

# Optimal Causal Inference: Estimating Stored Information and Approximating Causal Architecture

Susanne Still,<sup>1,\*</sup> James P. Crutchfield,<sup>2,3,†</sup> and Christopher J. Ellison<sup>2,‡</sup>

<sup>1</sup>*Information and Computer Sciences, University of Hawaii at Manoa, Honolulu, HI 96822*

<sup>2</sup>*Complexity Sciences Center and Physics Department,*

*University of California at Davis, One Shields Avenue, Davis, CA 95616*

<sup>3</sup>*Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501*

(Dated: August 19, 2010)

We introduce an approach to inferring the causal architecture of stochastic dynamical systems that extends rate distortion theory to use causal shielding—a natural principle of learning. We study two distinct cases of causal inference: optimal causal filtering and optimal causal estimation.

Filtering corresponds to the ideal case in which the probability distribution of measurement sequences is known, giving a principled method to approximate a system’s causal structure at a desired level of representation. We show that, in the limit in which a model complexity constraint is relaxed, filtering finds the exact causal architecture of a stochastic dynamical system, known as the *causal-state partition*. From this, one can estimate the amount of historical information the process stores. More generally, causal filtering finds a graded model-complexity hierarchy of approximations to the causal architecture. Abrupt changes in the hierarchy, as a function of approximation, capture distinct scales of structural organization.

For nonideal cases with finite data, we show how the correct number of underlying causal states can be found by optimal causal estimation. A previously derived model complexity control term allows us to correct for the effect of statistical fluctuations in probability estimates and thereby avoid over-fitting.

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

**Natural systems compute intrinsically and produce information. This organization, often only indirectly accessible to an observer, is reflected to varying degrees in measured time series. Nonetheless, this information can be used to build models of varying complexity that capture the causal architecture of the underlying system and allow one to estimate its information processing capabilities. We investigate two cases. The first is when a model builder wishes to find a more compact representation than the true one. This occurs, for example, when one is willing to incur the cost of a small increase in error for a large reduction in model size. The second case concerns the empirical setting in which only a finite amount of data is available. There one wishes to avoid over-fitting a model to a particular data set.**

## I. INTRODUCTION

Time series modeling has a long and important history in science and engineering. Advances in dynamical systems over the last half century led to new methods that attempt to account for the inherent nonlinearity in

many natural phenomena [1–7]. As a result, it is now well known that nonlinear systems produce highly correlated time series that are not adequately modeled under the typical statistical assumptions of linearity, independence, and identical distributions. One consequence, exploited in novel state-space reconstruction methods [8–10], is that discovering the hidden structure of such processes is key to successful modeling and prediction [11–14]. In an attempt to unify the alternative nonlinear modeling approaches, computational mechanics [15] introduced a minimal representation—the  $\epsilon$ -machine—for stochastic dynamical systems that is an optimal predictor and from which many system properties can be directly calculated. Building on the notion of state introduced in Ref. [8], a system’s effective states are those variables that *causally shield* a system’s past from its future—capturing, in the present, information from the past that predicts the future.

Following these lines, here we investigate the problem of learning predictive models of time series with particular attention paid to discovering hidden variables. We do this by using the information bottleneck method (IB) [16] together with a complexity control method discussed by Ref. [17], which is necessary for learning from finite data. Ref. [18] lays out the relationship between computational mechanics and the information bottleneck method. Here, we make the mathematical connection for times series, introducing a new method.

We adapt IB to time series prediction, resulting in a method we call *optimal causal filtering* (OCF) [44]. Since OCF, in effect, extends rate-distortion theory [19] to use

---

\*Electronic address: sstill@hawaii.edu

†Electronic address: chaos@cse.ucdavis.edu

‡Electronic address: cellison@cse.ucdavis.edu

causal shielding, in general it achieves an optimal balance between model complexity and approximation accuracy. The implications of these trade-offs for automated theory building are discussed in Ref. [20].

We show that in the important limit in which prediction is paramount and model complexity is not restricted, OCF reconstructs the underlying process's causal architecture, as previously defined within the framework of computational mechanics [15, 21, 22]. This shows that, in effect, OCF captures a source's hidden variables and organization. The result gives structural meaning to the inferred models. For example, one can calculate fundamental invariants—such as, symmetries, entropy rate, and stored information—of the original system.

To handle finite-data fluctuations, OCF is extended to *optimal causal estimation* (OCE). When probabilities are estimated from finite data, errors due to statistical fluctuations in probability estimates must be taken into account in order to avoid over-fitting. We demonstrate how OCF and OCI work on a number of example stochastic processes with known, nontrivial correlational structure.

## II. CAUSAL STATES

Assume that we are given a stochastic process  $P(\vec{X})$ —a joint distribution over a bi-infinite sequence  $\vec{X} = \overleftarrow{X} \overrightarrow{X}$  of random variables. The *past*, or *history*, is denoted  $\overleftarrow{X} = \dots X_{-3} X_{-2} X_{-1}$ , while  $\overrightarrow{X} = X_0 X_1 X_2 \dots$  denotes the *future* [45]. Here, the random variables  $X_t$  take on discrete values  $x \in \mathcal{A} = \{1, 2, \dots, k\}$  and the process as a whole is stationary. The following assumes the reader is familiar with information theory and the notation of Ref. [23].

Within computational mechanics, a process  $P(\vec{X})$  is viewed as a communication channel that transmits information from the past to the future, storing information in the present—presumably in some internal states, variables, or degrees of freedom [24]. One can ask a simple question, then: how much information does the past share with the future? A related and more demanding question is how we can infer a predictive model, given the process. Many authors have considered such questions. Refs. [18, 22, 25, 26] review some of the related literature.

The effective, or *causal*, states  $\mathcal{S}$  are determined by an equivalence relation  $\overleftarrow{x} \sim \overleftarrow{x}'$  that groups all histories together which give rise to the same prediction of the future [15, 22]. The equivalence relation partitions the space  $\overleftarrow{\mathbf{X}}$  of histories and is specified by the set-valued function:

$$\epsilon(\overleftarrow{x}) = \{\overleftarrow{x}' : P(\overrightarrow{X} | \overleftarrow{x}) = P(\overrightarrow{X} | \overleftarrow{x}')\} \quad (1)$$

that maps from an individual history to the equivalence class  $\sigma \in \mathcal{S}$  containing that history and all others which lead to the same prediction  $P(\overrightarrow{X} | \overleftarrow{x})$  of the future. A

causal state  $\sigma$  includes: (i) a label  $\sigma \in \mathcal{S}$ ; (ii) a set of histories  $\overleftarrow{X}_\sigma = \{\overleftarrow{x} : P(\overrightarrow{X} | \overleftarrow{x}) = P(\overrightarrow{X} | \sigma)\} \subset \overleftarrow{\mathbf{X}}$ ; and (iii) a future conditional distribution  $P(\overrightarrow{X} | \sigma)$  given the state [15, 22].

Any alternative model, called a *rival*  $\mathcal{R}$ , gives a probabilistic assignment  $P(\mathcal{R} | \overleftarrow{x})$  of histories to its states  $\rho \in \mathcal{R}$ . Due to the data processing inequality, a model can never capture more information about the future than shared between past and future:

$$I[\mathcal{R}; \overrightarrow{X}] \leq I[\overleftarrow{X}; \overrightarrow{X}] , \quad (2)$$

where  $I[V, W]$  denotes the mutual information between random variables  $V$  and  $W$  [23]. The quantity  $\mathbf{E} = I[\overleftarrow{X}; \overrightarrow{X}]$  has been studied by several authors and given different names, such as (in chronological order) convergence rate of the conditional entropy [27], excess entropy [28], stored information [29], effective measure complexity [30], past-future mutual information [31], and predictive information [32], amongst others. For a review see Ref. [25] and references therein.

The causal states  $\sigma \in \mathcal{S}$  are distinguished by the fact that the function  $\epsilon(\cdot)$  gives rise to a *deterministic* assignment of histories to states:

$$P(\sigma | \overleftarrow{x}) = \delta_{\sigma, \epsilon(\overleftarrow{x})} , \quad (3)$$

and, furthermore, by the fact that their future conditional probabilities are given by

$$P(\overrightarrow{X} | \sigma) = P(\overrightarrow{X} | \overleftarrow{x}) , \quad (4)$$

for all  $\overleftarrow{x}$  such that  $\epsilon(\overleftarrow{x}) = \sigma$ . As a consequence, the causal states, considered as a random variable  $\mathcal{S}$ , capture the full predictive information

$$I[\mathcal{S}; \overrightarrow{X}] = I[\overleftarrow{X}; \overrightarrow{X}] = \mathbf{E} . \quad (5)$$

More to the point, they *causally shield* the past and future—the past and future are independent given the causal state:  $P(\overleftarrow{X}, \overrightarrow{X} | \mathcal{S}) = P(\overleftarrow{X} | \mathcal{S})P(\overrightarrow{X} | \mathcal{S})$ .

The causal-state partition has, out of all *equally* predictive partitions, called *prescient rivals*  $\hat{\mathcal{R}}$  [33], the smallest entropy,  $C_\mu[\mathcal{R}] = H[\mathcal{R}]$ :

$$H[\hat{\mathcal{R}}] \geq H[\mathcal{S}] , \quad (6)$$

known as the *statistical complexity*,  $C_\mu := H[\mathcal{S}]$ . This is amount of historical information a process stores: A process communicates  $\mathbf{E}$  bits from the past to the future by storing  $C_\mu$  bits in the present.  $C_\mu$  is one of a process's key properties; the other is its entropy rate [23]. Finally, the causal states are *unique and minimal sufficient statistics* for prediction of the time series [15, 22].

### III. CONSTRUCTING CAUSAL MODELS OF INFORMATION SOURCES

Continuing with the communication channel analogy above, models, optimal or not, can be broadly considered to be a lossy compression of the original data. A model captures some regularity while making some errors in describing the data. Rate distortion theory [19] gives a principled method to find a lossy compression of an information source such that the resulting model is as faithful as possible to the original data, quantified by a *distortion function*. The specific form of the distortion function determines what is considered to be “relevant”—kept in the compressed representation—and what is “irrelevant”—can be discarded. Since there is no universal distortion function, it has to be assumed *ad hoc* for each application. The information bottleneck method [16] argues for explicitly keeping the relevant information, defined as the mutual information that the data share with a desired relevant variable [16]. With those choices, the distortion function can be derived from the optimization principle, but the relevant variable has to be specified *a priori*.

In time series modeling, however, there is a natural notion of relevance: the future data. For stationary time series, moreover, building a model with low generalization error is equivalent to constructing a model that accurately predicts future data from past data. These observations lead directly to an information-theoretic specification for reconstructing time series models: First, introduce general model variables  $\mathcal{R}$  that can store, in the present moment, the information transmitted from the past to the future. Any set of such variables specifies a stochastic partition of  $\vec{\mathbf{X}}$  via a probabilistic assignment rule  $P(\mathcal{R}|\vec{x})$ . Second, require that this partition be maximally predictive. That is, it should maximize the information  $I[\mathcal{R}; \vec{X}]$  that the variables  $\mathcal{R}$  contain about the future  $\vec{X}$ . Third, the so-constructed representation of the historical data should be a summary, i.e., it should not contain all of the historical information, but rather, as little as possible while still capturing the predictive information. The information kept about the past— $I[\vec{X}; \mathcal{R}]$ , the *coding rate*—measures the model complexity or bit cost. Intuitively, one wants to find the most predictive model at fixed complexity or, vice versa, the least complex model at fixed prediction accuracy. These criteria are equivalent, in effect, to causal shielding.

Writing this intuition formally reduces to the information bottleneck method, where the relevant information is information about the future. The constrained optimization problem one has to solve is:

$$\max_{P(\mathcal{R}|\vec{x})} \left\{ I[\mathcal{R}; \vec{X}] - \lambda I[\vec{X}; \mathcal{R}] \right\}, \quad (7)$$

where the parameter  $\lambda$  controls the balance between prediction and model complexity. The linear trade-off that  $\lambda$  represents is an *ad hoc* assumption [18]. Its justification

is greatly strengthened in the following by the rigorous results showing it leads to the causal states and the successful quantitative applications.

The optimization problem of Eq. (7) is solved subject to the normalization constraint:  $\sum_{\mathcal{R}} P(\mathcal{R}|\vec{x}) = 1$ , for all  $\vec{x} \in \vec{\mathbf{X}}$ . It then has a family of solutions [16], parametrized by the Lagrange multiplier  $\lambda$ , that gives the following optimal assignments of histories  $\vec{x}$  to states  $\rho \in \mathcal{R}$ :

$$P_{\text{opt}}(\rho|\vec{x}) = \frac{P(\rho)}{Z(\vec{x}, \lambda)} \exp \left( -\frac{1}{\lambda} \mathcal{D} \left( P(\vec{X}|\vec{x}) || P(\vec{X}|\rho) \right) \right), \quad (8)$$

with

$$P(\vec{X}|\rho) = \frac{1}{P(\rho)} \sum_{\vec{x} \in \vec{\mathbf{X}}} P(\vec{X}|\vec{x}) P(\rho|\vec{x}) P(\vec{x}) \quad \text{and} \quad (9)$$

$$P(\rho) = \sum_{\vec{x} \in \vec{\mathbf{X}}} P(\rho|\vec{x}) P(\vec{x}), \quad (10)$$

where  $\mathcal{D}(P||Q)$  is the *information gain* [23] between distributions  $P$  and  $Q$ . In the solution it plays the role of an “energy”, effectively measuring how different the predicted and true futures are. The more distinct, the more information one gains about the probabilistic development of the future from the past. That is, high energy models make predictions that deviate substantially from the process.

These self-consistent equations are solved iteratively [16] using a procedure similar to the Blahut-Arimoto algorithm [34, 35]. A connection to statistical mechanics is often drawn, and the parameter  $\lambda$  is identified with a (pseudo) temperature that controls the level of randomness; see, e.g., Ref. [36]. This is useful to guide intuition and, for example, has inspired *deterministic annealing* [37].

We are now ready for the first observation.

**Proposition 1.** *In the low-temperature regime ( $\lambda \rightarrow 0$ ) the assignments of pasts to states become deterministic and are given by:*

$$P_{\text{opt}}(\rho|\vec{x}) = \delta_{\rho, \eta(\vec{x})}, \quad \text{where} \quad (11)$$

$$\eta(\vec{x}) = \arg \min_{\rho} \mathcal{D} \left( P(\vec{X}|\vec{x}) || P(\vec{X}|\rho) \right). \quad (12)$$

**Proof.** Define the quantity

$$D(\rho) = \mathcal{D} \left( P(\vec{X}|\vec{x}) || P(\vec{X}|\rho) \right) - \mathcal{D} \left( P(\vec{X}|\vec{x}) || P(\vec{X}|\eta(\vec{x})) \right). \quad (13)$$

$D(\rho)$  is positive, by definition Eq. (12) of  $\eta(\vec{x})$ . Now,

write

$$P_{\text{opt}}(\eta(\overleftarrow{x})|\overleftarrow{x}) = \left( 1 + \sum_{\rho \neq \eta(\overleftarrow{x})} \frac{P(\rho)}{P(\eta(\overleftarrow{x}))} \exp \left[ -\frac{D(\rho)}{\lambda} \right] \right)^{-1}. \quad (14)$$

The sum in the r.h.s. tends to zero, as  $\lambda \rightarrow 0$ , assuming that  $P(\eta(\overleftarrow{x})) > 0$ . Via normalization, the assignments become deterministic.  $\square$

#### IV. OPTIMAL CAUSAL FILTERING

We now establish the procedure's fundamental properties by connecting the solutions it determines to the causal representations defined previously within the framework of computational mechanics. The resulting procedure transforms the original data to a causal representation and so we call it *optimal causal filtering* (OCF).

Note first that for deterministic assignments we have  $H[\mathcal{R}|\overleftarrow{X}] = 0$ . Therefore, the information about the past becomes  $I[\overleftarrow{X}; \mathcal{R}] = H[\mathcal{R}]$  and the objective function simplifies to

$$F_{\text{det}}[\mathcal{R}] = I[\mathcal{R}; \overleftarrow{X}] - \lambda H[\mathcal{R}]. \quad (15)$$

**Lemma 1.** *Within the subspace of prescient rivals, the causal-state partition maximizes  $F_{\text{det}}[\widehat{\mathcal{R}}]$ .*

**Proof.** *This follows immediately from Eqs. (5) and (6). They imply that*

$$\begin{aligned} F_{\text{det}}[\widehat{\mathcal{R}}] &= I[\mathcal{S}; \overleftarrow{X}] - \lambda H[\widehat{\mathcal{R}}] \\ &\leq I[\mathcal{S}; \overleftarrow{X}] - \lambda H[\mathcal{S}] \\ &= F_{\text{det}}[\mathcal{S}]. \end{aligned} \quad (16)$$

$\square$

The causal-state partition is the model with the largest value of the OCF objective function, because it is fully predictive at minimum complexity. We also know from Prop. 1 that in the low-temperature limit ( $\lambda \rightarrow 0$ ) OCF recovers a *deterministic* mapping of histories to states. We now show that this mapping is exactly the causal-state partition of histories.

**Theorem 1.** *OCF finds the causal-state partition of  $\overleftarrow{X}$  in the low-temperature limit,  $\lambda \rightarrow 0$ .*

**Proof.** *The causal-state partition, Eq. (1), always exists, and implies that there are groups of histories with*

$$P(\overrightarrow{X}|\overleftarrow{x}) = P(\overrightarrow{X}|\epsilon(\overleftarrow{x})). \quad (17)$$

We then have, for all  $\overleftarrow{x} \in \overleftarrow{X}$ ,

$$\mathcal{D}(P(\overrightarrow{X}|\overleftarrow{x})||P(\overrightarrow{X}|\epsilon(\overleftarrow{x}))) = 0, \quad (18)$$

and, hence,

$$\epsilon(\overleftarrow{x}) = \arg \min_{\rho} \mathcal{D}(P(\overrightarrow{X}|\overleftarrow{x})||P(\overrightarrow{X}|\rho)). \quad (19)$$

Therefore, we can identify  $\epsilon(\overleftarrow{x}) = \eta(\overleftarrow{x})$  in Eq. (12), and so the assignment of histories to the causal states is recovered by OCF:

$$P_{\text{opt}}(\rho|\overleftarrow{x}) = \delta_{\rho, \epsilon(\overleftarrow{x})}. \quad (20)$$

$\square$

Note that we have not restricted the size of the set  $\mathcal{R}$  of model states. Recall also that the causal-state partition is *unique* [22]. The Lemma establishes that OCF does *not* find prescient rivals in the low-temperature limit. The prescient rivals are suboptimal in the particular sense that they have smaller values of the objective function. We now establish that this difference is controlled by the model size with proportionality constant  $\lambda$ .

**Corollary 1.** *Prescient rivals are suboptimal in OCF. The value of the objective function evaluated for a prescient rival is smaller than that evaluated for the causal-state model. The difference  $\Delta F_{\text{det}}[\widehat{\mathcal{R}}] = F_{\text{det}}[\mathcal{S}] - F_{\text{det}}[\widehat{\mathcal{R}}]$  is given by:*

$$\Delta F_{\text{det}}[\widehat{\mathcal{R}}] = \lambda (C_{\mu}[\widehat{\mathcal{R}}] - C_{\mu}[\mathcal{S}]) \geq 0. \quad (21)$$

**Proof.**

$$\Delta F_{\text{det}}[\widehat{\mathcal{R}}] = F_{\text{det}}[\mathcal{S}] - F_{\text{det}}[\widehat{\mathcal{R}}] \quad (22)$$

$$= I[\mathcal{S}; \overleftarrow{X}] - I[\widehat{\mathcal{R}}; \overleftarrow{X}] - \lambda H[\mathcal{S}] + \lambda H[\widehat{\mathcal{R}}] \quad (23)$$

$$= \lambda (C_{\mu}[\widehat{\mathcal{R}}] - C_{\mu}[\mathcal{S}]). \quad (24)$$

Moreover, Eq. (6) implies that  $\Delta F_{\text{det}} \geq 0$ .  $\square$

So, we see that for  $\lambda = 0$ , causal states and all other prescient rival partitions are degenerate. This is to be expected as at  $\lambda = 0$  the model-complexity constraint disappears. Importantly, this means that maximizing the predictive information alone, without the appropriate constraint on model complexity does not suffice to recover the causal-state partition.

#### V. EXAMPLES

We study how OCF works on a series of example stochastic processes of increasing statistical sophistication. We compute the optimal solutions and visualize the trade-off between predictive power and complexity of the model by tracing out a curve similar to a rate-distortion curve [34, 35]: For each value of  $\lambda$ , we evaluate both the model's coding rate  $I[\overleftarrow{X}; \mathcal{R}]$  and its predicted information  $I[\mathcal{R}; \overleftarrow{X}]$  at the optimal solution and plot them

against each other. The resulting curve in the *information plane* [16] separates the feasible from the infeasible region: It is possible to find a model that is more complex at the same prediction error, but not possible to find a less complex model than that given by the optimum. In analogy to a rate-distortion curve, we can read off the maximum amount of information about the future that can be captured with a model of fixed complexity. Or, conversely, we can read off the smallest representation at fixed predictive power.

The examples in this and the following sections are calculated by solving the self-consistent Eqs. (8) to (10) iteratively [46] at each value of  $\lambda$ . To trace out the curves, a deterministic annealing [37] scheme is implemented, lowering  $\lambda$  by a fixed annealing rate. Smaller rates cost more computational time, but allow one to compute the rate-distortion curve in greater detail, while larger rates result in a rate-distortion curve that gets evaluated in fewer places and hence looks coarser. In examples, naturally, one can only work with finite length past and future sequences:  $\overleftarrow{x}^K$  and  $\overrightarrow{x}^L$ , where  $K$  and  $L$  give their lengths, respectively.

#### A. Periodic limit cycle: A predictable process

We start with an example of an exactly periodic process, a limit cycle oscillation. It falls in the class of deterministic and time reversible processes, for which the rate-distortion curve can be computed analytically—it lies on the diagonal [20]. We demonstrate this with a numerical example. Figure 1 shows how OCF works on a period-four process:  $(0011)^\infty$ . (See Figs. 1 and 2.) There are exactly two bits of predictive information  $I[\overleftarrow{X}; \overrightarrow{X}]$  to be captured about future words of length two (dotted horizontal line). This information describes the phase of the period-four cycle. To capture those two bits, one needs exactly four underlying causal states and a model complexity of  $C_\mu = 2$  bits (dotted vertical line).

The curve is the analog of a rate-distortion curve, except that the information plane swaps the horizontal and vertical axes—the coding rate and distortion axes. (See Ref. [20] for the direct use of the rate-distortion curve.) The value of  $I[\mathcal{R}; \overrightarrow{X}^2]$  (the “distortion”), evaluated at the optimal distribution, Eq. (8), is plotted versus  $I[\overleftarrow{X}^3; \mathcal{R}]$  (the “code rate”), also evaluated at the optimum. Those are plotted for different values of  $\lambda$  and, to trace out the curve, deterministic annealing is implemented. At large  $\lambda$ , we are in the lower left of the curve—the compression is extreme, but no predictive information is captured. A single state model, a fair coin, is found as expected. As  $\lambda$  decreases (moving to the right), the next distinct point on the curve is for a two-state model, which discards half of the information. This comes exactly at the cost of one predictive bit. Finally, OCF finds a four-state model that captures all of the predictive information at no compression. The numbers next to the curve indicate the first

time that the effective number of states increases to that value.

The four-state model captures the two bits of predictive information. But compressed to one bit (using two states), one can only capture one bit of predictive information. The information curve falls onto the diagonal—a straight line that is the worst case for possible beneficial trade-offs between prediction error and model complexity [20].

In Fig. 2, we show the best two-state model compared to the full (exact) four-state model. One of the future conditional probabilities captures zero probability events of “odd”  $\{01, 10\}$  words, assigning equal probability to the “even”  $\{00, 11\}$  words. The other one captures zero probability events of even words, assigning equal probability to the odd words. This captures the fundamental determinism of the process: an odd word never follows an even word and vice versa. The overall result illustrates how the actual long-range correlation in the completely predictable period-4 sequence is represented by a smaller *stochastic* model. While in the four-state model the future conditional probabilities are  $\delta$ -functions, in the two-state approximate model they are mixtures of those  $\delta$ -functions. In this way, OCF converts structure to randomness when approximating underlying states with a compressed model; cf. the analogous trade-off discussed in Ref. [25].

#### B. Golden Mean Process: A Markov chain

The Golden Mean (GM) Process is a Markov chain of order one. As an information source, it produces all binary strings with the restriction that there are never consecutive 0s. The GM Process generates 0s and 1s with equal probability, except that once a 0 is generated, a 1 is always generated next. One can write down a simple two-state Markov chain for this process; see, e.g., Ref. [25].

Figures 3 and 4 demonstrate how OCF reconstructs the states of the GM process. Figure 3 shows the behavior of OCF in the information plane. At very high temperature ( $\lambda \rightarrow \infty$ , lower left corner of the curve) compression dominates over prediction and the resulting model is most compact, with only one effective causal state. However, it contains no information about the future and so is a poor predictor. As  $\lambda$  decreases (moving right), OCF reconstructs increasingly more predictive and more complex models. The curve shows that the information about the future, contained in the optimal partition, increases (along the vertical axis) as the model increases in complexity (along the horizontal axis). There is a transition to two effective states: the number 2 along the curve denotes the first occurrence of this increase. As  $\lambda \rightarrow 0$ , prediction comes to dominate and OCF finds a fully predictive model, albeit one with the minimal statistical complexity, out of all possible state partitions that would retain the full predictive informa-

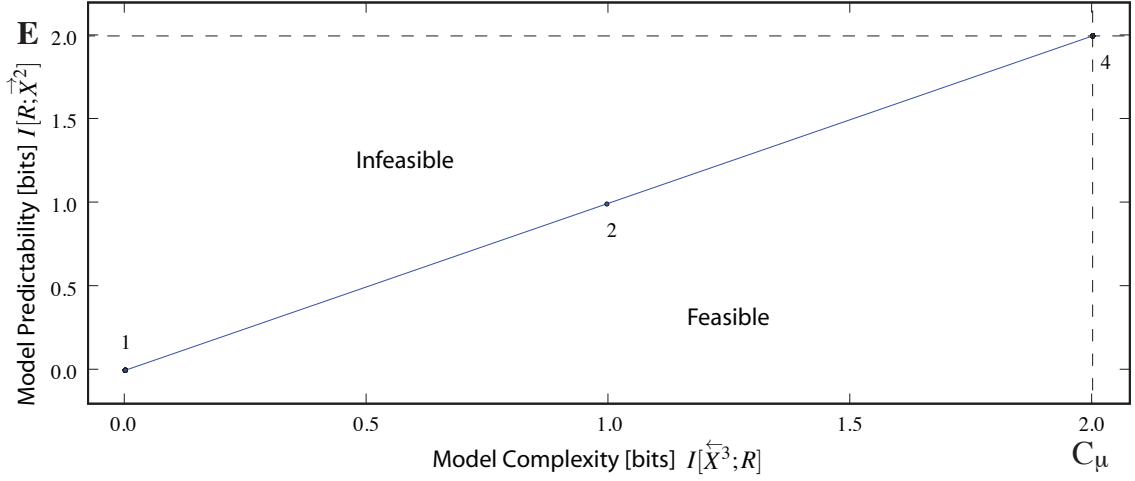


FIG. 1: Model predictability  $I[\mathcal{R}; \vec{X}^L]$  versus model complexity (size)  $I[\vec{X}^K; \mathcal{R}]$  trade-off under OCF for the exactly predictable period-4 process:  $(0011)^\infty$ . Monitored in the information plane. The horizontal dashed line is the full predictive information ( $\mathbf{E} = I[\vec{X}^3; \vec{X}^2] = 2$  bits) and the vertical dashed line is the block entropy ( $H[\vec{X}^3] = 2$  bits), which is also the statistical complexity  $C_\mu$ . The data points represent solutions at various  $\lambda$ . Lines connect them to help guide the eye only. Histories of length  $K = 3$  were used, along with futures of length  $L = 2$ . In this and the following information plane plots, the integer labels  $N_c$  indicate the first point at which the effective number of states used by the model equals  $N_c$ .

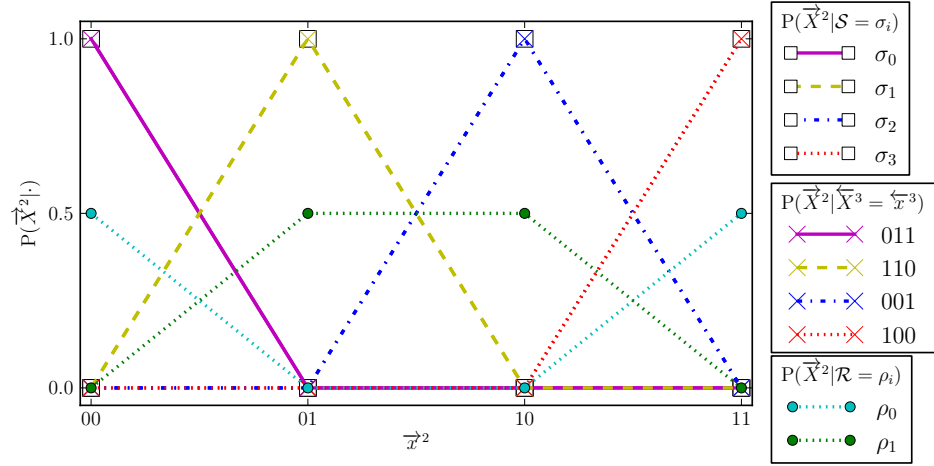


FIG. 2: Morphs  $P(\vec{X}^2 | \cdot)$  for the period-4 process: The 2-state approximation (circles) compared to the  $\delta$ -function morphs for the 4 causal states (boxes). The morphs  $P(\vec{X}^2 | \sigma)$  for the two-state approximation are  $(1/2, 0, 0, 1/2)$  and  $(0, 1/2, 1/2, 0)$  and for the four-state case  $(1, 0, 0, 0)$ ,  $(0, 1, 0, 0)$ ,  $(0, 0, 1, 0)$ , and  $(0, 0, 0, 1)$ . Histories of length  $K = 3$  were used, along with futures of length  $L = 2$  (crosses).

tion. The model's complexity— $C_\mu \approx 0.92$  bits—is 41% of the maximum, which is given by the entropy of all possible pasts of length 3:  $H[\vec{X}^3] \approx 2.25$  bits. The remainder (59%) of the information is nonpredictive and has been filtered out by OCF. Figure 4 shows the future conditional probabilities, associated with the partition found by OCF, as  $\lambda \rightarrow 0$ , corresponding to  $P(\vec{X}^2 | \rho)$  (circles). These future conditional probabilities overlap with the true (but not known to the algorithm) causal-state future conditional probabilities  $P(\vec{X}^2 | \sigma)$  (boxes) and so demonstrate that OCF finds the causal-state partition.

### C. Even Process: A hidden Markov chain

Now, consider a hidden Markov process: the *Even Process* [25], which is a stochastic process whose support (the set of allowed sequences) is a symbolic dynamical system called the *Even system*. The Even system generates all binary strings consisting of blocks of an even number of 1s bounded by 0s. Having observed a process's sequences, we say that a word (finite sequence of symbols) is *forbidden* if it never occurs. A word is an *irreducible forbidden word* if it contains no proper subwords which are themselves forbidden words. A system is *sofic* if its list of irreducible forbidden words is infinite. The Even system

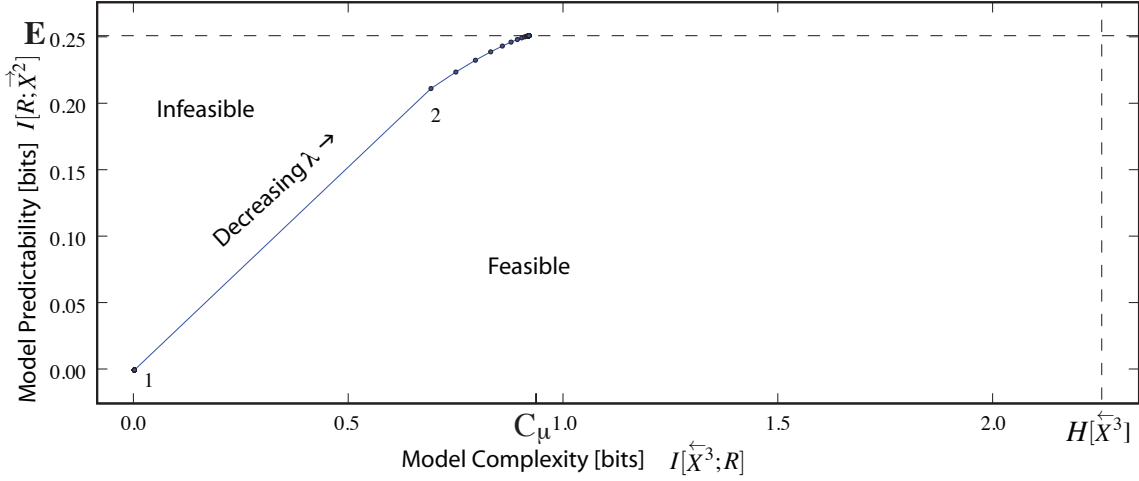


FIG. 3: OCF’s behavior monitored in the information plane— $I[\mathcal{R}; \vec{X}^2]$  versus  $I[\vec{X}^3; \mathcal{R}]$ —for the Golden Mean Process. The correct two-state model is found. Histories of length  $K = 3$  were used, along with futures of length  $L = 2$ . The horizontal dashed line is the full predictive information  $\mathbf{E} \approx I[\vec{X}^3; \vec{X}^2] = I[\mathcal{S}; \vec{X}^2] \approx 0.25$  bits which, as seen, is an upper bound on  $I[\mathcal{R}; \vec{X}^2]$ . The exact value is  $\mathbf{E} = I[\vec{X}; \vec{X}] = 0.2516$  bits [38]. Similarly, the vertical dashed line is the block entropy  $H[\vec{X}^3] \approx 2.25$  bits which is an upper bound on the retrodictive information  $I[\vec{X}^3; \mathcal{R}]$ . The statistical complexity  $C_\mu \approx 0.92$  bits, also an upper bound, is labeled. The annealing rate was 0.952.

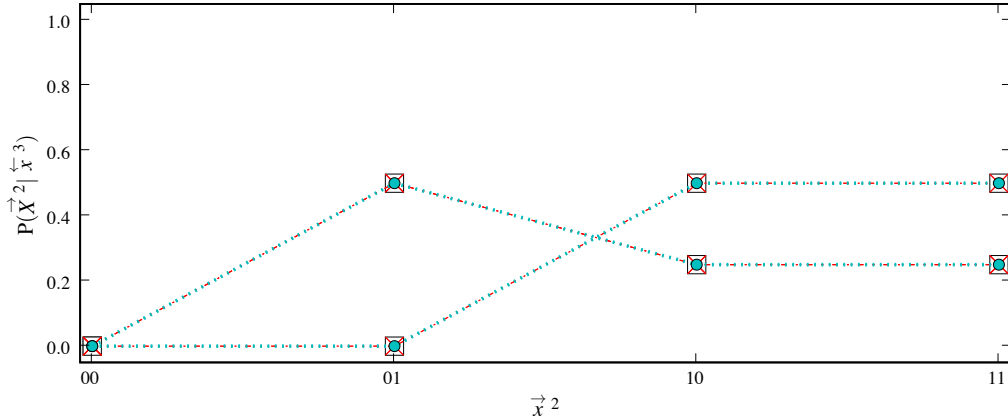


FIG. 4: Future conditional probabilities  $P(\vec{X}^2 | \cdot)$  conditioned on causal states  $\sigma \in \mathcal{S}$  (boxes) and on the OCF reconstructed states  $\rho \in \mathcal{R}$  (circles) for the Golden Mean Process. As an input to OCF, future conditional probabilities  $P(\vec{X}^2 | \vec{x}^3)$  calculated from histories of length  $K = 3$  were used (crosses).

is one such sofic system, since its set  $\mathcal{F}$  of irreducible forbidden words is infinite:  $\mathcal{F} = \{01^{2n+1}0, n = 0, 1, \dots\}$ . Note that no finite-order Markovian source can generate this or, for that matter, any other strictly sofic system [25]. The Even Process then associates probabilities with each of the Even system’s sequences by choosing a 0 or 1 with fair probability after generating either a 0 or a pair of 1s. The result is a *measure sofic process*—a distribution over a sofic system’s sequences.

As in the previous example, for large  $\lambda$ , OCF applied to the Even Process recovers a small, one-state model with poor predictive quality; see Fig. 5. As  $\lambda$  decreases there are transitions to larger models that capture increasingly more information about the future. (The numbers along the curve again indicate the points of first transition to

more states.) With a three-state model OCF captures the full predictive information at a model size of 56% of the maximum. This model is exactly the causal-state partition, as can be seen in Fig. 6 by comparing the future conditional probabilities of the OCF model (circles) to the true underlying causal states (boxes), which are not known to the algorithm.

The correct  $\epsilon$ -machine model of the Even Process has four causal states: two transient and two recurrent. At the finite past and future lengths used here, OCF picks up only one of the transient states and the two recurrent states. It also assigns probability to all three. This increases the effective state entropy ( $H[\mathcal{R}] \approx 1.48$  bits) above the statistical complexity ( $C_\mu = 0.92$  bits) which is only a function of the two recurrent states, since asymp-

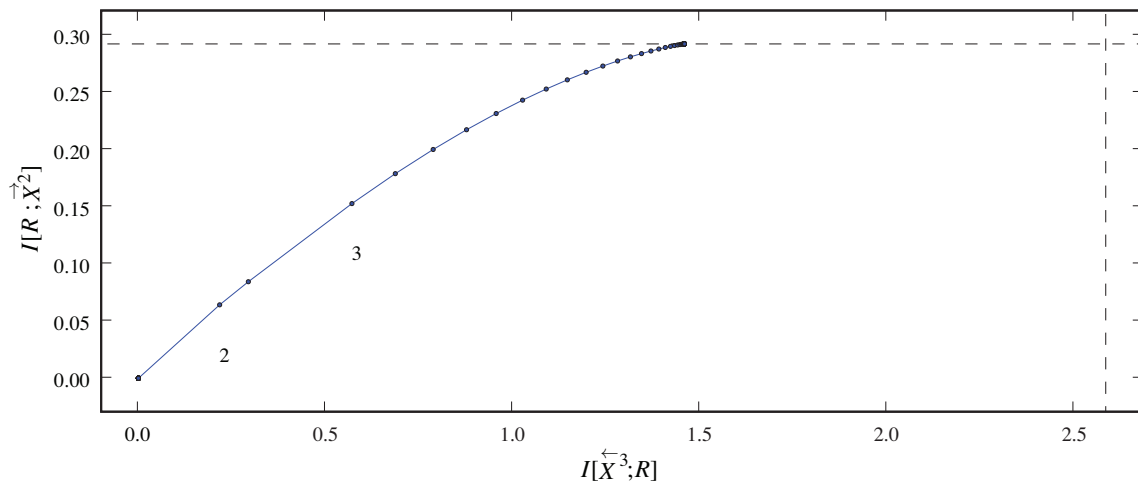


FIG. 5: OCF’s behavior inferring the Even Process: monitored in the information plane— $I[\mathcal{R}; \vec{X}^2]$  versus  $I[\vec{X}^3; \mathcal{R}]$ . Histories of length  $K = 3$  were used, along with futures of length  $L = 2$ . The horizontal dashed line is the full predictive information  $I[\vec{X}^3; \vec{X}^2] \approx 0.292$  bits which, as seen, is an upper bound on the estimates  $I[\mathcal{R}; \vec{X}^2]$ . Similarly, the vertical dashed line is the block entropy  $H[\vec{X}^3] \approx 2.585$  bits which is an upper bound on the retrodictive information  $I[\vec{X}^3; \mathcal{R}]$ .

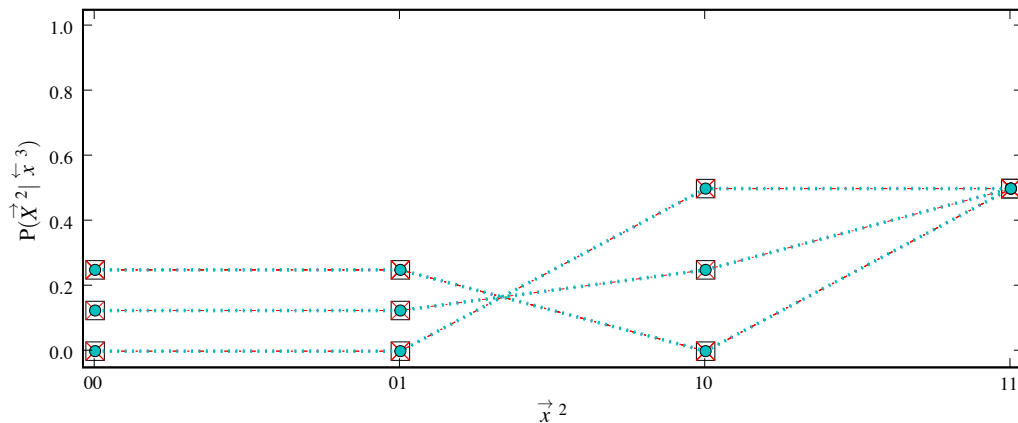


FIG. 6: Future conditional probabilities  $P(\vec{X}^2 | \cdot)$  conditioned on causal states  $\sigma \in \mathcal{S}$  (boxes) and on the OCF-reconstructed states  $\rho \in \mathcal{R}$  (circles) for the Even Process. As an input to OCF, future conditional probabilities  $P(\vec{X}^2 | \vec{x}^3)$  calculated from histories of length  $K = 3$  were used (crosses).

totically ( $K \rightarrow \infty$ ) the transient states have zero probability.

There is an important lesson in this example for general time-series modeling, not just OCF. Correct inference of even finite-state, but measure-sofic processes requires using hidden Markov models. Related consequences of this, and one resolution, are discussed at some length for estimating “nonhidden” Markov models of sofic processes in Ref. [39].

#### D. Random Random XOR: A structurally complex process

The previous examples demonstrated our main theoretical result: In the limit in which it becomes crucial to make the prediction error very small, at the expense

of the model size, the OCF algorithm captures all of the structure inherent in the process by recovering the causal-state partition.

However, if we allow (or prefer) a model with some finite prediction error, then we can make the model substantially smaller. We have already seen what happens in the worst case scenario, for a periodic process. There, each predictive bit costs exactly one bit in terms of model size. However, for highly structured processes, there exist situations in which one can compress the model substantially at essentially no loss in terms of predictive power. (This is called *causal compressibility* [20].) The Even Process is an example of such an information source: The statistical complexity  $H[\mathcal{S}]$  of the causal-state partition is smaller than the total available historical information—the entropy of the past  $H[\vec{X}^K]$ .

Now, we study a process that requires keeping *all* of the



historical information to be maximally predictive, which is the same as stating  $C_\mu(\mathcal{R}) = H[\vec{X}^K]$ . (Precisely, we mean given the finite past and future lengths we use.) Nonetheless, there is a systematic ordering of models of different size and different predictive power given by the rate-distortion curve, as we change the parameter  $\lambda$  that controls how much of the future fluctuations the model considers to be random; i.e., which fluctuations are considered indistinguishable. Naturally, the trade-off, and therefore the shape of the rate-distortion curve, depends on and reflects the source's organization.

As an example, consider the random-random XOR (RRXOR) process which consists of two successive random symbols chosen to be 0 or 1 with equal probability and a third symbol that is the logical Exclusive-OR (XOR) of the two previous. The RRXOR process can be represented by a hidden Markov chain with five recurrent causal states, but having a very large total number of causal states. There are 36 causal states, most (31) of which describe a complicated transient structure [25]. As such, it is a structurally complex process that an analyst may wish to approximate with a smaller set of states.

Figure 7 shows the information plane, which specifies how OCF trades off structure for prediction error as a function of model complexity for the RRXOR process. The number of effective states (again first occurrences are denoted by integers along the curve) increases with model complexity. At a history length of  $K = 3$  and future length of  $L = 2$ , the process has eight underlying causal states, which are found by OCF in the  $\lambda \rightarrow 0$  limit. The corresponding future conditional probability distributions are shown in Fig. 8.

The RRXOR process has a structure that does not allow for substantial compression. Fig. 7 shows that the effective statistical complexity of the causal-state partition is equal to the full entropy of the past:  $C_\mu(\mathcal{R}) = H[\vec{X}^3]$ . So, at  $L = 3$ , unlike the Even and Golden Mean Processes, the RRXOR process is not compressible. With half (4) of the number of states, however, OCF reconstructs a model that is only 33% as large, while capturing 50% of the information about the future. The corresponding conditional future probabilities of the (best) four-state model are shown in Fig. 9. They are mixtures of pairs of the eight causal states.

The rate-distortion curve informs the modeler about the (best possible) efficiency of predictive power to model complexity:  $I[\mathcal{R}; \vec{X}]/I[\vec{X}; \mathcal{R}]$ . This is useful, for example, if there are constraints on the maximum model size or, vice versa, on the minimum prediction error. For example, if we require a model of RRXOR to be 90% informative about the future, then we can read off the curve that this can be achieved at 70% of the model complexity. Generally, as  $\lambda$  decreases, phase transitions occur to models with a larger number of effective states [37].

## VI. OPTIMAL CAUSAL ESTIMATION: FINITE-DATA FLUCTUATIONS

In real world applications, we do not know a process's underlying probability density, but instead must estimate it from a *finite* time series that we are given. Let that time series be of length  $T$  and let us estimate the joint distribution of pasts (of length  $K$ ) and futures (of length  $L$ ) via a histogram calculated using a sliding window. Altogether we have  $M = T - (K + L - 1)$  observations. The resulting estimate  $\hat{P}(\vec{X}^K; \vec{X}^L)$  will deviate from the true  $P(\vec{X}^K; \vec{X}^L)$  by  $\Delta(\vec{X}^K, \vec{X}^L)$ . This leads to an overestimate of the mutual information [47]:  $\hat{I}[\vec{X}^K; \vec{X}^L] \geq I[\vec{X}^K; \vec{X}^L]$ . Evaluating the objective function at this estimate may lead one to capture variations that are due to the sampling noise and not to the process's underlying structure; i.e., OCF may over-fit. That is, the underlying process may appear to have a larger number  $N_c$  of causal states than the true number.

Following Ref. [17], we argue that this effect can be counteracted by subtracting from  $\hat{F}[\mathcal{R}]$  a model-complexity control term that approximates the error we make by calculating the estimate  $\hat{F}[\mathcal{R}]$  rather than the true  $F[\mathcal{R}]$ . If we are willing to assume that  $M$  is large enough, so that the deviation  $\Delta(\vec{X}^K, \vec{X}^L)$  is a small perturbation, then the error can be approximated by [17, Eq. (5.8)]:

$$\mathcal{E}(N_c) = \frac{k^L - 1}{2 \ln(2)} \frac{N_c}{M}, \quad (25)$$

in the low-temperature regime,  $\lambda \rightarrow 0$ . Recall that  $k^L$  is the total number of possible futures for alphabet size  $k$ . The optimal number  $N_c^*$  of hidden states is then the one for which the largest amount of mutual information is shared with the future, corrected by this error:

$$N_c^* := \arg \max_{N_c} \hat{I}[\vec{X}^K; \vec{X}^L]_{\lambda \rightarrow 0}^{\text{corrected}}(N_c), \quad (26)$$

with

$$\hat{I}[\vec{X}^K; \vec{X}^L]_{\lambda \rightarrow 0}^{\text{corrected}}(N_c) = \hat{I}[\vec{X}^K; \vec{X}^L]_{\lambda \rightarrow 0}(N_c) - \mathcal{E}(N_c). \quad (27)$$

This correction generalizes OCF to *optimal causal estimation* (OCE), a procedure that simultaneously accounts for the trade-off between structure, approximation, and sample fluctuations.

We illustrate OCE on the Golden Mean and Even Processes studied in Sec. V. With the *correct* number of underlying states, they can be predicted at a substantial compression. Figures 10 and 12 show the mutual information  $I[\mathcal{R}; \vec{X}^2]$  versus the number  $N_c$  of inferred states, with statistics estimated from time series of lengths  $T = 100$ . The graphs compare the mutual information  $\hat{I}[\mathcal{R}; \vec{X}^2]_{\lambda \rightarrow 0}$  evaluated using the estimate  $\hat{P}(\vec{X}^2; \vec{X}^3)$  (upper curve) to the corrected information

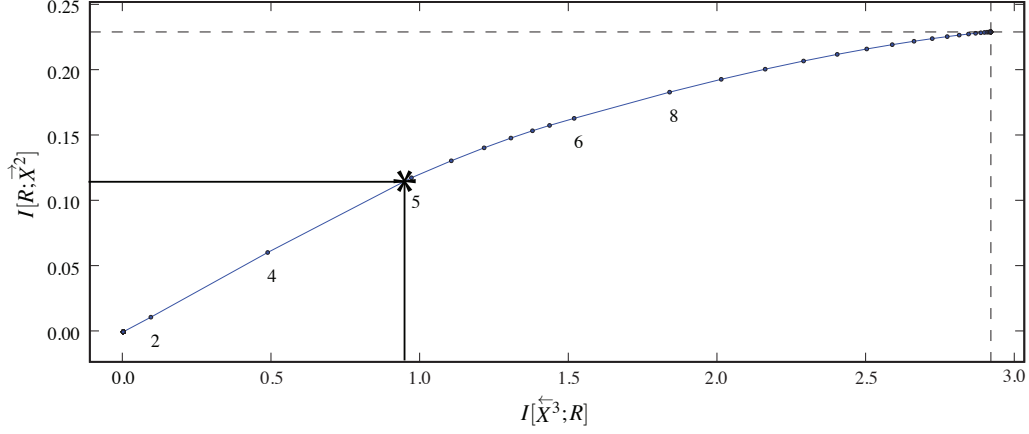


FIG. 7: Prediction versus structure trade-off under OCF for the random-random XOR (RRXOR) process, as monitored in the information plane. As above, the horizontal dashed line is the predictive information ( $\approx 0.230$  bits) and the vertical dashed line is the block entropy ( $\approx 2.981$  bits). Histories of length  $K = 3$  were used, along with futures of length  $L = 2$ . The asterisk and lines correspond to the text: they serve to show how the predictive power and the complexity of the best four state model, the future conditional probabilities of which are depicted in Fig. 9.

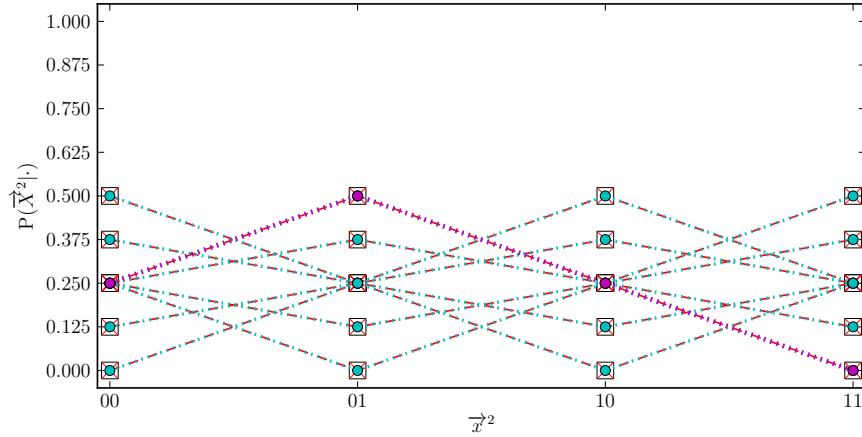


FIG. 8: Future conditional probabilities  $P(\vec{X}^2 | \cdot)$  for the RRXOR process: the 8-state approximation (circles) finds the causal states (boxes). For example, the heavier dashed line (purple) shows  $P(\vec{X}^2 | \rho) = (1/4, 1/2, 1/4, 0)$ . Histories of length  $K = 3$  were used, along with futures of length  $L = 2$ .

$\hat{I}[\mathcal{R}; \vec{X}^2]_{\lambda \rightarrow 0}^{\text{corrected}}$  calculated by subtracting the approximated error Eq. (25) with  $k^L = 4$  and  $M = 96$  (lower curve).

We see that the corrected information curves peak at, and thereby, select models with two states for the Golden Mean Process and three states for the Even Process. This corresponds with the true number of causal states, as we know from above (Sec. V) for the two processes. The true statistical complexity for both processes is  $C_\mu \approx 0.91830$ , while those estimated via OCE are  $C_\mu \approx 0.93773$  and  $C_\mu \approx 1.30262$ , respectively. (Recall that the overestimate for the latter was explained in Sec. VC.)

Figures 11 and 13 show the OCE future conditional probabilities corresponding to the (optimal) two- and three-state approximations, respectively. The input to

OCE are the future conditional probabilities given the histories  $\hat{P}(\vec{X}^2 | \vec{x}^3)$  (crosses), which are estimated from the full historical information. Those future conditional probabilities are corrupted by sampling errors due to the finite data set size and differ from the true future conditional probabilities (squares).

Compare the OCE future conditional probabilities (circles) to the true future conditional probabilities (squares), calculated with the knowledge of the causal states. (The latter, of course, is not available to the OCE algorithm.) In the case of the GM Process, OCE approximates the correct future conditional probabilities. For the Even Process there is more spread in the estimated OCE future conditional probabilities. Nonetheless, OCE reduced the fluctuations in its inputs and corrected in

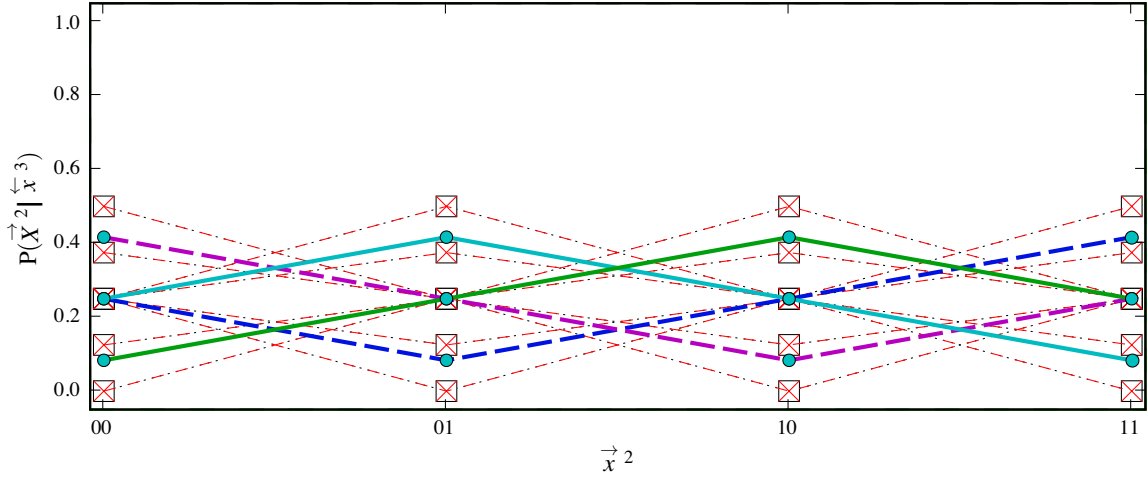


FIG. 9: Morphs  $P(\vec{X}^2 | \cdot)$  for the RRXOR process: the 4-state approximation (circles and colored lines: state 1 - cyan/full, 2 - green/full, 3 - blue/dashed, 4 - purple/dashed) compared to causal states (boxes). Histories of length  $K = 3$  were used, along with futures of length  $L = 2$ .

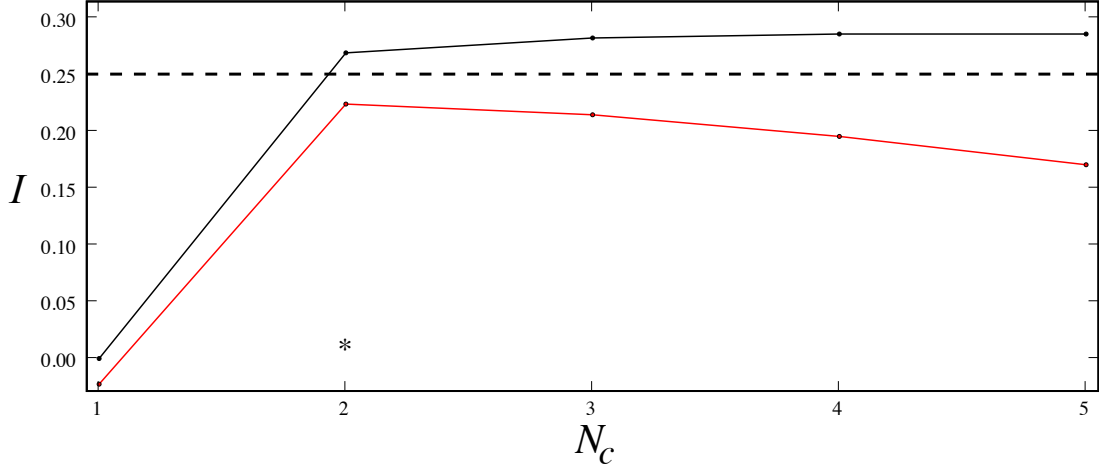


FIG. 10: Information  $I$  captured about the future versus the number  $N_c$  of reconstructed states, with statistics estimated from length  $T = 100$  time series sample from the Golden Mean Process. Upper line: plotted on the vertical axis is  $\hat{I}[\mathcal{R}; \vec{X}^2]_{\lambda \rightarrow 0}$  (not corrected); lower line: plotted on the vertical axis is the quantity  $\hat{I}[\mathcal{R}; \vec{X}^2]_{\lambda \rightarrow 0}^{\text{corrected}}$ , which is the retained predictive information, but corrected for estimation errors due to finite sample size. The dashed line indicates the actual upper bound on the predictive information  $I[\vec{X}^K; \mathcal{R}]$ , for comparison. This value is not known to the algorithm, it is computed from the true process statistics. Histories of length  $K = 3$  and futures of length  $L = 2$  were used. The asterisk denotes the optimal number ( $N_c = 2$ ) of effective states.

the direction of the true underlying future conditional probabilities.

## VII. CONCLUSION

We analyzed an information-theoretic approach to causal modeling in two distinct cases: (i) optimal causal filtering (OCF), where we have access to the process statistics and desire to capture the process's structure up to some level of approximation, and (ii) optimal causal estimation (OCE), in which, in addition, finite-data fluctuations need to be traded-off against approximation error and structure. The objective function used in both cases

follows from very simple first principles of information processing and causal modeling: a good model should minimize prediction error at minimal model complexity. The resulting principle of using small, predictive models follows from minimal prior knowledge that, in particular, makes no structural assumptions about a process's architecture: Find variables that do the best at causal shielding.

OCF stands in contrast with other approaches. Hidden Markov modeling, for example, assumes a set of states and an architecture [40]. OCF finds these states from the given data. In minimum description length modeling, to mention another contrast, the model complexity of a stochastic source diverges (logarithmically) with the data

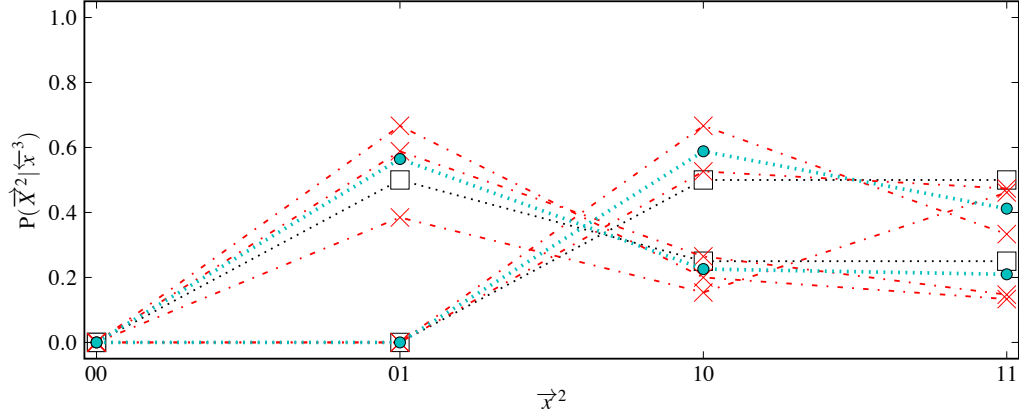


FIG. 11: OCE’s best two-state approximated future conditional probabilities (circles) for the Golden Mean Process. Compared to true (unknown) future conditional probabilities (squares). The OCE inputs are the estimates of  $\hat{P}(\vec{X}^2 | \vec{x}^3)$  (crosses).

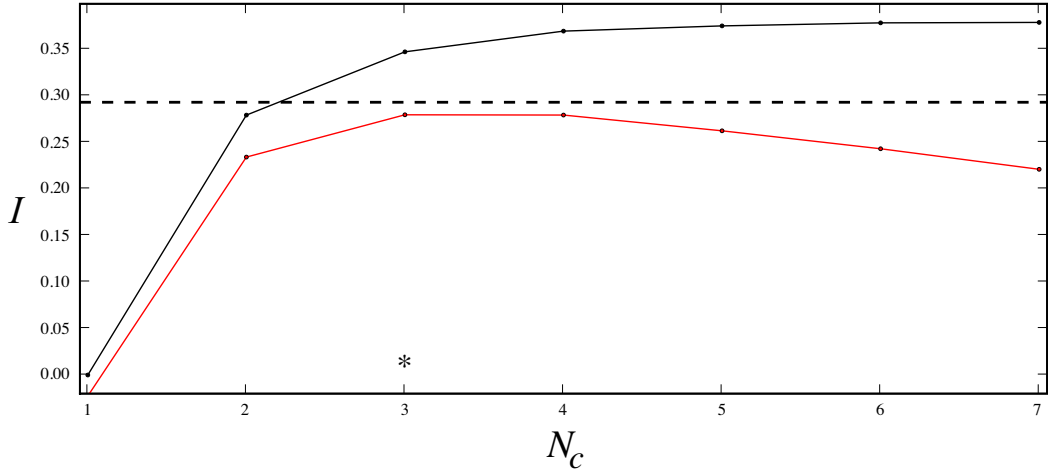


FIG. 12: Information  $I$  captured about the future versus the number  $N_c$  of reconstructed states, with statistics estimated from length  $T = 100$  time series sample from the Even Process. Upper line:  $\hat{I}[\mathcal{R}; \vec{X}^2]_{\lambda \rightarrow 0}$ , not corrected; lower line:  $\hat{I}[\mathcal{R}; \vec{X}^2]_{\lambda \rightarrow 0}^{\text{corrected}}$ , corrected for estimation error due to finite sample size. The dashed line indicates the actual upper bound on the predictive information, for comparison. This value is not known to the algorithm, it is computed from the true process statistics. Histories of length  $K = 3$  and futures of length  $L = 2$  were used. The asterisk denotes the optimal number ( $N_c = 3$ ) of effective states.

set size [41], as happens even when modeling the ideal random process of a fair coin. OCF, however, finds the simplest (smallest) models.

Our main result is that OCF reconstructs the causal-state partition, a representation previously known from computational mechanics that captures a process’s causal architecture and that allows important system properties, such as entropy rate and stored information, to be calculated [22]. This result is important as it gives a structural meaning to the solutions of the optimization procedure specified by the causal inference objective function. We have shown that in the context of time series modeling, where there is a *natural* relevant variable (the future), the IB approach [16] recovers the unique minimal sufficient statistic—the causal states—in the limit in which prediction is paramount to compression. Altogether, this allows us to go beyond plausibility arguments for the information-theoretic objective

function that have been used. We showed that this way (OCI) of phrasing the causal inference problem in terms of causal shielding results in a representation that is a sufficient statistic and minimal and, moreover, reflects the structure of the process that generated the data. OCI does so in a way that is meaningful and well grounded in physics and nonlinear dynamics. The optimal solutions to balancing prediction and model complexity take on meaning—asymptotically, they are the causal states.

The results also contribute to computational mechanics: The continuous trade-off allows one to extend the deterministic history-to-state assignments that computational mechanics introduced to “soft” partitions of histories. The theory gives a principled way of constructing stochastic approximations of the ideal causal architecture. The resulting approximated models can be substantially smaller and so will be useful in a number of applications.

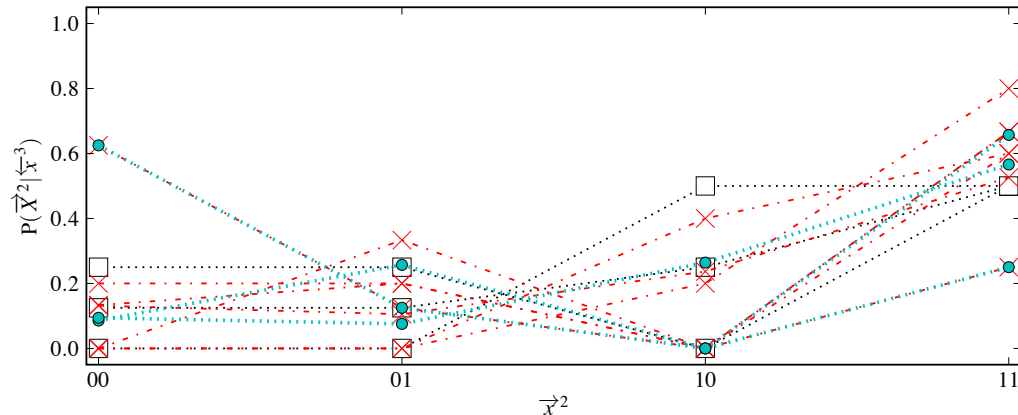


FIG. 13: OCE's best three-state approximated future conditional probabilities (circles) for the Even Process (d). Compared to true (unknown) future conditional probabilities (squares). The OCE inputs are the estimates of  $\hat{P}(X^2 | X^3)$  (crosses).

Finally, we showed how OCF can be adapted to correct for finite-data sampling fluctuations and so not over-fit. This reduces the tendency to see structure in noise. OCE finds the correct number of hidden causal states. This renders the method useful for application to real data.

### Acknowledgments

UC Davis and the Santa Fe Institute partially supported this work via the Network Dynamics Program

funded by Intel Corporation. It was also partially supported by the DARPA Physical Intelligence Program. CJE was partially supported by a Department of Education GAANN graduate fellowship. SS thanks W. Bialek, discussions with whom have contributed to shaping some of the ideas expressed, and thanks L. Bottou and I. Nemenmann for useful discussions.

- 
- [1] P. Berge, Y. Pomeau, and C. Vidal, *Order within chaos* (Wiley, New York, 1986).
  - [2] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields* (Springer-Verlag, New York, 1983).
  - [3] S. Wiggins, *Global Bifurcations and Chaos: analytical methods* (Springer-Verlag, New York, 1988).
  - [4] R. L. Devaney, *An Introduction to Chaotic Dynamical Systems* (Addison-Wesley, Redwood City, California, 1989).
  - [5] A. J. Lieberman and M. A. Lichtenberg, *Regular and Chaotic Dynamics* (Springer-Verlag, New York, 1993), 2nd ed.
  - [6] E. Ott, *Chaos in Dynamical Systems* (Cambridge University Press, New York, 1993).
  - [7] S. H. Strogatz, *Nonlinear Dynamics and Chaos: with applications to physics, biology, chemistry, and engineering* (Addison-Wesley, Reading, Massachusetts, 1994).
  - [8] N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw, *Phys. Rev. Lett.* **45**, 712 (1980).
  - [9] F. Takens, in *Symposium on Dynamical Systems and Turbulence*, edited by D. A. Rand and L. S. Young (Springer-Verlag, Berlin, 1981), vol. 898, p. 366.
  - [10] A. Fraser, in *Information Dynamics*, edited by H. Atmanspacher and H. Scheingraber (Plenum, New York, 1991), vol. Series B: Physics Vol. 256 of *NATO ASI Series*, p. 125.
  - [11] J. P. Crutchfield and B. S. McNamara, *Complex Systems* **1**, 417 (1987).
  - [12] M. Casdagli and S. Eubank, eds., *Nonlinear Modeling, SFI Studies in the Sciences of Complexity* (Addison-Wesley, Reading, Massachusetts, 1992).
  - [13] J. C. Sprott, *Chaos and Time-Series Analysis* (Oxford University Press, Oxford, UK, 2003), 2nd ed.
  - [14] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis* (Cambridge University Press, Cambridge, UK, 2006), 2nd ed.
  - [15] J. P. Crutchfield and K. Young, *Phys. Rev. Lett.* **63**, 105 (1989).
  - [16] N. Tishby, F. Pereira, and W. Bialek, in *Proceedings of the 37th Annual Allerton Conference*, edited by B. Hajek and R. S. Sreenivas (University of Illinois, 1999), pp. 368–377.
  - [17] S. Still and W. Bialek, *Neural Computation* **16**(12), 2483 (2004).
  - [18] C. R. Shalizi and J. P. Crutchfield, *Advances in Complex Systems* **5**, 1 (2002).
  - [19] C. E. Shannon, *Bell Sys. Tech. J.* **27** (1948), reprinted in C. E. Shannon and W. Weaver *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, 1949.
  - [20] S. Still and J. P. Crutchfield (2007), arxiv.org: 0708.0654

- [physics.gen-ph].
- [21] J. P. Crutchfield, *Physica D* **75**, 11 (1994).
  - [22] J. P. Crutchfield and C. R. Shalizi, *Physical Review E* **59**, 275 (1999).
  - [23] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley-Interscience, New York, 2006), 2nd ed.
  - [24] J. P. Crutchfield, C. J. Ellison, and J. R. Mahoney, *Phys. Rev. Lett.* **103**, 094101 (2009).
  - [25] J. P. Crutchfield and D. P. Feldman, *CHAOS* **13**, 25 (2003).
  - [26] W. Bialek, R. R. de Ruyter van Steveninck, and N. Tishby, in *Proceedings of the International Symposium on Information Theory* (2006), pp. 659–663.
  - [27] A. del Junco and M. Rahe, *Proc. AMS* **75**, 259 (1979).
  - [28] J. P. Crutchfield and N. H. Packard, *Physica* **7D**, 201 (1983).
  - [29] R. Shaw, *The Dripping Faucet as a Model Chaotic System* (Aerial Press, Santa Cruz, California, 1984).
  - [30] P. Grassberger, *Intl. J. Theo. Phys.* **25**, 907 (1986).
  - [31] W. Li, *Complex Systems* **5**, 381 (1991).
  - [32] W. Bialek and N. Tishby, *Predictive information* (1999), URL [arXiv:cond-mat/9902341v1](https://arxiv.org/abs/cond-mat/9902341v1).
  - [33] J. P. Crutchfield, C. J. Ellison, J. R. Mahoney, and R. G. James, *CHAOS* p. in press (2010).
  - [34] S. Arimoto, *IEEE Transactions on Information Theory* IT-18 pp. 14–20 (1972).
  - [35] R. E. Blahut, *IEEE Transactions on Information Theory* IT-18 pp. 460–473 (1972).
  - [36] K. Rose, E. Gurewitz, and G. C. Fox, *Phys. Rev. Lett.* **65**, 945 (1990).
  - [37] K. Rose, *Proc. of the IEEE* **86**, 2210 (1998).
  - [38] C. J. Ellison, J. R. Mahoney, and J. P. Crutchfield, *J. Stat. Phys.* **136**, 1005 (2009).
  - [39] C. C. Streliaoff, J. P. Crutchfield, and A. Hübner, *Phys. Rev. E* **76**, 011106 (2007).
  - [40] L. R. Rabiner and B. H. Juang, *IEEE ASSP Magazine* **January** (1986).
  - [41] J. Rissanen, *Stochastic Complexity in Statistical Inquiry* (World Scientific, Singapore, 1989).
  - [42] S. Still, *Euro. Phys. Lett.* **85**, 28005 (2009).
  - [43] N. Ay and J. P. Crutchfield, *J. Stat. Phys.* **210**, 659 (2005).
  - [44] A more general approach is taken in Ref. [42], where both predictive modeling and decision making are considered. The scenario discussed here is a special case.
  - [45] To save space and improve readability we use a simplified notation that refers to infinite sequences of random variables. The implication, however, is that one works with finite-length sequences into the past and into the future, whose infinite-length limit is taken at appropriate points. See, for example, Ref. [22] or, for measure-theoretic foundations, Ref. [43].
  - [46] The algorithm follows that used in the information bottleneck [16]. The convergence arguments there apply to the OCF algorithm.
  - [47] All quantities denoted with a  $\hat{\cdot}$  are evaluated at the estimate  $\hat{P}$ .