrightarrow{X})$ to the potentially tractable one of clustering in $(\mathcal{S}^-,\mathcal{S}^+)$. This is hugely beneficial when a process's causal state set is finite. However, many processes have an uncountable infinity of forward-time causal states or reverse-time causal states [4, 5]. Is Thm. 1 useless in these cases? Not necessarily. Predictive rate-distortion functions can be approximated to any desired accuracy by a finite or countable ϵ-machine. Additional work is required to understand how approximations of a process' minimal maximally predictive model map to approximations of its predictive rate-distortion function.

V. EXAMPLES

Theorem 1 suggests a new objective function to define lossy predictive features and predictive rate-distortion functions. It is unclear from theory alone how useful this new objective function might be. We now compare the results of an algorithm suggested by Coro. 1 to results of more commonly used PIB algorithms for several simple stochastic processes to investigate when and why moving to bidirectional model space proves useful.

To date, PIB algorithms cluster finite-length pasts to retain information about finite-length futures. For simplicity's sake, we assume that lengths of pasts and futures are both L. These algorithms find $\Pr(\mathcal{R}|\overleftarrow{X}^L)$ that maximize:

$$\mathcal{L}_\beta = \mathrm{I}[\mathcal{R};\overrightarrow{X}^L] - \beta^{-1}\,\mathrm{I}[\overleftarrow{X}^L;\mathcal{R}] \ , \qquad (10)$$

and vary the Lagrange multiplier β to achieve different distortions. We refer to such algorithms as *optimal causal filtering* (OCF). Using Coro. 1, we can instead search for a codebook $\Pr(\mathcal{R}|\mathcal{S}^+)$ that maximizes:

$$\mathcal{L}_\beta = \mathrm{I}[\mathcal{R};\mathcal{S}^-] - \beta^{-1}\,\mathrm{I}[\mathcal{S}^+;\mathcal{R}] \ , \qquad (11)$$

and again vary the Lagrange multiplier β to achieve different distortions. We refer to procedures that maximize this objective function as *causal information bottleneck* (CIB) algorithms. At large enough L, the approximated predictive features become indistinguishable from the true predictive features. However, several examples below give a rather sober illustration of the substantial errors that arise for OCF when operating at finite-L and do so for surprisingly simple processes.

We calculate solutions to both objective functions following Ref. [26]. For example, given $\Pr(\mathcal{S}^+,\mathcal{S}^-)$, then, one solves for the $\Pr(\mathcal{R}|\mathcal{S}^+)$ that maximizes the objective function in Eq. (11) at each β by iterating the dynamical system:

$$\Pr_t(r|\sigma^+) = \frac{\Pr_{t-1}(r)}{Z_t(\sigma^+,\beta)} e^{-\beta D_{\mathrm{KL}}[\Pr(\sigma^-|\sigma^+)||\Pr_{t-1}(\sigma^-|r)]} \tag{12}$$

$$\Pr_t(r) = \sum_{\sigma^+} \Pr_t(r|\sigma^+)\Pr(\sigma^+) \tag{13}$$

$$\Pr_t(\sigma^-|r) = \sum_{\sigma^+} \Pr(\sigma^-|\sigma^+)\Pr_t(\sigma^+|r) \ , \tag{14}$$

where $Z_t(\sigma^+,\beta)$ is the normalization constant for $\Pr_t(r|\sigma^+)$. Iterating Eqs. (12) and (14) at fixed β gives (i) one point on the function (R_β, D_β) and (ii) the explicit optimal lossy predictive features $\Pr(\mathcal{R}|\mathcal{S}^+)$.

We used a similar procedure to calculate finite-L approximations of information functions, but where σ^+ and σ^- are replaced by $x_{-L:0}$ and $x_{0:L}$, which are then replaced by finite-time causal states $\mathcal{S}^+_{L,L}$ and $\mathcal{S}^-_{L,L}$ using a finite-time variant of Coro. 1. The joint probability distribution of these finite-time causal states was calculated exactly by (i) calculating sequence distributions of length $2L$ directly from the ϵ-machine transition matrices and (ii) clustering these into finite-time causal states using the equivalence relation described in Sec. II A, except when the joint probability distribution was already analytically available. This procedure avoids the complications of finite sequence samples. As a result, differences between the algorithms derive entirely from a difference in objective function.

We display calculations in two ways. The first is the *information function*, a rate-distortion function that graphs the code rate $I[\overleftarrow{X};\mathcal{R}]$ versus the distortion $I[\overleftarrow{X};\overrightarrow{X}|\mathcal{R}]$ [46]. The second is a *feature curve* of code rate $I[\overleftarrow{X};\mathcal{R}]$ versus inverse temperature β. We recall that at zero temperature ($\beta \to \infty$) the code rate $I[\overleftarrow{X};\mathcal{R}] = C^+_\mu$ and the forward-time causal states are recovered: $\mathcal{R} \to \mathcal{S}^+$. At infinite temperature ($\beta = 0$) there is only a single state that provides no shielding and so the information distortion limits to $I[\overleftarrow{X};\overrightarrow{X}|\mathcal{R}] = \mathbf{E}$. As suggested by Sec. III, these extremes are useful references for monitoring convergence.

For each β, we chose 500 random initial $\Pr_0(r|\sigma^+)$, iterated Eqs. (12)-(14) 300 times, and recorded the solution with the largest $\mathcal{L}_\beta$. This procedure finds local maxima of $\mathcal{L}_\beta$, but does not necessarily find global maxima. Thus, if the resulting information function was nonmonotonic, we increased the number of randomly chosen initial $\Pr_0(r|\sigma^+)$ to 5000, increased the number of iterations to 500, and repeated the calculations. This brute force approach to the nonconvexity of the objective function was feasible here only due to analyzing processes with small ϵ-machines. Even so, the estimates might include suboptimal solutions in the lossier regime. A more sophisticated approach would leverage other results; e.g., using those of Refs. [47–49] to move carefully from high-β to low-β solutions.

Note that in contrast with deterministic annealing procedures that start at low β (high temperature) and add codewords to expand the codebook as necessary, we can also start at large β with a codebook with codewords $\mathcal{S}^+$ and decrease β, allowing the representation to naturally reduce its size. This is usually "naive" [50] due to the large number of local maxima of $\mathcal{L}_\beta$, but here, we know the zero-temperature result beforehand. More importantly, we are usually searching for the lossless predictive features at large β, but here, we are asking different questions. Of course, we could also start at low β and increase β. The key difference between the algorithm suggested by Coro. 1 and traditional predictive information bottleneck algorithms is not the algorithm itself, but the joint probability distribution of compressed and relevant variables—causal states versus sequences.

Section V A gives conditions on a process that guarantee that its information functions can be accurately calculated *without* first having a maximally-predictive model in hand. Section V B describes several processes that have first-order phase transitions in their feature curves at $\beta = 1$. Section V C describes how information functions and feature curves can change nontrivially under time reversal. Finally, Sec. V D shows how predictive features describe predictive "macrostates" for the process generated the symbolic dynamics of the chaotic Tent Map of the unit interval.

A. Unhidden and Almost Unhidden Processes

Predictive information bottleneck algorithms that cluster pasts of length $M \geq 1$ to retain information about futures of length $N \geq 1$ calculate accurate information functions when $\mathbf{E}(M,N) \approx \mathbf{E}$. (Recall Sec. III.) Such algorithms work exactly on order-R Markov processes when $M, N \geq R$, since $\mathbf{E}(R,R) = \mathbf{E}$. However, there are many processes that are "almost" order-R Markov, for which these algorithms should work quite well.

The quality of a process's approximation can be monitored by the convergence error $\mathbf{E} - \mathbf{E}(M,N)$, which is controlled by the elusive information $\sigma_\mu(L)$, defined as $I[\overleftarrow{X}; X_{L:}|X_{0:L}]$ [36]. To see this, we apply the mutual information chain rule repeatedly:

$$\begin{aligned}\mathbf{E} &= \mathrm{I}[X_{:0}; X_{0:}] \\ &= \mathrm{I}[X_{:0}; X_{0:N-1}] + \sigma_\mu(N) \\ &= \mathbf{E}(M,N) + \mathrm{I}[X_{:-M-1}; X_{0:N-1}|X_{-M-1:0}] + \sigma_\mu(N) \ .\end{aligned}$$

The last mutual information is difficult to interpret, but easy to bound:

$$\begin{aligned}\mathrm{I}[X_{:-M-1}; X_{0:N-1} | X_{-M-1:0}] \\ &\leq \mathrm{I}[X_{:-M-1}; X_{0:} | X_{-M-1:0}] \\ &= \sigma_\mu(M) \ ,\end{aligned}$$

And so, the convergence error is upper-bounded by the elusive information:

$$0 \leq \mathbf{E} - \mathbf{E}(M, N) \leq \sigma_\mu(N) + \sigma_\mu(M) \ . \quad (15)$$

The inequality of Eq. (15) suggests that, as far as accuracy is concerned, if a process has a small $\sigma_\mu(L)$ relative to its $\mathbf{E}$ for some reasonably small L, then sequences are effective states. This translates into the conclusion that for this class of process calculating information functions by first moving to causal state space is unnecessary.

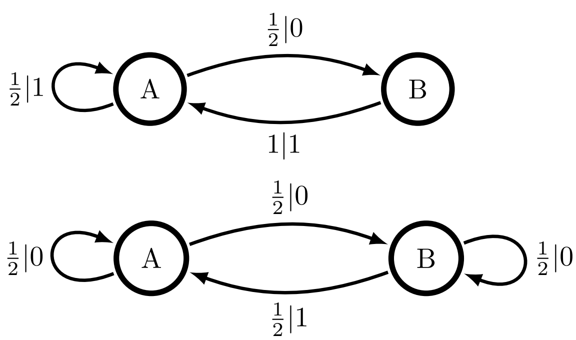


FIG. 2. (top) Golden Mean HMM, an ϵ-machine. (bottom) Simple Nonunifilar Source HMM presentation; not the SNS process's ϵ-machine.

Let's test this intuition. The prototypical example with $\sigma_\mu(1) = 0$ is the Golden Mean Process, whose HMM is shown in Fig. 2(top). It is order-1 Markov, so OCF with $L = 1$ is provably equivalent to CIB, illustrating one side of the intuition.

A more discerning test is an infinite-order Markov process with small σ_μ. One such process is the Simple Nonunifilar Source (SNS) whose (nonunifilar) HMM is shown in Fig. 2(bottom). As anticipated, Fig. 3(top) shows that OCF with $L = 1$ and CIB yield very similar information functions at low code rate and low β. In fact, many of SNS's statistics are well approximated by the Golden Mean HMM.

The feature curve in Fig. 3(bottom) reveals a slightly more nuanced story, however. The SNS is highly cryptic, in that it has a much larger C_μ than $\mathbf{E}$. As a result, OCF with $L = 1$ approximates $\mathbf{E}$ quite well but underestimates C_μ, replacing an (infinite) number of feature-discover transitions with a single transition. (More on these transitions shortly.)

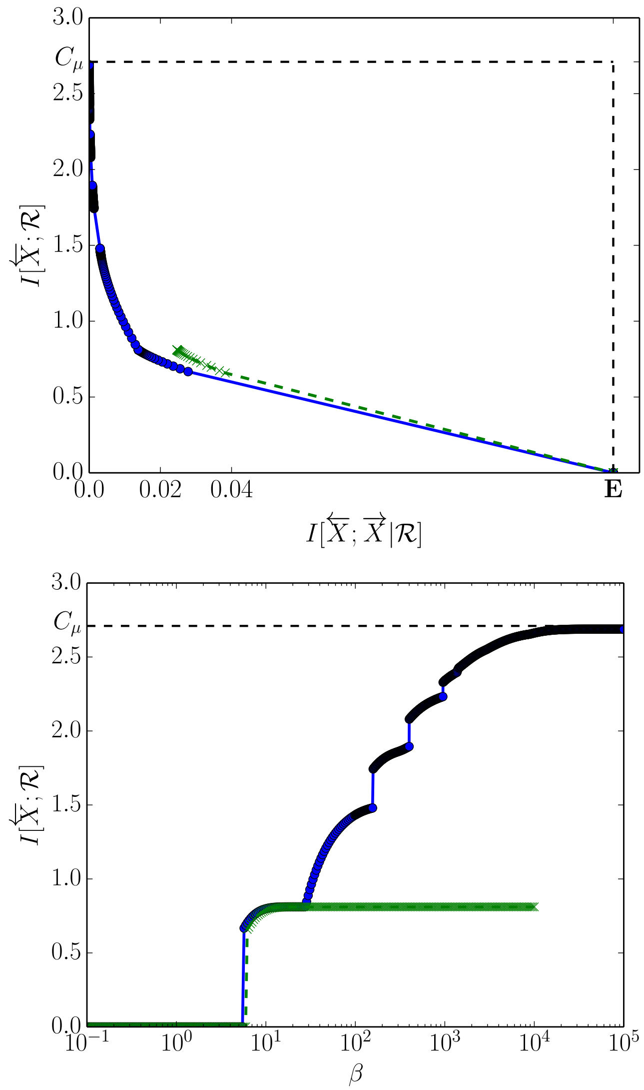


FIG. 3. Simple Nonunifilar Source: (Top panel) Information function: coding cost versus distortion. (Bottom panel) Feature curve: coding cost as a function of inverse temperature β. (Blue solid line, circles) CIB with a 10-state approximate ϵ-machine. (Green dashed line, crosses) OCF at $L = 1$.

This particular type of error—missing predictive features—only matters for predicting the SNS when low distortion is desired. Nonetheless, it is important to remember that the process implied by OCF with $L = 1$—the Golden Mean Process—is not the SNS. The Golden Mean Process is an order-1 Markov process. The SNS HMM is nonunifilar and generates an infinite-order Markov process and so provides a classic example [4] of how difficult it can be to exactly calculate information measures of stochastic processes.

Be aware that CIB cannot be directly applied to analyze the SNS, since the latter's causal state space is countably infinite; see Ref. [51]'s Fig. 3. Instead, we used finite-time causal states with finite past and future lengths and with the state probability distribution given in App. B of Ref. [51]. Here, we used $M, N = 10$, effectively approximating the SNS as an order-10 Markov process.

B. First-order Phase Transitions at $\beta = 1$

Feature curves have discontinuous jumps ("first-order phase transitions") or are nondifferentiable ("second-order phase transitions") at critical temperatures when new features or new lossy causal states are discovered. The effective dimension of the codebook changes at these transitions. Symmetry breaking plays a key role in identifying the type and temperature (β here) of phase transitions in constrained optimization [49, 52]. Using the infinite-order Markov Even Process of Sec. III, CIB allows us to explore in greater detail why and when first-order phase transitions occur at $\beta = 1$ in feature curves.

There are important qualitative differences between information functions and feature curves obtained via CIB and via OCF for the Even Process. First, as Fig. 4(top) shows, the Even Process CIB information function is a simple straight line, whereas those obtained from OCF are curved and substantially overestimate the code rate. Second, as Fig. 4(bottom) shows, the CIB feature curve is discontinuous at $\beta = 1$, indicating a single first-order phase transition and the discovery of highly predictive states. In contrast, OCF functions miss that key transition and incorrectly suggest several phase transitions at larger βs.

The first result is notable, as Ref. [9] proposed that the curvature of OCF information functions define natural scales of predictive coarse-graining. In this interpretation, linear information functions imply that the Even Process has *no* such intermediate natural scales. And, there are good reasons for this.

So, why does the Even Process exhibit a straight line? Recall that the Even Process's recurrent forward-time causal states code for whether or not one just saw an even number of 1's (state A) or an odd number of 1's (state B) since the last 0. Its recurrent reverse-time causal states (Fig. 2 in Ref. [29]) capture whether or not one will see an even number of 1's until the next 0 or an odd number of 1's until the next 0. Since one only sees an even number of 1's between successive 0's, knowing the forward-time causal state uniquely determines the reverse-time causal state and vice versa. The Even Process' forward causal-state distribution is $\Pr(\mathcal{S}^+) = \begin{pmatrix} 2/3 & 1/3 \end{pmatrix}$ and the conditional distribution of forward and reverse-time causal

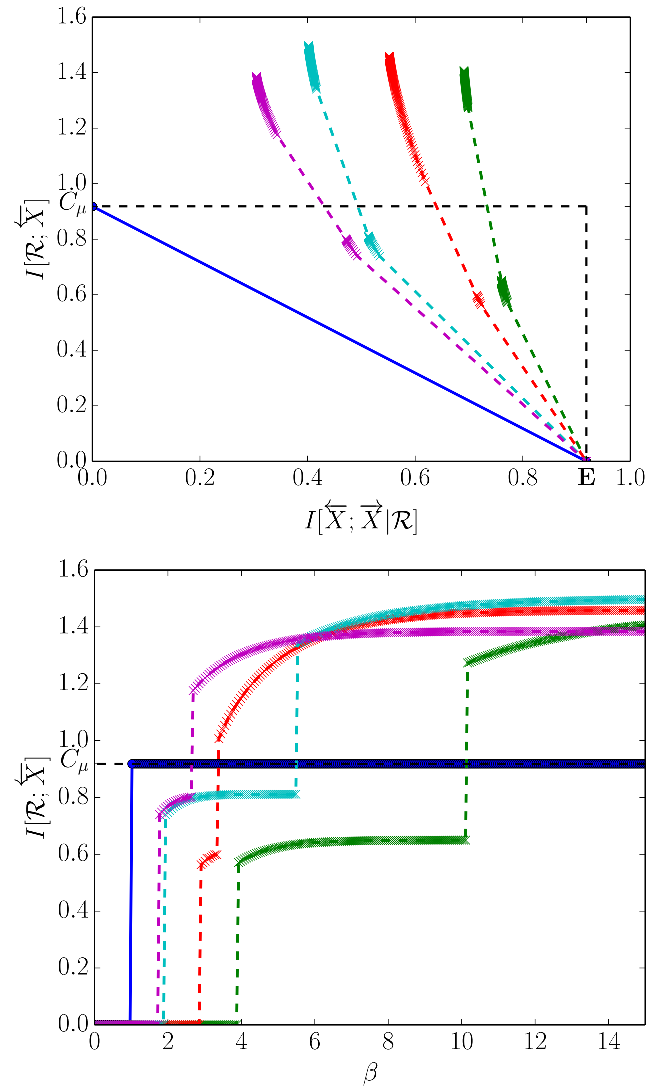


FIG. 4. Even Process analyzed with CIB (solid line, blue circles) and with OCF (dashed lines, colored crosses) at various values of $M = N = L$: (right to left) $L = 2$ (green), $L = 3$ (red), $L = 4$ (light blue), and $L = 5$ (purple). (top) Information functions. (bottom) Feature curves. At $\beta = 1$, CIB functions transition from approximating the Even Process as IID (biased coin flip) to identifying both causal states.

states is:

$$\Pr(\mathcal{S}^-|\mathcal{S}^+) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} .$$

Thus, there is an invertible transformation between $\mathcal{S}^+$ and $\mathcal{S}^-$, a conclusion that follows directly from the process's bidirectional machine. The result is that:

$$\mathrm{I}[\mathcal{R}; \mathcal{S}^+] = \mathrm{I}[\mathcal{R}; \mathcal{S}^-] . \tag{16}$$

And so, we directly calculate the information function from Eq. (7):

$$\begin{aligned} R(I_0) &= \min_{\mathrm{I}[\mathcal{R};\mathcal{S}^-] \geq I_0} \mathrm{I}[\mathcal{R};\mathcal{S}^+] \\ &= \min_{\mathrm{I}[\mathcal{R};\mathcal{S}^-] \geq I_0} \mathrm{I}[\mathcal{R};\mathcal{S}^-] \\ &= I_0 ~, \end{aligned}$$

for all $I_0 \leq \mathbf{E}$. Similar arguments hold for periodic process as described in Ref. [8, 9] and for general *cyclic* (noisy periodic) processes as well. However, periodic processes are finite-order Markov, whereas the infinite Markov-order Even Process hides its deterministic relationship between prediction and retrodiction underneath a layer of stochasticity. This suggests that the bidirectional machine's *switching maps* [29] are key to the shape of information functions.

The Even Process's feature curve in Fig. 4(bottom) shows a first-order phase transition at $\beta = 1$. Similar to periodic and cyclic processes, its lossy causal states are all-or-nothing. Iterating Eqs. (12) and (14) is an attempt to maximize the objective function of Eq. (11). However, Eq. (16) gives:

$$\mathcal{L}_\beta = (1 - \beta^{-1})\, \mathrm{I}[\mathcal{R};\mathcal{S}^+] ~.$$

Recall that $0 \leq \mathrm{I}[\mathcal{R};\mathcal{S}^+] \leq C_\mu$. For $\beta < 1$, on the one hand, maximizing $\mathcal{L}_\beta$ requires minimizing $\mathrm{I}[\mathcal{R};\mathcal{S}^+]$, so the optimal lossy model is a biased coin approximation of the Even Process—a single-state HMM. For $\beta > 1$, on the other, maximizing $\mathcal{L}_\beta$ requires maximizing $\mathrm{I}[\mathcal{R};\mathcal{S}^+]$, so the optimal lossy features are the causal states A and B themselves. At $\beta = 1$, though, $\mathcal{L}_\beta = 0$, and any representation $\mathcal{R}$ of the forward-time causal states $\mathcal{S}^+$ is optimal. In sum, the discontinuity of coding cost $\mathrm{I}[\mathcal{R};\mathcal{S}^+]$ as a function of β corresponds to a first-order phase transition and the critical inverse temperature is $\beta = 1$.

Both causal states in the Even Process are unusually predictive features: any increase in memory of such causal states is accompanied by a proportionate increase in predictive power. These states are associated with a one-to-one (switching) map between a forward-time and reverse-time causal state. In principle, such states should be the first features extracted by any predictive rate-distortion algorithm. More generally, when the joint probability distribution of forward- and reverse-time causal states can be permuted into diagonal block-matrix form, there should be a first-order phase transition at $\beta = 1$ with one new codeword for each of the blocks.

Many processes do not have probability distributions over causal states that can be permuted, even approximately, into a diagonal block-matrix form; e.g., most of those described in Refs. [51, 53]. However, we suspect that diagonal block-matrix forms for $\Pr(\mathcal{S}^+, \mathcal{S}^-)$ might be relatively common in the highly structured processes generated by low entropy-rate deterministic chaos, as such systems often have many irreducible forbidden words. Restrictions on the support of the sequence distribution easily yield blocks in the joint probability distribution of forward- and reverse-time causal states.

For example, the Even Process forbids words with an odd number of 1s, which is expressed by its *irreducible forbidden word* list $\mathcal{F} = \{01^{2k+1}0 : k = 0, 1, 2, \ldots\}$. Its causal states group pasts that end with an even (state A) or odd (state B) number of 1s since the last 0. Given the Even Process' forbidden words $\mathcal{F}$, sequences following from state A must start with an even number of ones before the next 0 and those from state B must start with an odd number of ones before the next 0. The restricted support of the Even Process' sequence distribution therefore gives its causal states substantial predictive power.

Moreover, many natural processes are produced by deterministic chaotic maps with added noise [54]. Such processes may also have $\Pr(\mathcal{S}^+, \mathcal{S}^-)$ in *nearly* diagonal block-matrix form. These joint probability distributions might be associated with sharp second-order phase transitions.

However, numerical results for the "four-blob" problem studied in Ref. [49] suggest the contrary. The joint probability distribution of compressed and relevant variables is "a discretization of a mixture of four well-separated Gaussians" [49] and has a nearly diagonal block-matrix form, with each block corresponding to one of the four blobs. If the joint probability distribution were exactly block diagonal—e.g., from a truncated mixture of Gaussians model—then the information function would be linear and the feature curve would exhibit a single first-order phase transition at $\beta = 1$ from the above arguments. The information function for the four-blob problem looks linear; see Fig. 5 of Ref. [49]. The feature curve (Fig. 4, there) is entirely different from the feature curves that we expect from our earlier analysis of the Even Process. Differences in the off-diagonal block-matrix structure allowed the annealing algorithm to discriminate between the nearly equivalent matrix blocks, so that there are three phase transitions to identify each of the four blobs. Moreover, none of the phase transitions are sharp. So, perhaps the sharpness of phase transitions in feature curves of noisy chaotic maps might have a singular noiseless limit, as is often true for information measures [53].

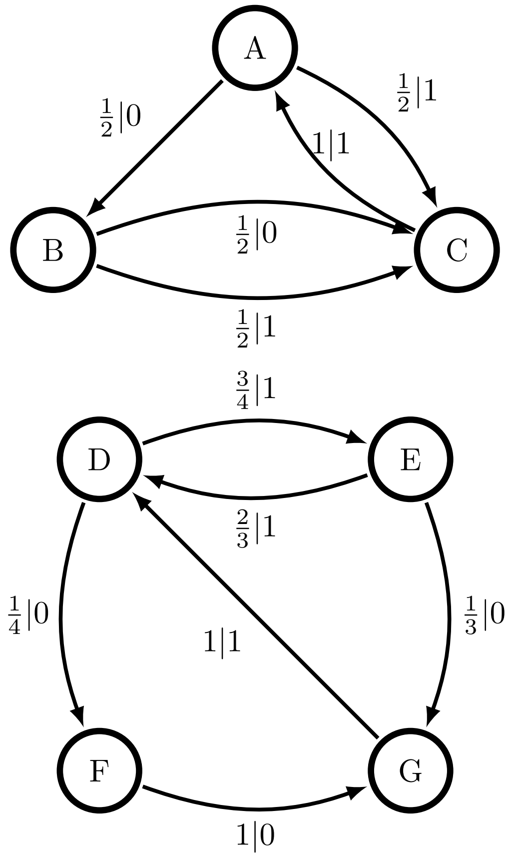


FIG. 5. Random Insertion Process (RIP): (Top) Forward-time ϵ-machine. (Bottom) Reverse-time ϵ-machine.

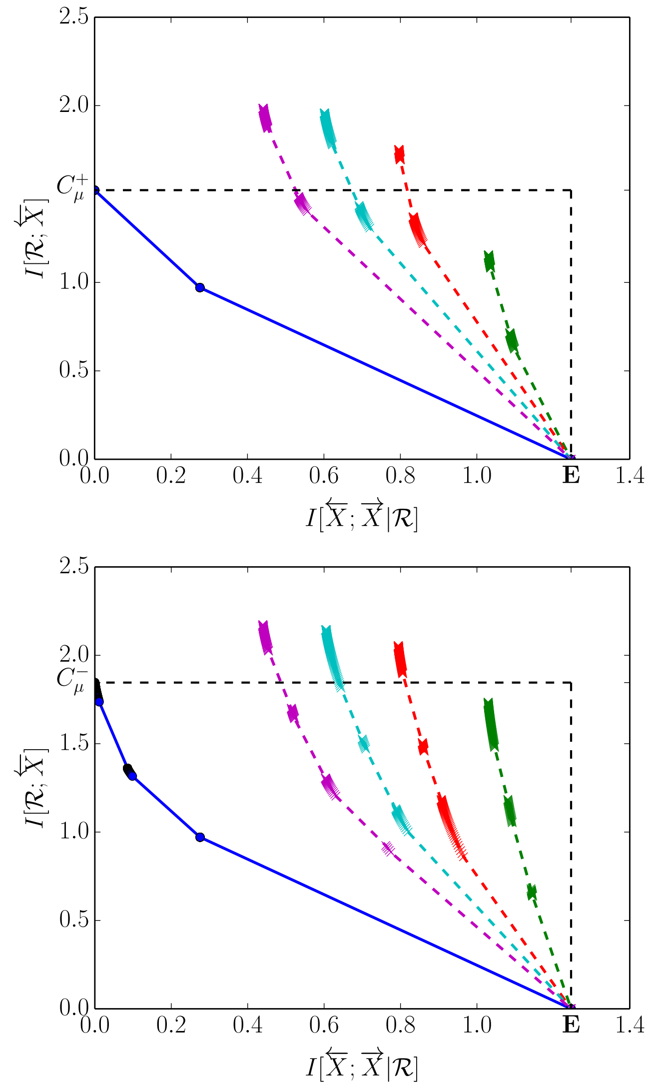


FIG. 6. Random Insertion Process (RIP) Information Functions: RIP is a causally irreversible process: $C_\mu^+ < C_\mu^-$. There are more causal states in reverse time than forward time, leading to more kinks in the reverse-time process' information function (bottom) than in the forward-time process' information function (top). Legend as in previous figure: (solid line, blue circles) CIB function and (dashed lines, colored crosses) OCF at various sequence lengths.

C. Temporal Asymmetry in Lossy Prediction

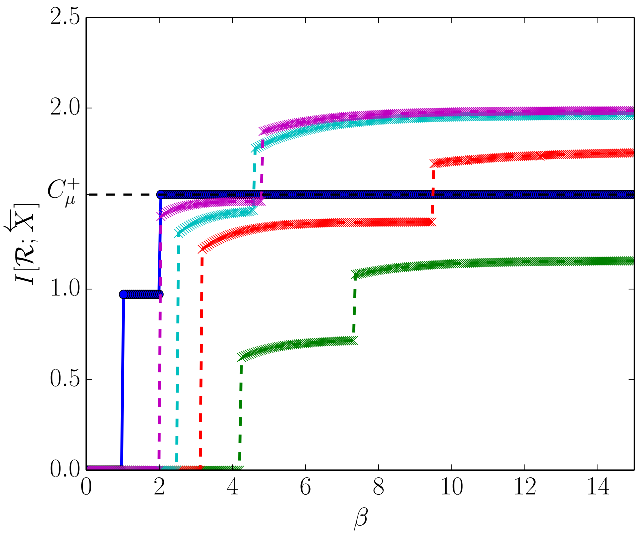
As Refs. [22, 29] describe, the resources required to losslessly predict a process can change markedly under time reversal. The prototype example is the Random Insertion Process (RIP), shown in Fig. 5. Its bidirectional machine is known analytically [22]. Therefore, we know the joint $\Pr(\mathcal{S}^+,\mathcal{S}^-)$ via $\Pr(\mathcal{S}^+) = \big(2/5\ 1/5\ 2/5\big)$ and:

$$\Pr(\mathcal{S}^-|\mathcal{S}^+) = \begin{pmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} .$$

There are three forward-time causal states and four reverse-time causal states. And, the forward-time statistical complexity and reverse-time statistical complexity are unequal, making the RIP causally irreversible. For instance, $C_\mu^+ \approx 1.8$ bits and $C_\mu^- \approx 1.5$ bits, even though the excess entropy $\mathbf{E} \approx 1.24$ bits is by definition time-reversal invariant.

However, it could be that the lossy causal states are somehow more robust to time reversal than the (lossless) causal states themselves. Let's investigate the difference in RIP's information and feature curves under time reversal. Figure 6 shows information functions for the forward-time and reverse-time processes. Despite RIP's causal irreversibility, information functions look similar until informational distortions of less than 0.1 bits. RIP's temporal correlations are sufficiently long-ranged so as to put OCF with $L \leq 5$ at a significant disadvantage relative to CIB, as the differences in the information functions demonstrate. OCF greatly underestimates $\mathbf{E}$ by about 30% and both underestimates and overestimates the correct C_μ.

The RIP feature curves in Fig. 7 reveal a similar story



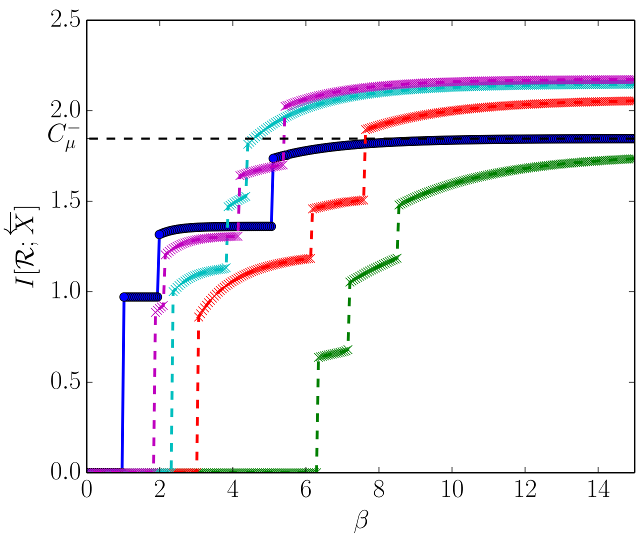


FIG. 7. Random Insertion Process (RIP) Feature Curves: Having more causal states in reverse time than forward time leads to more phase transitions in the reverse-time process' feature curve (bottom) than in the forward-time process' feature curve (top). Legend as in previous figure.

in that OCF fails to asymptote to the correct C_μ for any $L \leq 5$ in either forward or reverse time. Unlike the information functions, though, feature curves reveal temporal asymmetry in the RIP even in the lossy (low β) regime.

Both forward and reverse-time feature curves show a first-order phase transition at $\beta = 1$, at which point the forward-time causal state C and the reverse-time causal state D are added to the codebook, illustrating the argument of Sec. V B. (Forward-time causal state C and reverse-time causal state D are equivalent to the same bidirectional causal state C/D in RIP's bidirectional ϵ-machine. See Fig. 2 of Ref. [22].) This common bidirectional causal state is the main source of similarity in the information functions of Fig. 6.

Both feature curves also show phase transitions at $\beta = 2$, but similarities end there. The forward-time feature curve shows a first-order phase transition at $\beta = 2$, at which point both remaining forward-time causal states A and B are added to the codebook. The reverse-time feature curve has what looks to be a sharp second-order phase transition at $\beta = 2$, at which point the reverse-time causal state F is added to the codebook. The remaining two reverse-time causal states, E and G, are finally added to the codebook at $\beta = 5$. We leave solving for the critical temperatures and confirming the phase transition order using a bifurcation discriminator [47] to the future.

D. Predictive Hierarchy in a Dynamical System

Up to this point, the emphasis was analyzing selected prototype infinite Markov-order processes to illustrate the differences between CIB and OCF. In the following, instead we apply CIB and OCF to gain insight into a nominally more complicated process—a one-dimensional chaotic map of the unit interval—in which we emphasize the predictive features detected. We consider the symbolic dynamics of the Tent Map at the Misiurewicz parameter $a = \left(\sqrt[3]{9+\sqrt{57}} + \sqrt[3]{9-\sqrt{57}}\right)/\sqrt[3]{9}$, studied in Ref. [55]. Figure 8 gives both the Tent Map and the analytically derived ϵ-machine for its symbolic dynamics, from there. The latter reveals that the symbolic dynamic process is infinite-order Markov. The bidirectional ϵ-machine at this parameter setting is also known. Hence, one can directly calculate information functions as described in Sec. V.

From Fig. 9's information functions, one easily gleans natural coarse-grainings, scales at which there is new structure, from the functions' steep regions. As is typically true, the steepest part of the predictive information function is found at very low rates and high distortions. Though the information function of Fig. 9(top) is fairly smooth, the feature curve (Fig. 9(bottom)) reveals phase transitions where the feature space expands a lossier causal state into two distinct representations.

To appreciate the changes in underlying predictive features as a function of inverse temperature, Fig. 10 shows the probability distribution $\Pr(\mathcal{S}^+|\mathcal{R})$ over causal states given each compressed variable—the features. What we learn from such phase transitions is that some causal states are more important than others and that the most important ones are not necessarily intuitive. As we move from lossy to lossless ($\beta \to \infty$) predictive features, we add forward-time causal states to the representation in the order A, B, C, and finally D. The implication is that A is more predictive than B, which is more predic-

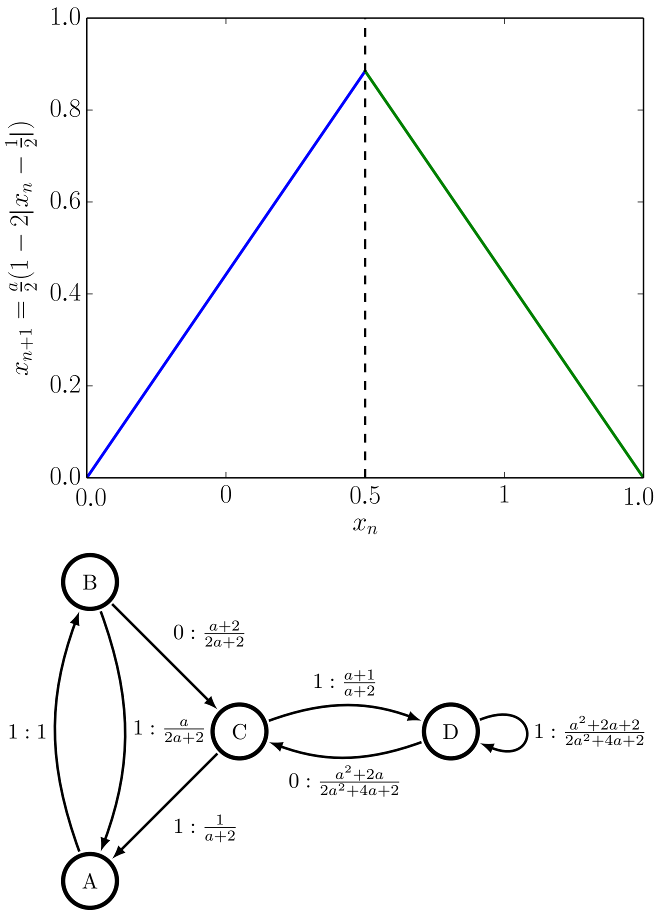


FIG. 8. Symbolic dynamics of the Tent Map at the Misiurewicz parameter a. (top) The map iterates points x_n in the unit interval $[0, 1]$ according to $x_{n+1} = \frac{a}{2}(1 - 2|x_n - \frac{1}{2}|)$, with $x_0 \in [0, 1]$. The symbolic dynamics translates the sequence $x_0, x_1, x_2, \ldots$ of real values to a 0 when $x_n \in [0, \frac{1}{2})$ and to a 1 when $x_n \in [\frac{1}{2}, 1]$. (bottom) Calculations described elsewhere [55] yield the ϵ-machine shown. (Reproduced from Ref. [55] with permission.)

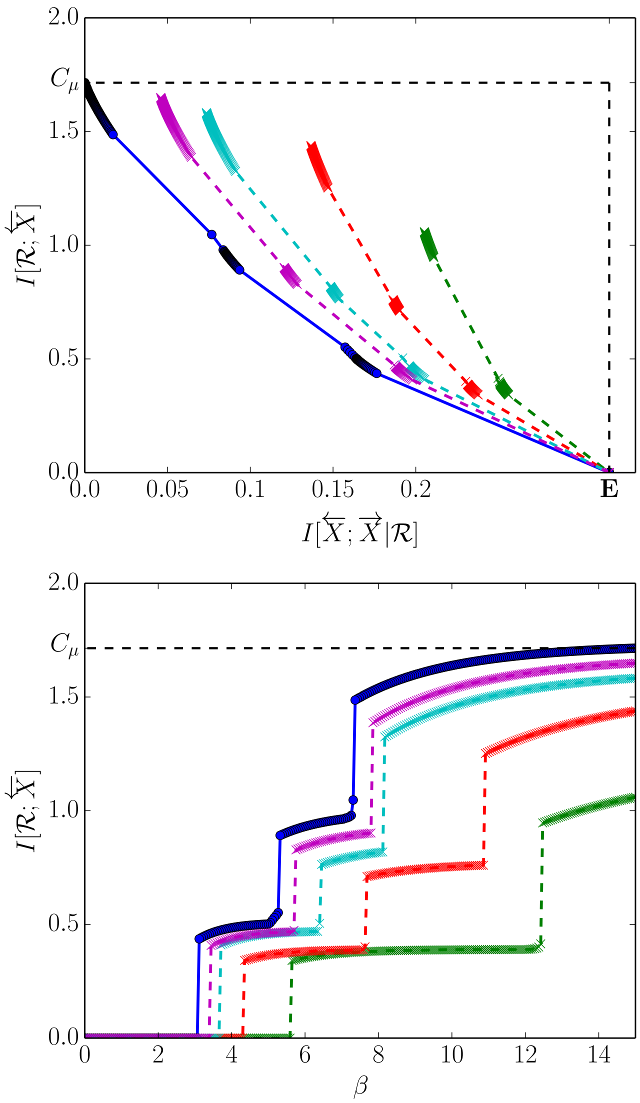


FIG. 9. Rate distortion analysis for symbolic dynamics of the Tent Map at the Misiurewicz parameter a given in the text. (top) Information functions. (bottom) Feature curves. Comparing CIB (solid line, blue circles) and OCF (dashed lines, colored crosses) at several values of L. Legend same as previous.

tive than C, which is more predictive than D. Note that this predictive hierarchy is not the same as a "stochastic hierarchy" in which one prefers causal states with smaller $\mathrm{H}[X_0|\mathcal{S}^+ = \sigma^+]$. The latter is equivalent to an ordering based on correctly predicting only one time step into the future. Such a hierarchy privileges causal state C over B based on the transition probabilities shown in Fig. 8(bottom), in contrast to how CIB orders them.

VI. CONCLUSION

We introduced a new relationship between predictive rate-distortion theory and computational mechanics [3]. Theorem 1 of Refs. [8, 9] say that the predictive information bottleneck can identify forward-time causal states, in theory. The analyses and results in Secs. III-V suggest that in practice, when studying time series with longer-range temporal correlations, we calculate substantially more accurate lossy predictive features and predictive rate-distortion functions by deriving or inferring an ϵ-machine first and working entirely within that model space.

The culprit is the curse of dimensionality for prediction: the number of possible sequences increases exponentially with their length. The longer-ranged the temporal correlations, the longer sequences need to be. And, as Secs. III and V demonstrated, a process need not have very long-ranged temporal correlations for the curse of dimensionality to rear its head. These lessons echo that

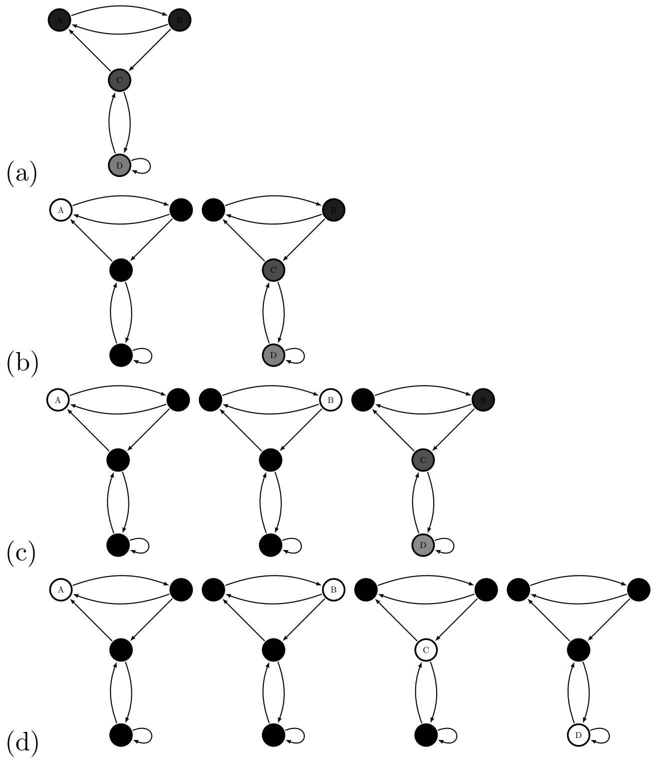


FIG. 10. Tent Map predictive features as a function of inverse temperature β: Each state-transition diagram shows the ϵ-machine in Fig. 8(bottom) with nodes gray-scaled by $\Pr(\mathcal{S}^+|\mathcal{R}=r)$ for each $r \in \mathcal{R}$. White denotes high probability and black low. Transitions are shown only to guide the eye. The four β are chosen to be close to the "critical β" at which the number of predictive features increases, shown by the β at which the feature curve in Fig. 9(bottom) appears to jump discontinuously. (a) $\beta = 0.01$: one state that puts unequal weights on states C and D. (b) $\beta = 1.9$: two states identified, A and a mixture of C and D. (c) $\beta = 3.1$: three states are identified, A, B, and the mixture of C and D. (d) $\beta \to \infty$: original four states identified, A, B, C, and D.

found when analyzing a process's large deviations [56]: Estimate a predictive model first and use it to estimate the probability of extreme events, events that almost by definition are not in the original data used for model inference.

This result is part of a larger body of work [36, 38, 57] that suggests prediction-related information properties are more accurately and more easily calculable from maximally predictive models, when available—the ϵ-machine or other prescient models [2]—than directly from trajectory distributions. These information measures are sometimes of interest to researchers, even when a model of the process is already known, because they summarize the intrinsic "uncertainty" or "predictability" of the process with a single number. A great deal of effort has been spent trying to correctly estimate such quantities from trajectory distributions [58]. Figure 11 outlines an alternative scheme to estimate such quantities: a theoretically derived, inferred, or already known model of the process is converted into a maximally predictive model using the mixed-state operator, and information measures are then estimated directly from labeled transition matrices of the maximally predictive model. In some cases, working with the so-obtained maximally predictive model may not be tractable, or the process may be effectively low-order Markov. Then, one will likely prefer to estimate information measures from trajectory distributions, simulating the process if one is initially given its model. In other cases—in particular, when the process is generated or approximately generated by finite ϵ-machines—the new scheme likely will outperform the latter.

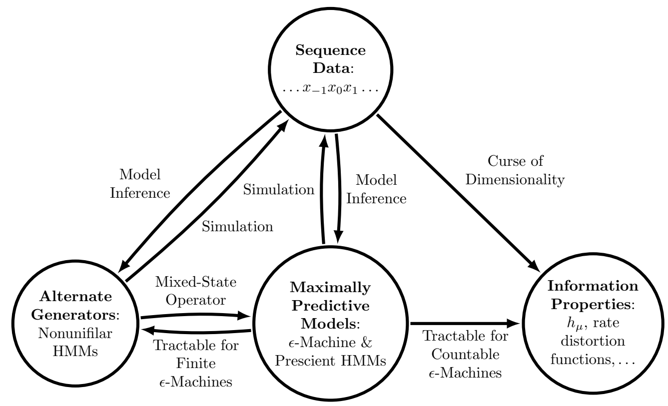


FIG. 11. **Prescient models and inferring information properties:** Estimating information measures directly from sequence data encounters a curse of dimensionality or, in other words, severe undersampling. Instead, one can calculate information measures in closed-form from (derived or inferred) maximally predictive (prescient) models [38]. Rate-distortion functions are now on the list of information properties that can be accurately calculated. Alternate generative models that are *not* prescient cannot be used directly, as Blackwell showed in the 1950s [59].

That said, cumbersome maximally predictive models are likely the norm, rather than the exception, and using approximate ϵ-machines can only yield approximate lossy predictive features. For instance, we approximated the SNS in Sec. V A by a 10-state unifilar HMM, even though the SNS technically has an infinite-state ϵ-machine. This approximation in model space led to incorrect information functions only at very low expected distortions. Future research could focus on relating distortions in model space (e.g., such as a distance between model and sequence data distributions) to errors in the rate-distortion functions. Such bounds will be important for applying CIB when only approximate ϵ-machines are known.

Section IV's methods can be directly extended to completely different rate-distortion settings, such as when the underlying minimal directed acyclic graphical model between compressed and relevant random variables is arbitrarily large and highly redundant. Also, though we mainly focused on informational distortions, Thm. 1 places fewer restrictions on the distortion measure. This opens up a wider range of applications; for example, those in which other properties, besides structure or prediction, are desired [43], including utility function optimization.

At first glance, the results presented here may seem rather unsurprising. It seems intuitive that one should be able to calculate more accurate lossy predictive features given lossless predictive features. Even so, until now, no theory or examples underlay this intuition.

At second glance, these results may also seem rather useless. Why would one want lossy predictive features when lossless predictive features are available? Accurate estimation of lossy predictive features could and have been used to further test whether or not biological organisms are near-optimal predictors of their environment [10]. Perhaps more importantly, lossless models can sometimes be rather large and hard to interpret, and a lossy model might be desired even when a lossless model is known.

Viewed in this way, the causal information bottleneck (CIB) is a new tool for accurately identifying emergent macrostates of a stochastic process [4]—lossy features relevant to interpreting biological, neurobiological, and social science phenomena in which the key emergent features are not known a priori or from first-principles calculation. In the context of neurobiological data, for example, such macrostates can provide approximately predictive models of neural spike trains [60, 61], perhaps eventually reducing large-scale simulations to more manageable models. In the context of social science data, in which "lossless" networks are often known, lossy features of various kinds might be related to new kinds of community organization. While it is encouraging to look forward, we appreciate that natural processes are quite complicated and that there is some way to go before we have fully automated detection of emergent macrostates.

ACKNOWLEDGMENTS

The authors thank C. Ellison, C. Hillar, W. Bialek, I. Nemenman, P. Riechers, and S. Still for helpful discussions. 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 contract number W911NF-13-1-0390. SM was funded by a National Science Foundation Graduate Student Research Fellowship and the U.C. Berkeley Chancellor's Fellowship.

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

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

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

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

[5] 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.

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

[7] C. E. Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE National Convention Record, Part 4*, 7:142–163, 623–656, 1959.

[8] S. Still and J. P. Crutchfield. Structure or noise? 2007. Santa Fe Institute Working Paper 2007-08-020; arxiv.org physics.gen-ph/0708.0654.

[9] S. Still, J. P. Crutchfield, and C. J. Ellison. Optimal causal inference: Estimating stored information and approximating causal architecture. *CHAOS*, 20(3):037111, 2010.

[10] S. E. Palmer, O. Marre, M. J. Berry, and W. Bialek. Predictive information in a sensory population. *Proc. Natl. Acad. Sci. USA*, 112(22):6908–6913, 2015.

[11] B. W. Andrews and P. A. Iglesias. An information-theoretic characterization of the optimal gradient sensing response of cells. *PLoS Comput. Bio.*, 3(8):1489–1497, 2007.

[12] C. R. Sims. The cost of misremembering: Inferring the loss function in visual working memory. *J. Vision*, 15(3):2, 2015.

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

[14] W. Bialek, I. Nemenman, and N. Tishby. Predictability, complexity, and learning. *Neural Comp.*, 13:2409–2463, 2001.

[15] M. Bar. Predictions: A universal principle in the operation of the human brain. *Phil. Trans. Roy. Soc. Lond. Ser. B: Bio. Sci.*, 364(1521):1181–1182, 2009.

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

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

[18] A. Paz. *Introduction to Probabilistic Automata.* Academic Press, New York, 1971.

[19] L. R. Rabiner and B. H. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, January, 1986.

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

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

[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, R. G. James, J. P. Crutchfield, and J. Reichardt. Information symmetries in irreversible processes. *CHAOS*, 21(3):037107, 2011.

[24] F. Creutzig, A. Globerson, and N. Tishby. Past-future information bottleneck in dynamical systems. *Physical Review E*, 79:041925, 2009.

[25] R. M. Gray. *Source Coding Theory*. Kluwer Academic Press, Norwell, Massachusetts, 1990.

[26] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *The 37th annual Allerton Conference on Communication, Control, and Computing*, 1999.

[27] P. Harremoës and N. Tishby. The information bottleneck revisited or how to choose a good distortion measure. In *IEEE Intl. Symposium on Information Theory. ISIT 2007*, pages 566–570. IEEE, 2007.

[28] The predictive information function $R(I_0)$ is the predictive rate-distortion function evaluated at $\mathbf{E} - I_0$.

[29] 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.

[30] S. Still. Information bottleneck approach to predictive inference. *Entropy*, 16(2):968–989, 2014.

[31] C. R. Shalizi and J. P. Crutchfield. Information bottlenecks, causal states, and statistical relevance bases: How to represent relevant information in memoryless transduction. *Adv. Comp. Sys.*, 5(1):91–95, 2002.

[32] F. Creutzig and H. Sprekeler. Predictive coding and the slowness principle: An information-theoretic approach. *Neural Computation*, 20:1026–1041, 2008.

[33] L. Gueguen and M. Datcu. Image time-series data mining based on the information-bottleneck principle. *IEEE Trans. Geo. Remote Sens.*, 45(4):827–838, 2007.

[34] L. Gueguen, C. Le Men, and M. Datcu. Analysis of satellite image time series based on information bottleneck. In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering (AIP Conference Proceedings)*, volume 872, pages 367–374, 2006.

[35] M. Rey and V. Roth. Meta-Gaussian information bottleneck. In *Adv. Neural Info. Proc. Sys.*, volume 25, pages 1925–1933. 2012.

[36] P. M. Ara, R. G. James, and J. P. Crutchfield. The elusive present: Hidden past and future dependence and why we build models. *Phys. Rev. E*, 93(2):022143, 2016. SFI Working Paper 15-07-024; arxiv.org:1507.00672 [cond-mat.stat- mech].

[37] More precisely, each element of $H(W^{\mathcal{A}})$ is the entropy in the next observation given that one is currently in the corresponding mixed state.

[38] J. P. Crutchfield, P. Riechers, and C. J. Ellison. Exact complexity: Spectral decomposition of intrinsic computation. *Phys. Lett. A*, 380(9-10):998–1002, 2015. Santa Fe Institute Working Paper 13-09-028; arXiv:1309.3792 [cond- mat.stat-mech].

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

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

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

[42] A. Banerjee, I. Dhillon, J. Ghosh, and S. Merugu. An information theoretic analysis of maximum likelihood mixture estimation for exponential families. In *Proc. Twenty-First Intl. Conf. Machine Learning*, page 8. ACM, 2004.

[43] N. Brodu. Reconstruction of ϵ-machines in predictive frameworks and decisional states. *Adv. Complex Sys.*, 14(05):761–794, 2011.

[44] J. P. Crutchfield and C. J. Ellison. The past and the future in the present. 2014. SFI Working Paper 10-12-034; arxiv.org:1012.0356 [nlin.CD].

[45] I. Csiszar. Information measures: A critical survey. In *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes, (1974)*, pages 73–86. Academia, 1977.

[46] These information functions are closely related to the more familiar information curves seen in Refs. [8, 9] and elsewhere, as the informational distortion is the excess entropy less the predictable information captured.

[47] A. E. Parker, T. Gedeon, and A. G. Dimitrov. Annealing and the rate distortion problem. In *Adv. Neural Info. Proc. Sys.*, pages 969–976, 2002.

[48] A. E. Parker and T. Gedeon. Bifurcation structure of a class of SN-invariant constrained optimization problems. *J. Dyn. Diff. Eq.*, 16(3):629–678, 2004.

[49] A. E. Parker, A. G. Dimitrov, and T. Gedeon. Symmetry breaking in soft clustering decoding of neural codes. *IEEE Trans. Info. Th.*, 56(2):901–927, 2010.

[50] G. Elidan and N. Friedman. The information bottleneck EM algorithm. In *Proc. Nineteenth Conf. Uncertainty in Art. Intel.*, UAI'03, pages 200–208, San Francisco, CA, USA, 2003. Morgan Kaufmann Publishers Inc.

[51] S. Marzen and J. P. Crutchfield. Informational and causal architecture of discrete-time renewal processes. *Entropy*, 17(7):4891–4917, 2015.

[52] K. Rose. A mapping approach to rate-distortion computation and analysis. *IEEE Trans. Info. Th.*, 40(6):1939–1952, 1994.

[53] S. Marzen and J. P. Crutchfield. Information anatomy of stochastic equilibria. *Entropy*, 16:4713–4748, 2014.

[54] J. P. Crutchfield, J. D. Farmer, and B. A. Huberman. Fluctuations and simple chaotic dynamics. *Phys. Rep.*,

92:45, 1982.

[55] R. G. James, K. Burke, and J. P. Crutchfield. Chaos forgets and remembers: Measuring information creation, destruction, and storage. *Phys. Lett. A*, 378:2124–2127, 2014.

[56] K. Young and J. P. Crutchfield. Fluctuation spectroscopy. *Chaos, Solitons, and Fractals*, 4:5 – 39, 1994.

[57] P. M. Riechers, J. R. Mahoney, C. Aghamohammadi, and J. P. Crutchfield. A closed-form shave from occam's quantum razor: Exact results for quantum compression. 2015.

[58] I. Nemenman, F. Shafee, and W. Bialek. Entropy and inference, revisited. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Adv. Neural Info. Proc. Sys.*, volume 14, Cambridge, Massachusetts, 2002. MIT Press.

[59] D. Blackwell. The entropy of functions of finite-state Markov chains. volume 28, pages 13–20, Publishing House of the Czechoslovak Academy of Sciences, Prague, 1957. Held at Liblice near Prague from November 28 to 30, 1956.

[60] R. Haslinger, K. L. Klinkner, and C. R. Shalizi. The computational structure of spike trains. *Neural Comp.*, 22:121–157, 2010.

[61] R. Watson. *The Structure of Dynamic Memory in a Simple Model of Inhibitory Neural Feedback.* PhD thesis, University of California, Davis, 2014.