Circumventing the Curse of Dimensionality in Prediction:  
Causal Rate-Distortion for Infinite-Order Markov Processes

Sarah Marzen\textsuperscript{1} and James P. Crutchfield\textsuperscript{2}

\textsuperscript{1}Department of Physics,  
Redwood Center for Theoretical Neuroscience  
University of California at Berkeley, Berkeley, CA 94720-5800  
\textsuperscript{2}Complexity Sciences Center and Department of Physics  
University of California at Davis, One Shields Avenue, Davis, CA 95616

(Dated: March 17, 2015)

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 show that algorithms which 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 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 causal rate-distortion theory substantially improves current predictive rate-distortion analyses.

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 maximally predictive features in a time series, or the forward-time causal states $S^+$ of computational mechanics. Storing these states costs on average $C^+_{\mu} = H[S^+]$ bits \cite{prl10,prl11}. However, for most processes \cite{pmlr13,pmlr14}, $C^+_{\mu}$ 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 \cite{shannon59,source}. 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 \cite{prl10,prl11}. This can yield reasonable estimates of predictive rate-distortion functions at sufficient lengths, but how long is long enough?

We introduce the causal rate-distortion (CRD) algorithm for calculating predictive rate-distortion functions and lossy predictive features. The heart of CRD 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 \cite{prl10,prl11}. The theorem allows us to calculate predictive rate-distortion functions within model space rather than within sequence space. This ameliorates, and sometimes eliminates, the aforementioned curse of dimensionality.

The algorithm’s usefulness naturally depends on the quality of the model with which one starts—computational mechanics’ $\epsilon$-machine. If calculated analytically (e.g., see Refs. \cite{garner88,crutchfield07}) then the $\epsilon$-machine is exact, but if estimated empirically (e.g., see Refs. \cite{garner88,crutchfield07}) then it will have sampling errors. In either case, though, we show that CRD substantially outperforms current predictive rate-distortion algorithms, illustrating this by calculating lossy predictive features of several processes.
with known ε-machines. A hierarchy of phase transitions associated with discovering new predictive features emerges as a function of approximation error, paralleling those described in Ref. [13]. As an illustration, we calculate a predictive “hierarchy” for an infinite Markov order process generated by a complex chaotic dynamical system.

Section II reviews computational mechanics and predictive rate-distortion theory. Section III describes how current predictive rate-distortion algorithms encounter the curse of dimensionality. Section IV introduces the CRD theorem that reformulates predictive rate-distortion analysis in terms of forward- and reverse-time causal states. Section V then describes the CRD algorithm for computing lossy causal states and illustrates its performance on infinite-order Markov processes. Section VII 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 [5 6].

A. Problem Statement

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 is certainly the circumstance in which most of measurement science finds itself. Instrumentation almost never exactly captures an experimental system’s states exactly. One can argue that this is even guaranteed by quantum uncertainty relations. Similarly for biological life and engineered adaptive systems, their sensory apparatus does not capture all of an environment’s information and organization. Indeed, in many cases they cannot due to sensory limitations. Even without such limitations, moreover, they may not have adequate representational capacity.

And, perhaps more profoundly, one does not want to do perfectly, as the “stammering grandeur” of Irenio Funes reminds us [14, Funes The Memorious]. Said positively, 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 been focused on determining memory and the ability to reproduce a given time series [15], but that memory may only be important to the extent that it affects the ability to predict the future; e.g., see Refs. [8 16 17].

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

First, we review causal states. Second, we review several information measures of stochastic processes. These, finally, lead us to describe what we mean by lossy causal states. The following assumes familiarity with information theory at the level of Refs. [13 19], information theory for complex processes at the level of Refs. [20 21], and computational mechanics at the level of Ref. [22].

B. Processes and Their Causal States

When predicting a system the main object is the process P it generates: the list of all of a system’s behaviors or realizations \{...x_{-2}, x_{-1}, x_0, x_1, ...\} as specified by their joint probabilities \text{Pr}(...X_{-2}, X_{-1}, X_0, X_1, ...). We denote a contiguous chain of random variables as \text{X}_{0:t} = X_0 X_1 ... X_{t-1}. Left indices are inclusive; right, exclusive. We suppress indices that are infinite. In this setting, the present \text{X}_{t:t+\ell} is the length-\ell chain beginning at t, the past is the chain \text{X}_{t:t-2} leading up the present, and the future is the chain following the present \text{X}_{t:t+\ell} = X_{t+\ell+1} X_{t+\ell+2} ... . When being more expository, we use arrow notation; for example, for the past \text{X} = \text{X}_0 and future \text{X} = \text{X}_0. We refer on occasion to the space \text{X} of all pasts. Finally, we assume a process is ergodic and stationary—\text{Pr}(\text{X}_{0:t}) = \text{Pr}(\text{X}_{t:t+1}) for all \text{t} \in \text{Z}— and the measurement symbols \text{x}_i range over a finite alphabet: \text{x} \in \text{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 [23].

Forward-time causal states \text{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^+:

\text{x}_0 \sim^+ \text{x}_0' \iff \text{Pr}(\text{X}_0 | \text{X}_0 = \text{x}_0) = \text{Pr}(\text{X}_0 | \text{X}_0 = \text{x}_0') . \quad (1)

As a shorthand, we denote a cluster of pasts so defined, a causal state, as \sigma^+ \in \text{S}^+. We implement Eq. (1) via the causal state map: \sigma^+ = \epsilon^+(\pi). Through it, each state \sigma^+ inherits a probability \text{Pr}(\sigma^+) from the process’s probability over pasts \text{Pr}(\text{X}_0). The forward-time statis-
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 \( E \) bits of future information that are predictable. Second, they shield the past and future from one another. That is:

\[
\Pr(X, \bar{X} | S^+) = \Pr(X | S^+) \Pr(\bar{X} | S^+) \quad \text{and}
\]

\[
\Pr(X, \bar{X} | S^-) = \Pr(X | S^-) \Pr(\bar{X} | S^-) ,
\]

even though \( S^+ \) and \( S^- \) are functions of \( \bar{X} \) and \( \bar{X} \), respectively. Thus, the excess entropy vanishes if one conditions on the causal states: e.g., \( I[X; \bar{X} | S^+] = 0 \). Figure 1 illustrates shielding the future \( X \) from the past \( \bar{X} \), given by the forward causal states \( S^+ \), and the past from the future, given by the reverse causal states \( S^- \). The result is a series of Markov chains forming a Markov loop that illustrates the relationship between prediction and retrodiction.

### C. Lossy Predictive Features

Lossy predictive features are naturally defined via predictive rate-distortion (PRD) or its information-theoretic instantiations, optimal causal inference (OCI) \(^7\) \(^8\) and the past-future information bottleneck (PFIB) \(^31\). Appendix \(^3\) briefly reviews rate-distortion theory, but interested readers should also refer to Refs. \(^5\) \(^6\) \(^32\) or Ref. \(^19\) Ch. 8 for detailed expositions. The presentation here is adapted to serve our focus on prediction.

The basic setting of rate-distortion theory is depicted schematically in Fig. 2. The 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 \( \bar{X} \) with realizations \( \bar{x} \) and the relevant variable is its future \( X \). That is, we have the
These determine candidate random variables $\mathcal{R}_\beta$ from the past. Relations: $\Rightarrow$ denotes “function of”, $\rightarrow$ denotes a Markov chain, and $Y \sim Z$ indicates that random variable $Z$ is the “relevant” variable for $Y$. The latter does not imply a Markov chain relationship.

Markov chain $\mathcal{R} \rightarrow \hat{X} \rightarrow \hat{X}$ in Fig. 2(bottom). Our distortion measures have the form:

$$d(\mathcal{R}, r) = d(\Pr(\hat{X} | X = \mathcal{R}), \Pr(\hat{X} | \mathcal{R} = r)).$$

The minimal code rate $R$ for a given expected distortion $D$ is given by the rate-distortion function:

$$R(D) = \min_{d(\mathcal{R}, r) | \mathcal{R}, r \leq D} I[\mathcal{R}; \hat{X}] .$$

Determining the optimal $\Pr(\mathcal{R}|\hat{X})$ that achieve these limits is the goal of predictive rate distortion (PRD) theory.

If we restrict to informational distortions, such as the conditional excess entropy $I[X; \hat{X} | \mathcal{R}]$, we find the associated information rate-distortion function:

$$R(I_0) = \min_{I[X; \hat{X}] \geq I_0} I[\mathcal{R}; \hat{X}] , \quad (5)$$

and its accompanying objective function:

$$\mathcal{L}_\beta = I[\mathcal{R}; \hat{X}] - \beta^{-1} I[\hat{X}; \mathcal{R}] . \quad (6)$$

These determine candidate random variables $\mathcal{R}_\beta$ that have a minimum level of predictability $(I[\mathcal{R}; \hat{X}])$ while minimizing the amount of information $(I[X; \hat{X}])$ kept from the past. $\beta$ controls the trade-off between prediction and memory. Calculating the $\Pr(\mathcal{R}|\hat{X})$ that maximize Eq. (6) constitutes optimal causal inference (OCI) [7 8] or, for linear dynamical systems, the past-future information bottleneck [31]. We refer to these methods generally as the predictive information bottleneck (PIB), emphasizing that an information distortion measure for PRD has been chosen.

The choice of method name can lead to confusion since the recursive information bottleneck (RIB) introduced in Ref. [33] is an information bottleneck approach to predictive inference that does not take the form of Eq. (6). However, RIB is a departure from the original IB framework since its objective function explicitly infers lossy machines rather than lossy statistics [34].

Generally, rate-distortion analysis reveals how a process’s structure (captured in the associated codebooks) varies under coarse-graining, with the shape of the rate-distortion functions $R(D)$ identifying those regimes in which smaller codebooks are good approximations. To implement this analysis, one graphs a process’s information function $R(I_0)$ and its accompanying feature curve $(\beta, R_\beta)$, corresponding to optimizing $\mathcal{L}_\beta$ above. Again, previous results established that the zero-distortion predictive features are a process’s causal states and so the maximal $R(I_0) = C^{++}_\mu$ and this code rate occurs at an $I_0 = E$ [28 30].

Figure 2(bottom) gives the context in which the approximately predictive states $\mathcal{R}_\beta$ are soft clusters of the past $\hat{X}$ constrained by predicting the future $\hat{X}$ at a fidelity controlled by $\beta$. $\mathcal{R}_\beta$ there indicates which random variable is “relevant” and this will be useful for explaining Thm. 1 via Fig. 4(b).

### III. CURSE OF DIMENSIONALITY IN PREDICTION

Let’s consider the performance of any PRD 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 $I[X_{-M:0}; X_{0:N}] = E(M, N)$ of the total predictable information $I[X; \hat{X}] = E$ at a coding cost of $C^{++}_\mu(M, N)$. As $M, N \rightarrow \infty$, they should recover the forward-time causal states giving predictability $E$ and coding cost $C^{++}_\mu$. 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 $|A|^{M+N}$ probabilities.

More to the point, applying PRD algorithms at small distortions requires storing and manipulating a matrix of dimension $|A|^{M} \times |A|^{N}$. This leads to obvious practical limitations—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 $M, N \leq \log_{|A|} 10^5$ (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 $|A| = 2$, one is practically limited to $M, N \leq 16$. Notably, $M, N \leq 5$ are more often used in practice [7 8 35 37]. 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 \( E - E(M, N) \) 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 [38]. This calls for an independent calibration for convergence. We address this by calculating its information function.

\[
E - E(M, N) = \sum_{i: \lambda_i \neq 1} \frac{\lambda_i^M + \lambda_i^{N+1} - \lambda_i^{M+N+1}}{1 - \lambda_i} \langle \delta_{\pi|i} W_{\lambda_i} | H(W^A) \rangle, \tag{7}
\]

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

Figure 3 (bottom) shows \( E(M, N) \) as a function of \( M \) and \( N \) for the Even Process, whose \( \epsilon \)-machine is displayed in the top panel. The process’s spectral gap is \( \approx 0.3 \) bits and, correspondingly, we see \( E(M, N)/E \) 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. [39], by way of contrast, Even Process information functions were calculated using \( M = 3 \) and \( N = 2 \). As a consequence, the OCF estimates there captured only 27% of the full \( E \).

The Even Process is generated by a simple two-state HMM, so it is notable that computing its information function (done shortly in Sec. 4) is at all challenging. Then again, the Even Process is an infinite-order Markov process.

The difficulty can easily become extreme. Altering the

![Figure 3](image-url)
IV. CAUSAL RATE DISTORTION THEORY

Circumventing the curse of dimensionality requires an alternative approach to predictive rate distortion (PRD)—causal rate distortion (CRD) theory—that leverages the structural information about a process captured by its forward and reverse causal states [44]. Prop. 1, our first main result, establishes that CRD is equivalent to PRD under certain conditions, leading to a new and efficient method to calculate information functions for a broad range of distortion measures. Our second main result adapts CRD to informational distortions, introducing an efficient PIB alternative—causal information bottleneck (CIB). It shows that compressing \( \hat{X} \) to retain information about \( \hat{X} \) is equivalent to compressing \( S^+ \) to retain information about \( S^- \). These results combine and generalize two previous theorems in Refs. [7, 8, 28, 30]. This section first describes the relationship between PRD and CRD and then that between PIB and CIB in an intuitive way. Appendix D gives more precise statements and their proofs. Section V's examples illustrate CRD’s benefits, partly by comparing to previous methods and partly through new analytical insights into information functions.

To start, inspired by the previous finding that PIB recovers the forward-time causal states in the zero-temperature (\( \beta \to \infty \)) limit [7] [8], we argue that compressing either the past \( \hat{X} \) or forward-time causal states \( S^+ \) should yield the same predictive features at any temperature. This leads us to consider two different rate-distortion settings. In the first, traditional PRD setting we compress the past \( \hat{X} \) to minimize expected distortion about the future \( \hat{X} \). In the second, CRD setting we compress the forward-time causal states \( S^+ \) to retain information about the future \( \hat{X} \). Though somewhat intuitive, there is no a priori reason that the two objective functions are related. The Markov chain of Fig. 4 (top) lays out the implied interdependencies. For many distortion functions, in fact, they are not related. Lemma 1, however, says that they are equivalent for certain types of predictive distortion measures.

**Lemma 1.** Compressing the past \( \hat{X} \) to minimize expected predictive distortion is equivalent to compressing the forward-time causal states \( S^+ \) to minimize expected predictive distortion if the distortion measure is expressible as \( d(\Pr(\hat{X} | R = r), \Pr(X | \hat{X} = \hat{x})) \).

Many distortion measures can be expressed in this form, though the mapping can be nonobvious. For instance, a distortion measure that penalizes the Hamming distance between the maximum likelihood estimates of the most likely length-\( L \) future is actually expressible in terms of conditional probability distributions of futures given pasts or features. Lemma 1 then suggests that PRD will recover the forward-time causal states in the zero-temperature limit for a great many prediction-related distortion measures, not just the information distortions considered in Refs. [7, 8]. However, for distortion measures that ask for optimal predictions of an arbitrary coarse-graining of futures, the forward-time causal states will be no longer be sufficient statistics; cf. decisional states [45]. Then, PRD’s zero-temperature limit will not recover the forward-time causal states.

Interestingly, in a nonprediction setting, Ref. 10 states Lemma 1 in their Eq. (2.2) without conditions on the distortion measure. At least in the context of PRD, this can be an oversimplification. An entirely reasonable predictive distortion measure could penalize the difference between the output of a particular prediction algorithm applied to the true past versus an estimated past. Without proper conditions, these prediction algorithms can incorporate aspects of the past \( \hat{X} \) that are entirely useless for prediction. A version of Lemma 1 will still apply, but the variable that replaces \( \hat{X} \) depends on the particular prediction algorithm. For instance, predictions of future sequences of the Even Process (Sec. III) based on certain ARIMA estimators will store unnecessary information about the past. This makes the forward-time causal states insufficient statistics with re-

![FIG. 4. Causal Rate Distortion (CRD) Markov chain relationships between the past, future, causal states \( S^+ \) and \( S^- \), and lossy predictive features \( R, \hat{R} \). Diagrammatic notation as in Fig. 2 (bottom).](image-url)
spective to the process and the predictor; see Sec. \ref{V.B}. Ideally, prediction algorithms should be well matched to the stochastic process’ model class.

When the distortion measure takes a particular special form, then we can simplify the objective function further. Our inspiration comes from Refs. \cite{28,30,47} which showed that the mutual information between past and future is identical to the mutual information between forward and reverse-time causal states: \( I[\tilde{X}; \tilde{X}] = I[S^+; S^-] \). In other words, forward-time causal states \( S^+ \) are the only features needed to predict the future as well as possible, and reverse-time causal states \( S^- \) are features one can predict about the future.

And so, we now consider a third objective function that compresses the forward-time causal states to minimize expected distortion about the reverse-time causal states. Proposition \ref{prop1} says that this objective function is equivalent to compressing the past to minimize expected distortion for a class of distortion measures.

**Proposition 1** (Causal Rate Distortion). Compressing the past \( \tilde{X} \) to minimize expected distortion of the future \( \tilde{X} \) is equivalent to compressing the forward-time causal states \( S^+ \) to minimize expected distortion of reverse-time causal states \( S^- \), if the distortion measure is expressible as:

\[
d'(\tilde{x}, r) = d(\Pr(\tilde{X} | \mathcal{R} = r), \Pr(\tilde{X} | \tilde{X} = \tilde{x}))
\]

and satisfies:

\[
d(\Pr(\tilde{X} | \mathcal{R} = r), \Pr(\tilde{X} | \tilde{X} = \tilde{x})) = d(\Pr(S^- | \mathcal{R} = r), \Pr(S^- | \tilde{X} = \tilde{x})).
\]

Distortion measures that do not satisfy Prop. \ref{prop1}’s conditions, such as mean squared-error distortion measures, in effect emphasize predicting one reverse-time causal state over another. Even then, inferring a maximally-predictive model can improve calculational accuracy for nearly any distortion measure on sequence distributions. For details, see App. \ref{appB}.

Informational distortion measures, though, treat all reverse-time causal states equally. Leveraging this, Thm. \ref{thm1} follows as a particular application of Prop. \ref{prop1}.

**Theorem 1** (Causal Information Bottleneck). Compressing the past \( \tilde{X} \) to retain information about the future \( \tilde{X} \) is equivalent to compressing \( S^+ \) to retain information about \( S^- \).

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 \cite{2} can be used in place of \( S^+ \) and \( S^- \) in any of the statements above.

Appendix \ref{appB}’s proofs follow almost directly from the definitions of forward- and reverse-time causal states. Variations or portions of Lemma \ref{lem1}, Prop. \ref{prop1} and Thm. \ref{thm1} may seem intuitive. Appendix \ref{appB} points this out when clearly the case. That said, to the best of our knowledge, Lemma \ref{lem1}, Prop. \ref{prop1} and Thm. \ref{thm1} are new.

Theorem \ref{thm1} and Prop. \ref{prop1} reduce the numerically intractable problem of clustering in the infinite-dimensional sequence space \((\tilde{X}, \tilde{X})\) to the potentially tractable one of clustering in \((S^-, S^+)\). This is 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 \cite{3,4}. Is Theorem \ref{thm1} useless in these cases? Not at all. A practical approach is that information functions can be approximated to any desired accuracy by a finite or countable \( \epsilon \)-machine. However, additional work is required to understand how model approximations map to information-function approximations.

**V. EXAMPLES**

To illustrate CRD and CIB, we find lossy causal states and calculate information functions for processes generated by known \( \epsilon \)-machines. For several, the \( \epsilon \)-machines are sufficiently simple that the information functions can be obtained analytically using the above results. These allow us to comment more generally on the shape of information functions for several broad classes of process. The examples also provide a setting in which to compare CIB information functions to those from optimal causal filtering (OCF) \cite{2,8}. Though, Thm. \ref{thm1} says that OCF and CIB yield identical results in the \( L \to \infty \) limit, the examples give a rather sober illustration of the substantial errors that arise using OCF at finite \( L \).

We display the results of the PRD analyses in two ways \cite{48}. The first is the information function, a rate-distortion function that graphs the code rate \( I[\tilde{X}; \mathcal{R}] \) versus the distortion \( I[\tilde{X}; \tilde{X} | \mathcal{R}] \). The second is a feature curve of code rate \( I[\tilde{X}; \mathcal{R}] \) versus inverse temperature \( \beta \). We recall that at zero temperature \((\beta \to \infty)\) the code rate \( I[\tilde{X}; \mathcal{R}] = C^+_{\mu} \) and the forward-time causal states are recovered: \( \mathcal{R} \leftrightarrow S^+ \). At infinite temperature \((\beta = 0)\) there is only a single state that provides no shielding and so the information distortion limits to \( I[\tilde{X}; \tilde{X} | \mathcal{R}] = 0 \). As we will see, these extremes are useful references for monitoring convergence.

We calculate information functions and feature curves following Ref. \cite{49}. (So that the development here is self-contained, App. \ref{appC} reviews this approach, but as adapted to our focus on prediction.) Given \( \Pr(\sigma^-; \sigma^+) \), then, one
solves for the $\Pr(r|\sigma^+)$ and $\Pr(\sigma^+)$ that maximize the CIB objective function at each $\beta$:

$$L_\beta = I[\mathcal{R}; S^-] - \beta^{-1} I[S^+; \mathcal{R}] ,$$

by iterating the dynamical system:

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

(9)

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

(10)

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

(11)

where $Z_t(\sigma^+, \beta)$ is the normalization constant for $\Pr_t(r|\sigma^+)$. Iterating Eqs. (9) and (11) at fixed $\beta$ gives (i) one point on the function $(R_\beta, D_\beta)$ and (ii) the explicit optimal lossy predictive features $\mathcal{R}_\beta = \{ \Pr_t(r|\sigma^+) \}$.

For each $\beta$, we chose 500 random initial $\Pr_0(r|\sigma^+)$, iterated Eqs. (9)-(11) 300 times, and recorded the solution with the largest $L_\beta$. This procedure finds local maxima of $L_\beta$, but does not necessarily find global maxima. Thus, if the resulting information function was non-monotonic, 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 $\epsilon$-machines. Even so, the estimates might include suboptimal solutions in the lossier regime. A more sophisticated approach could leverage the results of Ref. [1] to move carefully from high-$\beta$ to low-$\beta$ solutions.

We used a similar procedure to calculate OCF functions, but $\sigma^+$ and $\sigma^-$ were replaced by $x_{-L:L}$ and $x_{0:L}$, which were then replaced by finite-time causal states $S^+_{L,L}$ and $S^-_{L,L}$ using a finite-time variant of CIB. The joint probability distribution of these finite-time causal states was calculated exactly by (i) calculating sequence distributions of length 2$L$ directly from the $\epsilon$-machine transition matrices and (ii) clustering these into finite-time causal states using the equivalence relation described in Sec. II B, 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 results produced by CIB and OCF are entirely a difference in the objective function.

Note that in contrast with deterministic annealing procedures that start at low $\beta$ (high temperature) and add codewords to expand the codebook as necessary, CRD algorithms could also start at large $\beta$ and decrease $\beta$, allowing the representation to naturally reduce its size. This is usually “naive” due to the large number of local maxima of $L_\beta$, but here, we know the zero-temperature result beforehand. Of course, CRD algorithms could also start at low $\beta$ and increase $\beta$. The key difference between CRD and PRD, and between CIB and PIB, is not the algorithm itself, but the joint probability distribution of compressed and relevant variables.

Section V A gives conditions on a process which 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.

A. Unhidden and Almost Unhidden Processes

PIB algorithms that cluster pasts of length $M \geq 1$ to retain information about futures of length $N \geq 1$ calculate accurate information functions when $E(M, N) \approx E$. (Recall Sec. III.) Such algorithms work exactly on order-R Markov processes when $M, N \geq R$, since $E(R, R) = E$. However, there are many processes that are “almost” order-R Markov, for which algorithms based on finite-length pasts and futures should work quite well.

The quality of a process’s approximation can be monitored by the convergence error $E - E(M, N)$, which is controlled by the elusive information $\sigma_{\mu}(\ell)$ [29]. To see this, we apply the mutual information chain rule repeatedly:

$$E = I[X_0; X_0]$$

$$= I[X_0; X_{0:N-1}] + \sigma_{\mu}(N)$$

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

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

$$I[X_{-M-1}; X_{0:N-1}|X_{-M-1}]$$

$$\leq I[X_{-M-1}; X_{0}|X_{-M-1}]$$

$$= \sigma_{\mu}(M) ,$$

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

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

The inequality of Eq. [12] suggests that, as far as accu-
racy is concerned, if a process has a small $\sigma_\mu(L)$ relative
to its $E$ 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.

![Diagram](attachment:image.png)

FIG. 5. (top) Golden Mean HMM, an $\epsilon$-machine. (bottom)
Simple Nonunifilar Source nonunifilar HMM presentation; not
the SNS process’s $\epsilon$-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. 5(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 pro-
cess with small $\sigma_\mu$. One such process is the Simple
Nonunifilar Source (SNS) whose (nonunifilar) HMM is
shown in Fig. 5(bottom). As anticipated, Fig. 6(top)
shows that OCF with $L = 1$ and CIB yield very simi-
lar information functions at low code rate and low $\beta$.
In fact, many of SNS’s statistics are well approximated by
the Golden Mean HMM.

The feature curve in Fig. 6(bottom) reveals a slightly
more nuanced story, however. The SNS is highly cryptic,
in that it has a much larger $C_\mu$ than $E$. As a result,
OCF with $L = 1$ approximates $E$ quite well but under-
estimates $C_\mu$, replacing an (infinite) number of feature-
discover transitions with a single transition. (More on
these transitions shortly.)

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 [3] of
how difficult it can be to exactly calculate information
measures of stochastic processes.

Be aware that CIB cannot be directly applied to an-
alyze the SNS, since the latter’s causal state space is
countably infinite; see Ref. [54]’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. [54]. Here, we used $M, N = 10$, ef-
fectively 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 of phase transitions in constrained optimization [13, 52]. Using the infinite-order Markov Even Process of Sec. [11] 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. 7(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. 7(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 $\beta$s.

The first result is notable, as Ref. [8] 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. [30]) 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 $\Pr(S^+) = (2/3 \ 1/3)$ and the conditional distribution of forward and reverse-time causal states is:

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

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

$$I[R; S^+] = I[R; S^-].$$  \hspace{1cm} (13)

And so, we directly calculate the information function from Eq. (5):

$$R(I_0) = \min_{I[R; S^-] \geq I_0} I[R; S^+]$$

$$= \min_{I[R; S^-] \geq I_0} I[R; S^-]$$

$$= I_0,$$

for all $I_0 \leq E$. Similar arguments hold for periodic process as described in Ref. [14] 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 [30] are key to the shape of information functions.

The Even Process’s feature curve in Fig. 7(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. (9) and (11) is an attempt to maximize the objective function of Eq. (8). However, Eq. (13) gives:

$$L_\beta = (1 - \beta^{-1}) I[R; S^+] .$$

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

Both causal states in the Even Process are unusualy 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 PRD 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, 55]. However, we suspect that diagonal block-matrix forms for $\Pr(S^+, S^-)$
FIG. 7. 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.

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 \( \mathcal{F} = \{01^{2k+1} : k = 0, 1, 2, \ldots \} \). 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 \( \Pr(S^+, S^-) \) 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. \[52\] suggest the contrary. The joint probability distribution of compressed and relevant variables there 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. \[52\]. 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 \[55\].

C. Temporal Asymmetry in Lossy Prediction

As Refs. \[28, 30\] 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. 8. Its bidirectional machine is known analytically \[28\]. Therefore, we know the joint \( \Pr(S^+, S^-) \) via \( \Pr(S^+) = \left( \begin{array}{c} 2/5 \\ 1/5 \end{array} \right) \). And:

\[
\Pr(S^-|S^+) = \left( \begin{array}{cccc} 0 & 1/2 & 0 & 1/2 \\ 0 & 1/2 & 1/2 & 0 \\ 1 & 0 & 0 & 0 \end{array} \right).
\]

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 $E \approx 1.24$ 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 9 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 $E$ by about 30% and both underestimates and overestimates the correct $C_\mu$. 

The RIP feature curves in Fig. 10 reveal a similar story in that OCF fails to asymptote to the correct $C_\mu$ 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 $\beta$) 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 $\epsilon$-machine. See Fig. 2 of Ref. [28] for more details.) This common bidirectional causal state is the main source of similarity in the information functions of Fig. 9.

Both feature curves also show phase transitions at $\beta = 2$, but similarities end there. The forward-time feature curve gives a more accurate estimate of $C_\mu$ due to the increased number of causal states and the longer temporal correlations. OCF, on the other hand, remains significantly off, illustrating its limitations in capturing the full complexity of the RIP process.
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 $a = \left(\sqrt[3]{9} + \sqrt[3]{57} + \sqrt[3]{9} - \sqrt[3]{57}\right)/\sqrt[3]{9}$, studied in Ref. [10]. Figure 11 gives both the Tent Map and the analytically derived $\epsilon$-machine for its symbolic dynamics, from there. The latter reveals that the symbolic dynamic process is infinite-order Markov. The bidirectional $\epsilon$-machine at this parameter setting is also known. Hence, one can directly calculate information functions as described in Sec. V.

From Fig. 12’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. 12(top) is fairly smooth, the feature curve (Fig. 12(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. 13 shows the probability distribution $\Pr(S^+|R)$ 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 predictive 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 $H[X_0|S^+ = s^+]$. 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. 11(bottom), in contrast to how CIB orders them.

VI. COMPUTATIONAL CHALLENGES

Proposition I and Thm. I says that one can calculate PRD functions (including PIB functions) in three steps:

1. Empirically estimate an $\epsilon$-machine [7,11,33,57,60]...
or theoretically derive a system’s $\epsilon$-machine using Eq. (3):

2. Calculate the joint probability distribution over forward- and reverse-time causal states and:

3. Apply a rate-distortion algorithm to compress $\mathcal{S}^+$ to minimize expected distortion about $\mathcal{S}^-$ for a desired distortion.

From the computational complexity viewpoint each step is hard.

Recall that NP refers to the class of decision problems for which a deterministic Turing machine can verify a “yes” answer in polynomial-time; an NP-hard problem is one that is as hard as the hardest problems in NP. Such problems are quite familiar. Many computations related to spin glass Ising models are NP-hard. Inferring an $\epsilon$-machine from finite data is NP-hard, though finite $\epsilon$-machines are PAC-learnable. Calculating $\Pr(\mathcal{S}^+, \mathcal{S}^-)$ from an $\epsilon$-machine typically has run time and storage requirements exponential in $|\mathcal{S}^+|$. Solving the associated PRD optimization is known to be NP-hard.

It may seem as though the three-step approach is overly complicated, especially when compared to the more common approach in which one directly clusters finite-length pasts to retain information about finite-length futures. In point of fact, algorithms based on clustering pasts and futures of length $L$ themselves cannot avoid the basic complexities, either. Rather, they implic-
the $\epsilon$-verse temperature

**FIG. 13.** Tent Map predictive features as a function of inverse temperature $\beta$: Each state-transition diagram shows the $\epsilon$-machine in Fig. 11(bottom) with nodes gray-scaled by $\Pr(S^+|R = r)$ for each $r \in R$. White denotes high probability and black low. Transitions are shown only to guide the eye. (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) $\beta \rightarrow \infty$: original four states identified, $A$, $B$, $C$, and $D$.

We introduced a new relationship between predictive rate-distortion and causal states. Theorem 1 of Refs. [7, 8] say that the predictive information bottleneck can identify forward-time causal states, in theory. The analyses and results in Secs. [III] suggest that in practice, when studying time series with longer-range temporal correlations, we calculate substantially more accurate predictive information functions by deriving or inferring an $\epsilon$-machine first.

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 Sec. [III] demonstrated, a process need not have very long-ranged temporal correlations for the curse of dimensionality to rear its head. These lessons echo that found when analyzing a process’s large deviations [67]. 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.

There may be applications that only need temporally local predictive feature extraction—e.g., that in Ref. [85]—and, then, by assumption there is no such curse of dimensionality. Even then, one often assumes that temporally local learning rules will yield temporally long-ranged predictive features. In this situation, causal rate distortion could provide benchmarks for testing the performance of temporally local predictive feature extractors on infinite-order Markov processes.

Section [VIA] showed that building a model may be unnecessary if the underlying process effectively has small, finite Markov order, since sequences are adequate proxies for a process’s effective states. Sections [VIB, VID] however, then showed that when the underlying process has relatively long-range temporal correlations—either larger Markov order than sequence lengths used or infinite Markov order—then computing PRD functions directly from sequence distributions produces quantitatively inaccurate and structurally misleading results. Since an exhaustive survey [68] shows that infinite Markov order dominates the space of processes generated by even finite HMMs, these problems are likely generic and could very well have affected a number of existing calculations of predictive rate-distortion functions.

That said, using approximate $\epsilon$-machines for CRD or CIB can only yield approximate lossy predictive features. For instance, we approximated the SNS in Sec. [VIA] by a 10-state unifilar HMM, even though the SNS technically has an infinite-state $\epsilon$-machine. This approximation in model space leads to incorrect information functions at very low expected distortions. Future research should 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 and CRD when only approximate $\epsilon$-machines are known.

The relationship between predictive rate-distortion and causal states runs much deeper than we probed here.
At a minimum, CRD is a useful theoretical tool for analyzing coarse-grained features. One technical aspect appeared in our discussion of how the bidirectional machine’s switching maps determine much of an information function’s shape. Perhaps more importantly, though, the predictive features identified using these methods are statistics and not full machines with both a state space and dynamic \cite{34}. (Although one can certainly build lossy machines from these predictive features they are likely to be nonunifilar and not even generate the given process.) Inferring lossy predictive machines requires a different approach, such as the recursive information bottleneck \cite{35}. Success in developing this should prove useful in sensorimotor loop applications, where one learns predictive models actively rather than passively \cite{34,35,70,72}.

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, despite the restrictions that Prop. 1 places on the distortion measures for which CRD and PRD are equivalent, App. B suggests that CRD can be used instead of PRD if one correctly modifies the distortion measure. This opens up a wider range of applications; for example, those in which other properties, besides structure or prediction, are desired \cite{35}, including utility function optimization.

Information rate-distortion functions have already been used to test whether or not the retina predicts its inputs \cite{73}. Another immediate application—the construction of a predictive hierarchy—was suggested in Sec. V.D using the stochastic process defined by symbolic dynamics of a chaotic dynamical system. We saw that rate-distortion analysis provided a principled way to determine which temporal or spatial structures are emergent. In other words, CIB is a new tool for accurately identifying emergent macrostates of a stochastic process \cite{3}.

ACKNOWLEDGMENTS

The authors thank C. Ellison, C. Hillar, W. Bialek, I. Nemenman, and S. Still for helpful discussions and the Santa Fe Institute for its hospitality during visits. JPC is an SFI External Faculty member. 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.

Appendix A: Review of rate-distortion theory

Figure 2 (top) shows the rate-distortion theory setting of Ref. 19 that combines the encoder and noisy channel used in Shannon’s communication channel model \cite{5,6} into a single encoder. In this framing, an information source generates \( n \) successive symbols, a word \( x_{0:n} \in A^n \). The word is presented to the encoder, which outputs \( r_i = 1, \ldots, M \). The collection \( C = \{ (x_{0:n}, r_i) : i = 1, \ldots, M \} \), mapping words to codewords, is the channel codebook. Since sending one of \( M \) codewords requires \( \log_2 M \) bits, the code rate \( R(C) = n^{-1} \log M \) is the number of communicated bits per source symbol.

Given a codeword \( r_i \), the decoder makes its best estimate \( \hat{x}_{0:n} \in A^n \) as to the original source word \( x_{0:n} \). Each estimate is evaluated using a distortion measure \( d(\hat{x}_{0:n}, x_{0:n}) \geq 0 \), that quantifies the error between the
given codeword \( \hat{x}_{0:n} \) and the true word \( x_{0:n} \). Over long periods, the estimates lead to an expected distortion rate:

\[
\mathbb{E} d(x, \hat{x}) = \lim_{n \to \infty} \frac{\mathbb{E} d(x_{0:n}, \hat{x}_{0:n})}{n},
\]

that depends on the codebook \( \mathcal{C} \) and its associated code rate \( R(\mathcal{C}) \). We assume that the distortion measure is normal:

\[ (\hat{x}_{0:n}) \quad d(x_{0:n}, \hat{x}_{0:n}) = 0. \]

There is a natural trade-off between the code rate and the expected distortion: the smaller the desired expected distortion, the larger the required code rate to achieve it. To achieve no expected distortion, \( \mathbb{E} d(x, \hat{x}) = 0 \), requires some minimal code rate \( R_{\text{max}} \), which depends on the distortion measure. For instance, when reconstructing a given time series as well as possible, \( R_{\text{max}} \) is the process’s entropy rate \( h_\mu \) \cite{2}. When the information source consists of successive semi-infinite pasts \( X \) of a time series, then for many prediction-related distortion measures, \( R_{\text{max}} \) is the forward-time statistical complexity \( C_\mu^+ \) \cite{2,3,8}. More generally, when the distortion measure is an informational distortion of a source \( X \) with respect to some “relevant variable” \( Y \) \cite{49,77}, then \( R_{\text{max}} \) is the entropy of the inherited probability distribution of the minimal sufficient statistics of \( X \) with respect to \( Y \) \cite{18}.

A channel’s capacity \( \mathcal{C} \) is the largest information transmission rate it can sustain over all possible information sources \( X \):

\[
\mathcal{C} = \sup_{\mathcal{P}_X} I[X; Y],
\]

where \( Y \) is the channel’s output process. If the encoder’s capacity is large enough—\( \mathcal{C} \geq R_{\text{max}} \)—then there exists a codebook such that the decoder can reconstruct the input sequence with arbitrarily small error probability. If the encoder’s capacity is not large enough—\( \mathcal{C} < R_{\text{max}} \)—however, it cannot. There is an irreducible positive error rate \cite{19}. As the Introduction noted, for PRD applications one is in this regime when \( C_\mu^+ \) is infinite \cite{3,4}.

In this positive error-rate regime, we ask for the rate-distortion function:

\[
R(D) = \inf_{\mathbb{E} d(x, \hat{x}) \leq D} R(\mathcal{C}). \tag{A1}
\]

For simplicity, we typically limit ourselves to single-symbol distortion measures, so that the distortion between decoded and input word is:

\[
d(x_{0:n}, \hat{x}_{0:n}) = \sum_{i=0}^{n-1} d(x_i, \hat{x}_i).
\]

Equation (A1) is a difficult optimization, given that to determine \( R(D) \) for each \( D \) requires enumerating a combinatorially large space of codebooks. Instead, one views the information source as a random variable \( X \) with realizations \( x \) and the (potentially stochastic) codebook’s output as a random variable \( \mathcal{R} = \mathcal{C}(X) \) with realizations \( r \in \{1, \ldots, M\} \).

Then, according to the Rate-Distortion Theorem \cite{19}, one has:

\[
R(D) = \min_{I[X; \mathcal{R}] \leq D} I[X; \mathcal{R}]\ .
\]

One can solve numerically for \( R(D) \) using annealing to find a \( \mathcal{P}(\mathcal{R}|X) \) that minimizes the objective function:

\[
\mathcal{L}_\beta = I[X; \mathcal{R}] + \beta \langle d(x, r) \rangle_{X, \mathcal{R}}.
\]

As \( \beta \) varies, one obtains different rates \( R_\beta = I[X; \mathcal{R}] \) and expected distortions \( D_\beta = \langle d(x, r) \rangle_{X, \mathcal{R}} \). In this way, one traces out the rate-distortion function \( R(D) \) parametrically.

Note that, in the preceding, \( X \) was not a measurement symbol of a stochastic process, as it was in Sec. II B. In the present context, \( X \) is the random variable denoting the output of any information source, which could be, but need not be, a string of symbols from a stochastic process.

A natural question arises, which distortion measure should one use? Historically, it was chosen to be a more or less familiar statistic, such as the mean-squared error and the like. More recently, though, information measures have been used, mirroring the definition of the code rate in the objective function. For example, Ref. \cite{49} posited that one should retain some aspect \( \mathcal{R} \) of the input \( X \) that is related to another “relevant” random variable \( Y \). This is tantamount to assuming a Markov chain relationship: \( \mathcal{R} \to X \to Y \). A natural distortion measure then is \cite{79}:

\[
d(x, r) = D_{KL}[\mathcal{P}(Y|X = x)||\mathcal{P}(Y|\mathcal{R} = r)],
\]

where \( D_{KL}(P||Q) \) is the relative entropy between distributions \( P \) and \( Q \). Though, one might consider others based on the distance between the conditional probability distributions \( \mathcal{P}(Y|X = x) \) and \( \mathcal{P}(Y|\mathcal{R} = r) \). It is straightforward to show that:

\[
\langle D_{KL}[\mathcal{P}(Y|X = x)||\mathcal{P}(Y|\mathcal{R} = r)] \rangle = I[X; Y|\mathcal{R}],
\]

since the Markov chain gives \( \mathcal{P}(Y|X, \mathcal{R}) = \mathcal{P}(Y|X) \). And, since \( I[X; Y|\mathcal{R}] = I[X; Y] - I[\mathcal{R}; Y] \) due to the same Markov chain assumption, we define the informa-
tion function:

\[ R(I_0) = \min_{I[R;Y] \geq I_0} I[X;R] \]

with objection function:

\[ \mathcal{L}_\beta = I[R;Y] - \beta^{-1} I[X;R] \]

Finding the \( \Pr(R|X) \) that maximize \( \mathcal{L}_\beta \) is the basis for the information bottleneck (IB) method \[39,77].

Appendix B: Causal Rate-Distortion Proofs

We assume familiarity with rate-distortion theory on the level of Refs. \[19\] Ch. 8 and \[32\].

We designed the proofs to be as elementary as possible; they are basically repeated applications of the Markov loop in Fig. [1] and the pullback method from probability theory. They are also, in effect, detailed sketches more than they are proofs. This allows us to highlight their constructive nature. All statements can almost be proven through repeated applications of the Markov loop in Fig. [1] to the formal solution for optimal stochastic codebooks given by Theorem 4 of Ref. \[49\]. The “almost” comes from the fact that not all solutions to the annealing objective functions associated with PIB and CIB are the same. Rather, it only finds equivalence between solutions to the annealing objectives of PIB and CIB are the same. Furthermore, a given causal-state codebook \( \mathcal{R} \) implies a causal-state codebook \( \mathcal{R}_{\hat{\epsilon}} \) with realizations \( r_{\hat{\epsilon}} \in \epsilon^+(\mathbf{X}) \):

\[ \Pr(\mathcal{R}_{\hat{\epsilon}}|S^+ = \sigma^+) = \sum_{\mathbf{x} \in \mathbf{X}} \Pr(\mathcal{R}_{\hat{\epsilon}}|\mathbf{x} = \mathcal{X}|S^+ = \sigma^+) \Pr(\mathbf{x}|S^+ = \sigma^+) = \sum_{\mathbf{x} : \epsilon^+\left(\mathbf{x}\right) = \sigma^+} \Pr(\mathcal{R}_{\hat{\epsilon}}|\mathbf{x} = \mathcal{X}) \Pr(\mathbf{x} = \mathcal{X}) . \]

This is essentially the pullback method of probability theory. Similarly, a given causal-state codebook \( \mathcal{R} \) can be naturally extended to a process codebook \( \mathcal{R}_{\mathcal{X}} \) with realizations \( \mathcal{X} \in \mathbf{X} \):

\[ \Pr\left(\mathcal{R}_{\mathcal{X}}|\mathbf{X} = \mathcal{X}\right) = \Pr\left(\mathcal{R}|S^+ = \epsilon^+(\mathcal{X})\right) . \]

In this way, we compare \( \mathcal{R} \) to \( \mathcal{R}_{\mathcal{X}} \), using \( \mathcal{R}_{\hat{\epsilon}} \). To simplify these comparisons, if \( \mathcal{Q} \) is some codebook, in the following \( \mathcal{Q}_{\hat{\epsilon}} \) and \( \mathcal{Q}_{\hat{\epsilon}} \) denote a codebook over causal states and \( \mathcal{Q}_{\mathcal{X}} \) a codebook over process pasts.

Here, we only consider distortion measures of the form:

\[ d(\mathcal{X}, r) = f\left(\Pr(\mathcal{X}|\mathbf{X} = \mathcal{X}), \Pr(\mathbf{X}|\mathcal{R} = r)\right) \]

and

\[ \hat{d}(\sigma^+, \hat{\epsilon}) = f\left(\Pr\left(\mathcal{X}|S^+ = \sigma^+\right), \Pr\left(\mathcal{X}|\mathcal{R} = \hat{\mathcal{X}}\right)\right) , \]

for some unspecified \( f(\cdot, \cdot) \) that quantifies differences between probability distributions; e.g., Kullback-Liebler or Shannon-Jensen divergence. Causal shielding implies that:

\[ \Pr(\mathcal{X}|\mathbf{X} = \mathcal{X}) = \Pr(\mathcal{X}|S^+ = \epsilon^+(\mathcal{X})) \]

and:

\[ \Pr\left(\mathcal{X}|\mathcal{R}\right) = \sum_{\sigma^+ \in S^+} \Pr\left(\mathcal{X}|S^+ = \sigma^+\right) \Pr(S^+ = \sigma^+|\mathcal{R}) = \sum_{\sigma^+ \in S^+} \Pr\left(\mathcal{X}|S^+ = \sigma^+\right) \Pr(S^+ = \sigma^+|\mathcal{R}_{\hat{\epsilon}}) . \]

By construction, then:

\[ d(\mathcal{X}, r) = \hat{d}(\sigma^+, \hat{\epsilon}(\mathcal{X}), r_{\hat{\epsilon}}) \]

and

\[ \hat{d}(\sigma^+, \hat{\epsilon}) = d(\mathcal{X}, \hat{\mathcal{R}}) , \]

Using the fact that the two source sample spaces and codebooks are intimately related via the causal-state map \( \sigma^+ = \epsilon^+(\mathbf{X}) \), we construct codebooks in one sample space out of codebooks from the other. For instance, a given process codebook \( \mathcal{R} \) implies a causal-state codebook \( \mathcal{R}_{\hat{\epsilon}} \) with realizations \( r_{\hat{\epsilon}} \in \epsilon^+(\mathbf{X}) \):

\[ \Pr(\mathcal{R}_{\hat{\epsilon}}|S^+ = \sigma^+) \]

\[ = \sum_{\mathbf{x} \in \mathbf{X}} \Pr(\mathcal{R}_{\hat{\epsilon}}|\mathbf{x} = \mathcal{X}|S^+ = \sigma^+) \Pr(\mathbf{x}|S^+ = \sigma^+) \]

\[ = \sum_{\mathbf{x} : \epsilon^+\left(\mathbf{x}\right) = \sigma^+} \Pr(\mathcal{R}_{\hat{\epsilon}}|\mathbf{x} = \mathcal{X}) \Pr(\mathbf{x} = \mathcal{X}) . \]
for all $\sigma$ such that $\epsilon^+(\sigma) = \sigma^+$. On the one hand, when the information source is $\hat{X}$, we consider the following rate-distortion function:

$$R(D) = \min_{\langle d(\sigma, r) \rangle_{\bar{R}, \sigma^+ \leq D}} I[\hat{R}; \hat{X}]$$

and its associated optimal codebook:

$$\bar{R}^*_D = \arg\min I[\bar{R}; \hat{X}].$$

On the other hand, when the information source is $S^+$, we consider the following rate-distortion function:

$$\tilde{R}(D) = \min_{\langle \tilde{d}(\sigma, r) \rangle_{\bar{R}, \sigma^+ \leq D}} I[\tilde{R}; S^+]$$

and its associated optimal codebook:

$$\tilde{R}^*_D = \arg\min I[\tilde{R}; S^+].$$ (B1)

The lemma below, a more precise version of the main text’s Lemma 1, states their relationship.

**Lemma 1.** $R(D) = \tilde{R}(D)$ and $\bar{R}^*_D \simeq \tilde{R}^*_D$ for all achievable $D$.

**Proof.** First, to establish the isomorphism we construct a map between the process and causal-state codebooks, showing that $\bar{R}^*_D = ((\bar{R}^*_+)_D)_{\hat{X}}$, via a proof by contradiction. Suppose they are not equal. As described earlier:

$$D = \langle d(\sigma, r) \rangle_{\bar{R}^*_D, \hat{X}}$$

$$= \langle \tilde{d}(\sigma, r) \rangle_{\bar{R}^*_+}, S^+$$

$$= \langle \tilde{d}(\sigma, r) \rangle_{\tilde{R}^*_D, S^+}.$$

So, the two codebook random variables have the same expected distortion. Since $\bar{R}^*_D \neq ((\bar{R}^*_+)_D)_{\hat{X}}$, $\Pr(\bar{R}^*_D, X | S^+) \neq \Pr(\bar{R}^*_+ S^+) \Pr(X | S^+)$ and thus $I[\bar{R}^*_D; X | S^+] > 0$. This implies that $((\bar{R}^*_+))_{\hat{X}}$ has a lower coding cost:

$$I[\bar{R}^*_D; \hat{X}] = I[\bar{R}^*_D; S^+] + I[\bar{R}^*_D; \hat{X} | S^+]$$

$$> I[\bar{R}^*_D; S^+]$$

$$= I[\bar{R}^*_+; S^+]$$

$$= I[(\bar{R}^*_+)_D; S^+]$$

Since $((\bar{R}^*_+)_D)_{\hat{X}}$ is a codebook with the same expected distortion as $\bar{R}^*_D$ and lower coding cost, our assumption that they are not equal is incorrect and so $\bar{R}^*_D = ((\bar{R}^*_+)_D)_{\hat{X}}$.

Second, to show that $\hat{R}^*_D = \bar{R}^*_D$, we only need to show that $\hat{R}^*_D = \bar{R}^*_D$ since the extension from codebooks $\hat{R}$ to codebooks $\bar{R}$ is injective. The reduced codebook $(\bar{R}^*_+)_D$ inherits a constraint:

$$\langle d(\hat{X}, r) \rangle_{\bar{R}^*_D, \hat{X}} = \langle \tilde{d}(\sigma^+, r) \rangle_{\bar{R}^*_+}; S^+ \leq D$$

and must minimize coding cost $I[\bar{R}^*_D; \hat{X}] = I[(\bar{R}^*_+)_D; S^+]$, as described above. In short, $(\bar{R}^*_+)_D$ satisfies:

$$\bar{R}^*_D = \arg\min I[(\bar{R}^*_+)_D; S^+].$$ (B2)

Comparing Eq. (B1) to Eq. (B2) yields:

$$\tilde{R}^*_D = \bar{R}^*_D$$

$$\hat{R}^*_D = \bar{R}^*_D,$$

as desired. Using earlier manipulations, we also see that:

$$R(D) = I[\bar{R}^*_D; \hat{X}]$$

$$= I[(\bar{R}^*_+)_D; S^+]$$

$$= I[\bar{R}^*_D; S^+]$$

$$= \tilde{R}(D).$$

completing the proof sketch.

Next, consider an alternate class of causal-state codebooks $\tilde{R}$ and realizations $\tilde{r}$ for coding information sources $S^+$. These are distinguished by a new distortion measure:

$$\tilde{d}(\sigma^+, \tilde{r}) = f\left(\Pr(S^- | S^+ = \sigma^+), \Pr(S^- | \tilde{R} = \tilde{r})\right).$$

The associated rate-distortion function is:

$$\tilde{R}(D) = \min_{\langle \tilde{d}(\sigma, \tilde{r}) \rangle_{\bar{R}, \sigma^+ \leq D}} I[\tilde{R}; S^+]$$

with optimal codebooks:

$$\tilde{R}^*_D = \arg\min I[\tilde{R}; S^+].$$ (B3)

We would like to know the relationship between $R(D)$ and $\tilde{R}(D)$, among other things. The proposition below, a more precise version of the main text’s Proposition 1, states their relationship.

**Proposition 1.** $R(D) = \tilde{R}(D)$ and $\tilde{R}^*_D = \bar{R}^*_D$ for all achievable $D$ if:

$$f\left(\Pr(\tilde{X} | S^+ = \sigma^+), \Pr(\tilde{X} | \tilde{R} = \tilde{r})\right)$$

$$= f\left(\Pr(S^- | S^+ = \sigma^+), \Pr(S^- | \tilde{R} = \tilde{r})\right).$$
Lemma 1 above.

Finally, we concentrate on a particularly useful distortion measure that satisfies the above constraint. Then \(\hat{R}^*_{I_0} = \hat{R}^*_{D} \) and \(\hat{R}(D) = \hat{R}(D)\). The conclusion follows from Lemma 1 above.

\[
d(\bar{x}, r) = \sum_{\bar{x}} (\Pr(\bar{X} = \bar{x} | X = \bar{x}) - \Pr(\bar{X} = \bar{x} | R = r))^2
\]

\[
= \sum_{\sigma} \left( \sum_{\bar{x} \in \epsilon^{-}\sigma} \Pr(\bar{X} = \bar{x} | S^- = \sigma^-) \right)^2 \left( \Pr(S^- = \sigma^- | X = \bar{x}) - \Pr(S^- = \sigma^- | R = r) \right)^2.
\]

This weighting factor typically vanishes if the futures are semi-infinite, since:

\[
\sum_{\bar{x} \in \epsilon^{-}\sigma^-} \Pr(\bar{X} = \bar{x} | S^- = \sigma^-) = 1
\]

and \(0 \leq \Pr(\bar{X} = \bar{x} | S^- = \sigma^-)\). This issue is easily addressed by rescaling the weighting factor for finite-length futures by an appropriate function of the length and then taking the limit of this rescaled weighting factor as the length tends to infinity.

So, again, distortion measures that do not satisfy Prop. 1’s conditions implicitly privilege one reverse-time causal state over another.

Finally, we concentrate on a particularly useful distortion measure that satisfies the above constraint. OCI compresses \(\bar{X}\) to retain information about \(\bar{X}\), resulting in the information function:

\[
R(I_0) = \min_{I|R;X \geq I_0} I[\bar{X} ; R]
\]

and the associated representation is:

\[
\hat{R}^*_{I_0} = \arg\min_{I|R;\bar{X} \geq I_0} I[\bar{X} ; R].
\]

There are many distortion measures that do not satisfy the constraints in Prop. 1. As an example, we focus on mean squared error and still can use the Markov chains \(\mathcal{R} \rightarrow \mathcal{S}^- \rightarrow \bar{X}\) and \(\bar{X} \rightarrow \mathcal{S}^- \rightarrow \bar{X}\) to some benefit:

\[
d(\bar{x}, r) = D_{KL}[\Pr(\bar{X} | X = \bar{x}) || \Pr(\bar{X} | R = r)] = \sum_{\sigma^-} \Pr(\bar{X} = \bar{x} | S^- = \sigma^- , \bar{X} = \bar{x}) \Pr(S^- = \sigma^- | X = \bar{x})
\]

\[
= \Pr(\bar{X} = \bar{x} | S^- = \epsilon^-(\bar{x})) \Pr(S^- = \epsilon^-(\bar{x}) | \bar{X} = \bar{x}).
\]
Similarly, since $\mathcal{R} \rightarrow S^- \rightarrow \overrightarrow{X}$:

$$\Pr(\overrightarrow{X} = \overrightarrow{x} | \mathcal{R} = r) = \Pr(\overrightarrow{X} = \overrightarrow{x} | S^- = \epsilon^{-}(\overrightarrow{x})) \Pr(S^- = \epsilon^{-}(\overrightarrow{x}) | \mathcal{R} = r) .$$

The $D_{KL}$ distortion measure satisfies the conditions of Prop. 1:

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

$$= \sum_{\overrightarrow{x}} \Pr(\overrightarrow{X} = \overrightarrow{x} | \overrightarrow{X} = \overrightarrow{x}) \log \frac{\Pr(\overrightarrow{X} = \overrightarrow{x} | \overrightarrow{X} = \overrightarrow{x})}{\Pr(\overrightarrow{X} = \overrightarrow{x} | \mathcal{R} = r)}$$

$$= \sum_{\overrightarrow{x}} \Pr(\overrightarrow{X} = \overrightarrow{x} | S^- = \epsilon^{-}(\overrightarrow{x})) \Pr(S^- = \epsilon^{-}(\overrightarrow{x}) | \overrightarrow{X} = \overrightarrow{x}) \log \frac{\Pr(S^- = \epsilon^{-}(\overrightarrow{x}) | \overrightarrow{X} = \overrightarrow{x})}{\Pr(S^- = \epsilon^{-}(\overrightarrow{x}) | \mathcal{R} = r)}$$

$$= \sum_{\overrightarrow{x}} \Pr(S^- = \sigma^{-}|\overrightarrow{X} = \overrightarrow{x}) \log \frac{\Pr(S^- = \sigma^{-}|\overrightarrow{X} = \overrightarrow{x})}{\Pr(S^- = \sigma^{-}|\mathcal{R} = r)}$$

$$= D_{KL}[\Pr(S^- | \overrightarrow{X} = \overrightarrow{x}) || \Pr(S^- | \mathcal{R} = r)] .$$

Thus, from Prop. 1, $(\overrightarrow{R}_D)_{\overrightarrow{x}} = \mathcal{R}_D^*$ for each achievable $D$. However, and it is straightforward to show, the expected distortion for this measure is $I[\overrightarrow{X}; \mathcal{R}]$. Hence, upper bounding the expected distortion $\langle d(\overrightarrow{x}, r) \rangle = I[\overrightarrow{X}; \mathcal{R}] \leq D$ is equivalent to lower bounding $I[\overrightarrow{X}; \mathcal{R}] = I[\overrightarrow{X}; \mathcal{R}] - D$; cf. Ref. [27]. Thus, $(\overrightarrow{R}_I)_{\overrightarrow{x}} = \mathcal{R}_I^*$ for each achievable $I_0$ and the information functions are equivalent:

$$R(I_0) = I[\overrightarrow{X}; \mathcal{R}_I^*]$$

$$= I[\overrightarrow{X}; (\overrightarrow{R}_I)_{\overrightarrow{x}}]$$

$$= I[S^+; \overrightarrow{R}_I^*]$$

$$= \hat{R}(I_0) .$$

Note that there are other $f(\cdot, \cdot)$ satisfying Prop. 1’s conditions, including:

- $f(q_1(Y), q_2(Y)) = \sum_y |q_1(y) - q_2(y)|$ and

- $f(q_1(Y), q_2(Y)) = D_{KL}[\omega_1 q_1(Y) + (1 - \omega_1) q_2(Y) || \omega_2 q_1(Y) + (1 - \omega_2) q_2(Y)]$ for any $0 \leq \omega_1, \omega_2 \leq 1$.

Again, for other distortion measures, we might find that they can be expressed as a weighted average distortion over reverse-time causal states.

Let’s close with a caveat on notation. Throughout, we cavalierly manipulated semi-infinite pasts and futures and their conditional and joint probability distributions—e.g., $\Pr(\overrightarrow{X} | \overrightarrow{X})$. 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 Prop. 1 and Thm. 1 for this objective function and a version of CIB with finite-time forward and reverse-time causal states, and (iii) take the limit as those lengths go to infinity; e.g., as in Ref. 17. 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 Prop. 1 and Thm. 1. In the service of clarity, such care was abandoned to a shorthand, leaving the task of an expanded measure-theoretic development to elsewhere.

Appendix C: Optimizing the CIB Objective Function

To make CRD’s development self-contained, we explain how to derive the formal solutions and algorithms of Eqs. (9)-(11) from the objective function Eq. (8). What follows is essentially a short review of Thm. 4 of Ref. 49 when the variable to compress is $S^+$ and the relevant variable is $S^-$. Theorem 1 identifies codebooks that minimize coding rate $I[\mathcal{R}; S^+]$ given a lower bound on predictive power.
Unsurprisingly, optimal codebooks saturate the given bounds on predictive power [10]. Then, the method of Lagrange multipliers implies that optimal codebooks maximize the annealing function:

\[ \mathcal{L}_\beta = I[R; S^+ | R] - \beta^{-1} I[S^+; R] \]

\[ - \sum_{\sigma^+} \mu(\sigma^+) \sum_r \Pr(R = r| S^+ = \sigma^+) , \quad (C1) \]

where \( \beta \) and \( \mu(\sigma^+) \) are Lagrange multipliers. The Lagrange multiplier \( \beta \) controls the trade-off between coding cost and predictive power, while the Lagrange multipliers \( \mu(\sigma^+) \) enforce normalization constraints:

\[ \sum_r \Pr(R = r| S^+ = \sigma^+) = 1 . \]

This is a slightly different objective function than in the main text. The only difference is that in Eq. (8), we search only over \( \Pr(R = r| S^+ = \sigma^+) \) that are normalized. The objective function in Eq. (C1) explicitly enforces this constraint.

Optimal codebooks satisfy \( \partial \mathcal{L}_\beta / \partial \Pr(R|S^+) = 0 \). For notational ease, we use \( p(r|\sigma^+) \) to denote \( \Pr(R = r| S^+ = \sigma^+) \) and so on. We rewrite \( \partial \mathcal{L}_\beta / \partial \Pr(r|\sigma^+) = 0 \) in a more useful way. The first step is to rewrite \( I[R; S^-] = H[R] - H[R|S^-] \) and \( I[S^+; R] = H[S^+] - H[S^+|R] \). Also:

\[ \frac{\partial}{\partial p(r|\sigma^+)} \sum_{\sigma^+} \mu(\sigma^+) \sum_r p(r|\sigma^+) = \mu(\sigma^+) . \]

We combine these simple manipulations to obtain:

\[ 0 = (1 - \beta^{-1}) \frac{\partial H[R]}{\partial p(r|\sigma^+)} - \frac{\partial H[R|S^-]}{\partial p(r|\sigma^+)} + \beta^{-1} \frac{\partial H[R|S^+]}{\partial p(r|\sigma^+)} - \mu(\sigma^+) . \quad (C2) \]

The expression \( \partial H[R|S^+] / \partial p(r|\sigma^+) \) is straightforward to calculate, since:

\[ H[R|S^+] = - \sum_{\sigma^+} p(\sigma^+) \sum_r p(r|\sigma^+) \log p(r|\sigma^+) \]

and \( \partial p(r'|\sigma^+)/ \partial p(r|\sigma^+) = \delta_{r,r'} \):

\[ \frac{\partial H[R|S^+]}{\partial p(r|\sigma^+)} = -p(\sigma^+) (\log p(r|\sigma^+) + 1) . \quad (C3) \]

Next, to determine \( \partial H[R] / \partial p(r|\sigma^+) \), we note that \( p(r) = \sum_{\sigma^+} p(\sigma^+) p(r|\sigma^+) \), so \( \partial p(r) / \partial p(r|\sigma^+) = p(\sigma^+) \). Thus, we have:

\[ \frac{\partial H[R]}{\partial p(r|\sigma^+)} = \frac{\partial H[R]}{\partial p(r)} \frac{\partial p(r)}{\partial p(r|\sigma^+)} = -(\log p(r) + 1)p(\sigma^+) . \quad (C4) \]

Finally, to calculate \( \partial H[R|S^-]/\partial p(r|\sigma^-) \), we recall that \( R \to S^+ \to S^- \) is a Markov chain. Hence:

\[ p(r|\sigma^-) = \sum_{\sigma^+} p(r|\sigma^+) p(\sigma^+|\sigma^-) , \]

implying that:

\[ \frac{\partial p(r|\sigma^-)}{\partial p(r|\sigma^+)} = p(\sigma^+|\sigma^-) . \]

These give:

\[ \frac{\partial H[R|S^-]}{\partial p(r|\sigma^+)} = \sum_{\sigma^-} \frac{\partial H[R|S^-]}{\partial p(r|\sigma^-)} \frac{\partial p(r|\sigma^-)}{\partial p(r|\sigma^+)} \]

\[ = - \sum_{\sigma^-} (1 + \log p(r|\sigma^-)) p(\sigma^+|\sigma^-) \]

\[ = -p(\sigma^+) - p(\sigma^+) \sum_{\sigma^-} p(\sigma^-|\sigma^+) \log p(r|\sigma^-) . \]

Substituting Eqs. (C3), (C5) into Eq. (C2) and dividing through by \( p(\sigma^+) \) yields:

\[ \mu(\sigma^+) = -(1 - \beta^{-1}) \log p(r) + \sum_{\sigma^-} p(\sigma^-|\sigma^+) \log p(r|\sigma^-) \]

\[ - \beta^{-1} \log p(r|\sigma^+) , \]

where \( \mu(\sigma^+) \) replaced \( \mu(\sigma^+)/p(\sigma^+) \). (Since \( \mu(\sigma^+) \) was an unknown constant to start with, this abuse of notation is somewhat justified.) If we recall that:

\[ D_{KL}[p(\sigma^-|\sigma^+)||p(\sigma^-|r)] = \sum_{\sigma^-} p(\sigma^-|\sigma^+) \log \frac{p(\sigma^-|\sigma^+)}{p(\sigma^-|r)} , \]

we can use Bayes’ Rule—\( p(r|\sigma^-) = p(\sigma^-|r)p(r)/p(\sigma^-) \)—and algebra not shown here to further rewrite:

\[ \sum_{\sigma^-} p(\sigma^-|\sigma^+) \log p(r|\sigma^-) = -D_{KL}[p(\sigma^-|\sigma^+)||p(\sigma^-|r)] \]

\[ + \sum_{\sigma^-} p(\sigma^-|\sigma^+) \log \frac{p(\sigma^-|\sigma^+)}{p(\sigma^-)} + \log p(r) . \]

After multiplying through by \( \beta \) and allowing \( \mu(\sigma^+) \) to absorb various unimportant other constants, this implies:

\[ \mu(\sigma^+) = -\log p(r) - \beta D_{KL}[p(\sigma^-|\sigma^+)||p(\sigma^-|r)] - \log p(r|\sigma^+) . \]
Finally, a slight rearrangement gives:

\[ p(r|\sigma^+) = \frac{p(r)}{Z(\sigma^+)} e^{-\beta D_{KL}[p(\sigma^-|\sigma^+)||p(\sigma^-|r)]}, \]  

(C6)

where \( Z(\sigma^+) = e^{-\mu(\sigma^+)} \). Enforcing normalization constraints, \( \sum_r p(r|\sigma^+) = 1 \), means that \( Z(\sigma^+) \) is a partition function. Meanwhile, \( p(r) \) is set implicitly via \( p(r) = \sum_{\sigma^+} p(\sigma^+|r)p(r|\sigma^+) \).

Equation (C6) is a formal solution for the optimal codebook \( p(r|\sigma^+) \). The right-hand side of this formal solution can be viewed as a map on \( p(r|\sigma^+) \), taking one codebook to a new codebook. By iterating this map, we eventually converge to a codebook that satisfies Eq. (C6). Equations (9), (10) then turn this realization into an algorithm. However, there are many suboptimal codebooks for a given \( \beta \) that are also extrema of \( \mathcal{L}_\beta \). So, practically, one is either careful about initial conditions or one starts from a variety of initial conditions, choosing the converged codebook with maximal \( I[R;S^+;R] \) or (preferably) both.

25


