

Optimal Causal Inference

Susanne Still

*Information and Computer Sciences, University of Hawaii at Manoa, Honolulu, HI 96822, USA.
sstill@hawaii.edu*

James P. Crutchfield and **Christopher J. Ellison**

*Center for Computational Science & Engineering and Physics Department,
University of California Davis, One Shields Avenue, Davis, CA 95616, USA.
chaos@cse.ucdavis.edu
cellison@cse.ucdavis.edu*

Editor: TBD

Abstract

We consider an information-theoretic objective function for statistical modeling of time series that embodies a parametrized trade-off between the predictive power of a model and the model's complexity. We study two distinct cases of optimal causal inference, which we call optimal causal filtering (OCF) and optimal causal estimation (OCE). OCF corresponds to the ideal case of having infinite data. We show that OCF leads to the exact causal architecture of a stochastic process, in the limit in which the trade-off parameter tends to zero, thereby emphasizing prediction. Specifically, the filtering method reconstructs exactly the hidden, causal states. More generally, we establish that the method leads to a graded model-complexity hierarchy of approximations to the causal architecture. We show for nonideal cases with finite data (OCE) that the correct number of states can be found by adjusting for statistical fluctuations in probability estimates.

1. 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 information sources (Guckenheimer and Holmes, 1983; Berge et al., 1986; Ott, 1993; Strogatz, 1994). 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 (Packard et al., 1980; Takens, 1981), is that discovering the hidden structure of such processes is key to successful modeling and prediction (Casdagli and Eubank, 1992; Sprott, 2003; Kantz and Schreiber, 2006).

Following these lines, here we investigate the problem of predictive modeling of time series with particular attention paid to discovering hidden variables. Using a framework (Still and Crutchfield, 2007) inspired by rate distortion theory (Shannon, 1948) and the information bottleneck method (IB) (Tishby et al., 1999) and using estimation methods discussed in Still and Bialek (2004), we analyze the trade-offs between model complexity, approximation accuracy, and the effect of finite-data fluctuations. We apply IB to time

series prediction, thereby introducing the *optimal causal filtering* (OCF) method. OCF achieves an optimal balance between model complexity and approximation. In the limit in which prediction is paramount and model complexity is not restricted, we prove that OCF reconstructs the underlying process’s causal architecture, as defined within the framework of computational mechanics (Crutchfield and Young, 1989).

This shows, in effect, that OCF captures a source’s hidden variables and structure. We then extend OCF to handle finite-data fluctuations, introducing *optimal causal estimation* (OCE). In this case, errors due to statistical fluctuations in probability estimates must be taken into account in order to avoid over-fitting. We demonstrate how the methods work on a number of example stochastic processes with known, nontrivial correlational structure.

2. Causal Models of Information Sources

Assume that we are given a *process* $P(\overleftrightarrow{X})$ —a joint distribution over a bi-infinite sequence $\overleftrightarrow{X} = \overleftarrow{X}\overrightarrow{X}$ of random variables. The *past*, or *history*, is denoted $\overleftarrow{X} = \dots X_{-3}X_{-2}X_{-1}$, while $\overrightarrow{X} = X_0X_1X_2\dots$ denotes the *future*.¹ Here the random variables X_t take on discrete values $x \in \mathcal{A}$ and the process as a whole is stationary. The following assumes the reader is familiar with information theory and the notation of Cover and Thomas (2006).

We can consider any process $P(\overleftrightarrow{X})$ to be a communication channel that transmits information from the past to the future, by storing information in the present—presumably in some internal states, variables, or degrees of freedom. 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.²

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 (Crutchfield and Young, 1989). 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 \overleftarrow{x} 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 has three aspects: a label $\sigma \in \mathcal{S}$; a set of histories $\overleftarrow{X}_\sigma = \{\overleftarrow{x} : P(\overrightarrow{X} | \overleftarrow{x}) = P(\overrightarrow{X} | \sigma)\} \subset \overleftarrow{\mathbf{X}}$; and a *future morph* which is the conditional distribution $P(\overrightarrow{X} | \sigma)$ of the futures that can be seen from the state. The causal states along with their transitions \mathcal{T} form the ϵ -*machine* representation of the process as specified within computational mechanics (Crutchfield and Young, 1989; Crutchfield, 1994; Crutchfield and Shalizi, 1999; Ay and Crutchfield, 2005).

1. 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, Crutchfield and Shalizi (1999).

2. Many authors have considered such questions; see, for example, Crutchfield and McNamara (1987) and Bialek et al. (2001). The review in Crutchfield and Feldman (2003), and references therein, gives an account of the related literature.

Any alternative model, *rival* \mathcal{R} , generally 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 (Crutchfield and Shalizi, 1999):

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

where $I[V, W]$ denotes the mutual information between random variables V and W (Cover and Thomas, 2006).³

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 morphs 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 ϵ -machine is an optimal predictor (Crutchfield and Young, 1989). That is, the causal states, considered as a random variable \mathcal{S} , capture the full *excess entropy*

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

Said differently, the ϵ -machine minimizes the uncertainty about the future compared to any rival \mathcal{R} : $H(\overrightarrow{X}|\mathcal{S}) \leq H(\overrightarrow{X}|\mathcal{R})$. *Prescient rivals* $\widehat{\mathcal{R}}$, for which the equality holds, are as predictive as the causal-state model:

$$I[\widehat{\mathcal{R}}; \overrightarrow{X}] = I[\mathcal{S}; \overrightarrow{X}] = I[\overleftarrow{X}; \overrightarrow{X}] . \quad (6)$$

Out of all prescient rivals, the ϵ -machine has the smallest *statistical complexity*, $C_\mu[\mathcal{R}] = H[\mathcal{R}]$ (Crutchfield and Young, 1989; Crutchfield and Shalizi, 1999):

$$H[\widehat{\mathcal{R}}] \geq H[\mathcal{S}] . \quad (7)$$

Importantly, the ϵ -machine allows for the exact computation of the amount of information that the process communicates from the past to the future, by storing it in the present: $C_\mu = H[\mathcal{S}]$. In addition, it allows for the exact computation of the process's *entropy rate* h_μ which, for example, gives the optimal compression rate for the source and an upper bound on its predictability (Crutchfield and Young, 1989).

The general conclusion is that the ϵ -machine captures all of a process's structure—symmetry, regularity, and organization. And so, it should be the goal of model inference.

3. The quantity $I[\overleftarrow{X}; \overrightarrow{X}]$ has been studied by several authors and, over the years, assigned various names, including *excess entropy* (Crutchfield and Packard, 1983), *stored information* (Shaw, 1984), *predictive information* (Bialek et al., 2001), and others. See Crutchfield and Feldman (2003) and references therein for a review.

3. Objective Function For Causal Filtering

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 (Shannon, 1948) 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 thrown away. Since there is no universal distortion function, it has to be assumed *ad hoc* for each application. The information bottleneck method (Tishby et al., 1999) argued for a special distortion function that allows for explicitly keeping relevant information—defined by Tishby et al. (1999) as the mutual information that the data share with a desired relevant variable. However, this relevant variable also has to be specified *ad hoc*.

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. The first step is to 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 $P(\mathcal{R}|\vec{x})$. The second step is to 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} . However, we do not want to keep all of the historical information. Instead, and this is the final step, we keep in \mathcal{R} as little of the *retrodictive information* $I[\vec{X}; \mathcal{R}]$ about the past as possible. This quantity monitors the statistical complexity, or bit cost, of the model \mathcal{R} .

Following the information bottleneck method then, the constrained optimization problem to solve to find the most predictive model at fixed model complexity $I[\vec{X}; \mathcal{R}]$ is⁴

$$\max_{P(\mathcal{R}|\vec{X})} F[\mathcal{R}] \tag{8}$$

with the objective function:

$$F[\mathcal{R}] = I[\mathcal{R}; \vec{X}] - \lambda I[\vec{X}; \mathcal{R}] , \tag{9}$$

where the parameter λ controls the balance between prediction and model complexity.

The optimization problem given by Eqs. (8) and (9) has a family of solutions, parametrized by the Lagrange multiplier λ , that gives the following optimal assignments of histories \vec{x} to states $\rho \in \mathcal{R}$:⁵

$$P_{\text{opt}}(\rho|\vec{x}) = \frac{P(\rho)}{Z(\vec{x}, \lambda)} \exp \left(-\frac{1}{\lambda} \mathcal{D} \left(P(\vec{X}|\vec{x}) || P(\vec{X}|\rho) \right) \right) \tag{10}$$

4. The approach is similar to that taken in Still and Bialek (2006), where both predictive modeling and decision making are considered. Here we focus only on predictive time series modeling.

5. Formally, the derivation follows Tishby et al. (1999).

with

$$P(\vec{X} | \rho) = \frac{1}{P(\rho)} \sum_{\overleftarrow{x} \in \overleftarrow{\mathbf{X}}} P(\vec{X} | \overleftarrow{x}) P(\rho | \overleftarrow{x}) P(\overleftarrow{x}) \quad (11)$$

$$P(\rho) = \sum_{\overleftarrow{x} \in \overleftarrow{\mathbf{X}}} P(\rho | \overleftarrow{x}) P(\overleftarrow{x}), \quad (12)$$

where $\mathcal{D}(P||Q)$ is the information gain (Cover and Thomas, 2006) between distributions P and Q . These self-consistent equations are solved iteratively (Tishby et al., 1999) using a procedure similar to the Blahut-Arimoto algorithm (Blahut, 1972; Arimoto, 1972).

Observe that in the *low-temperature regime* ($\lambda \rightarrow 0$) the assignments of pasts to states become deterministic and are given by:

$$P_{\text{opt}}(\rho | \overleftarrow{x}) = \delta_{\rho, \eta(\overleftarrow{x})}, \text{ where} \quad (13)$$

$$\eta(\overleftarrow{x}) = \arg \min_{\rho} \mathcal{D} \left(P(\vec{X} | \overleftarrow{x}) || P(\vec{X} | \rho) \right). \quad (14)$$

4. Optimal Causal Filtering

We now establish the procedure's fundamental properties by connecting the solutions it determines to the causal representations afforded by computational mechanics. The resulting procedure, in effect, 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 retrodictive information 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 *The causal-state partition of the ϵ -machine maximizes $F_{\text{det}}[\widehat{\mathcal{R}}]$.*

Proof This follows immediately from Eqs. (6) and (7). They imply that

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

■

The Lemma tells us that the causal-state partition is the fully predictive model with the largest value of the OCF objective function. We also know from Eq. (14) 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 ϵ -machine's causal-state partition of $\overleftarrow{\mathbf{X}}$ in the low-temperature limit.*

Proof Given that the ϵ -machine always exists, there are groups of histories $\overleftarrow{X}_\sigma \subset \overleftarrow{\mathbf{X}}$ such that

$$P(\overrightarrow{X} | \overleftarrow{x}) = P(\overrightarrow{X} | \sigma) , \forall \overleftarrow{x} \in \overleftarrow{X}_\sigma . \quad (17)$$

Thus,

$$\sigma = \arg \min_{\rho} \mathcal{D} \left(P(\overrightarrow{X} | \overleftarrow{x}) || P(\overrightarrow{X} | \rho) \right) , \quad (18)$$

with

$$\mathcal{D} \left(P(\overrightarrow{X} | \overleftarrow{x}) || P(\overrightarrow{X} | \sigma) \right) = 0 , \quad (19)$$

for all $\overleftarrow{x} \in \overleftarrow{X}_\sigma$. Therefore, the assignment of histories to the causal states is recovered:

$$P_{\text{opt}}(\mathcal{R} | \overleftarrow{x}) = P(\sigma | \overleftarrow{x}) = \delta_{\sigma, \epsilon(\overleftarrow{x})} . \quad (20)$$

■

Note that we have not restricted the size of the set, \mathcal{R} , of model states. Note furthermore, that the Lemma establishes that, other than the causal states, OCF does *not* find prescient rival models in the low-temperature limit. The prescient rival models are suboptimal, as they have a smaller value of the objective function than the causal states. We now establish that this difference is controlled by the model size with proportionality constant λ .

Corollary 1 *Non-causal-state, prescient rival models are suboptimal in OCF. The value of the objective function evaluated for a prescient rival model 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*

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

Proof

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

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

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

Moreover, Eq. (7) implies that $\Delta F_{\text{det}} \geq 0$. ■

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. However, in the *very low* temperature regime, as $\lambda \rightarrow 0$, OCF finds the deterministic assignment rule which recovers the causal-state partition of the ϵ -machine.

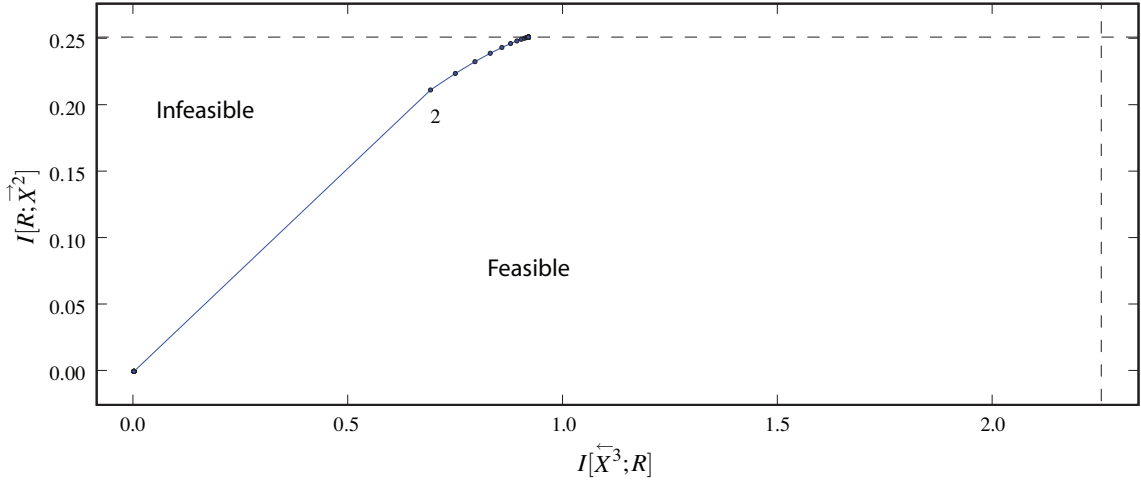


Figure 1: OCF’s behavior as monitored in the information plane— $I[\mathcal{R}; \vec{X}^2]$ versus $I[\vec{X}^3; \mathcal{R}]$ —for the Golden Mean process. Histories of length 3 were used, along with futures of length 2. The horizontal dashed line is the full excess entropy $I[\vec{X}^3; \vec{X}^2] = I[\sigma; \vec{X}^2] \approx 0.25$ bits which, as seen, is an upper bound on $I[\mathcal{R}; \vec{X}^2]$. 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 annealing rate was 0.952. In this and the following information plane plots the integer labels $N_c(\geq 2)$ indicate the first point at which the effective number of states used by the model equals N_c .

5. Examples

We study how OCF works on a series of example stochastic processes of increasing statistical sophistication. In examples, naturally, one can only work with finite length sequences. The results established in the previous section hold for finite histories and finite futures; denote their lengths as t_p and t_f , respectively. 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: For each value of λ , we evaluate both terms in the objective function at the optimal solution and plot them against each other. The resulting curve in the information plane separates the feasible from the infeasible region: we can find a model that is more complex at the same prediction error, but we cannot 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.

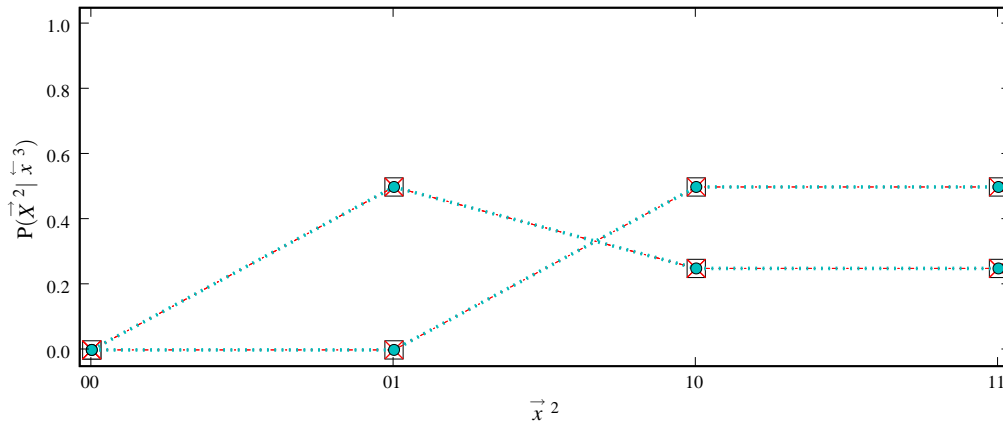


Figure 2: Golden Mean process future morphs $P(\vec{X}^2 | \cdot)$ conditioned on causal states $\sigma \in \mathcal{S}$ (boxes) and on the OCF reconstructed states $\rho \in \mathcal{R}$ (circles). As an input to OCF, morphs $P(\vec{X}^2 | \vec{x}^3)$ calculated from histories of length 3 were used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

The examples in this and the following sections are calculated by solving the self-consistent Eqs. (10) to (12) iteratively⁶ at each value of λ . To lower λ a deterministic annealing scheme is implemented following Rose (1998).

5.1 Golden Mean Process: 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., Young and Crutchfield (1994).

Figures 1 and 2 demonstrate how OCF reconstructs the states of the GM process. Figure 1 shows the behavior of OCF in the information plane. The curve seen there is traced out as λ decreases from high to low temperature during annealing. 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 little or no information about the future and so is a poor predictor. As λ decreases, 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 this increase. As $\lambda \rightarrow 0$, prediction comes to dominate and OCF finds a fully predictive model, albeit one

6. The algorithm follows that used in the information bottleneck (Tishby et al., 1999). The convergence arguments there apply to the OCF algorithm.

with the minimal statistical complexity, out of all possible state partitions that would retain the full predictive information. The model’s complexity is 41% of the maximum, which is given by $H[\vec{X}^3] \approx 2.25$ bits.

Figure 2 shows the future morphs, associated with the partition found by OCF, as $\lambda \rightarrow 0$, corresponding to $P(\vec{X}^2 | \rho)$ (circles), compared to true (but not known to the algorithm) causal-state future morphs $P(\vec{X}^2 | \sigma)$ (boxes). These morphs overlap and so demonstrate that OCF finds the causal-state partition as $\lambda \rightarrow 0$.

5.2 Even Process: Hidden Markov Chain

Now consider a hidden Markov process: the *Even process* (Crutchfield, 1992; Crutchfield and Feldman, 2003), which is a stochastic process whose support (the set of allowed sequences) is a symbolic dynamical system called the *Even system* (Weiss, 1973). 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 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 (Weiss, 1973). 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 λ , OCF applied to the Even process recovers a small, one-state model with poor predictive quality (Fig. 3). As λ decreases there are transitions to larger models that capture increasingly more information about the future. (The numbers along the curve indicate the transitions to more states.) With a three-state model OCF captures the full excess entropy at a model size of 56% of the maximum.

This model is exactly the causal-state partition, as can be seen in Fig. 4 by comparing the future morphs of the OCF model (circles) to the true underlying causal states (boxes), which are not known to the algorithm.

6. Trading model size against prediction error

It is now clear that the Lagrange multiplier λ controls the trade-off between the amount of structure that the model captures of the past data, on the one hand, and the model’s predictive power, on the other. That is, it controls how much of the future fluctuations the model considers to be random; i.e., which fluctuations are indistinguishable. We just showed that 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.

What happens, though, if we allow (or prefer) a model with some finite prediction error? Can we make the model substantially smaller? Is there some systematic ordering of models of different size and different predictive power given by OCF, as we change the parameter λ ? Naturally, the trade-off depends on, and even reflects, the source’s organization.

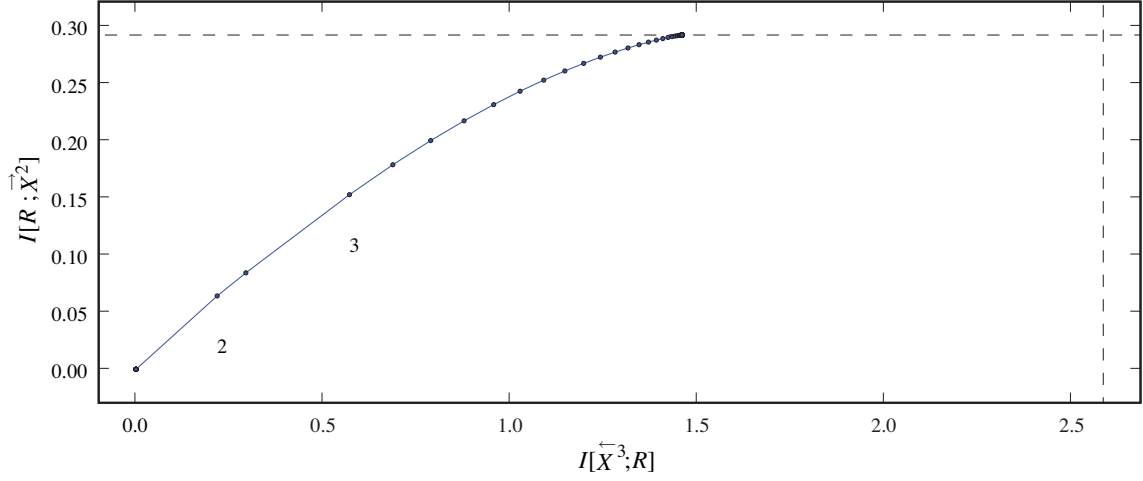


Figure 3: OCF’s behavior inferring the Even process: monitored in the information plane— $I[\mathcal{R}; \vec{X}^2]$ versus $I[\overleftarrow{X}^3; \mathcal{R}]$. Histories of length 3 were used, along with futures of length 2. The horizontal dashed line is the full excess entropy $I[\overleftarrow{X}^3; \vec{X}^2] \approx 0.292$ 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 $H[\overleftarrow{X}^3] \approx 2.585$ bits which is an upper bound on the retrodictive information $I[\overleftarrow{X}^3; \mathcal{R}]$.

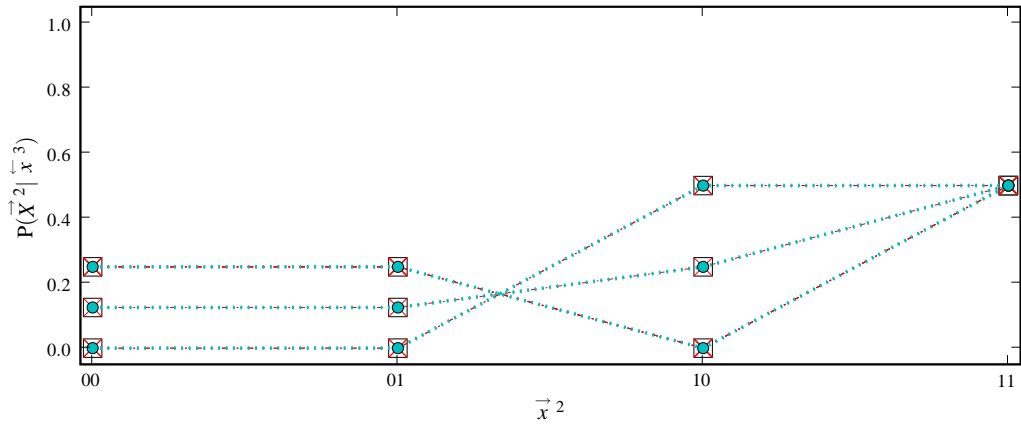


Figure 4: Even process future morphs $P(\vec{X}^2 | \cdot)$ conditioned on causal states $\sigma \in \mathcal{S}$ (boxes) and on the OCF-reconstructed states $\rho \in \mathcal{R}$ (circles). As an input to OCF, morphs $P(\vec{X}^2 | \overleftarrow{x}^3)$ calculated from histories of length 3 were used (crosses).

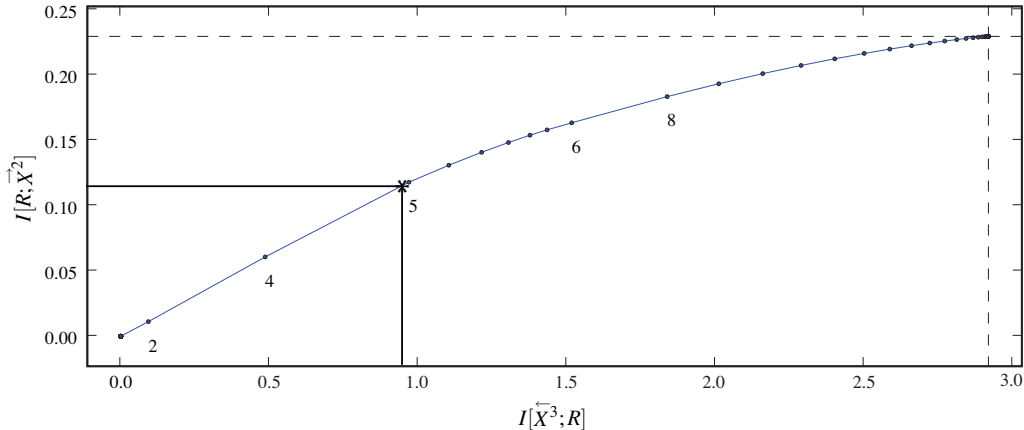


Figure 5: Structure versus prediction trade-off under OCF for the random-random XOR (RRXOR) process, as monitored in the information plane. As above, the horizontal dashed line is full excess entropy (≈ 0.230 bits) and the vertical dashed line is the block entropy (≈ 2.981 bits). Histories of length 3 were used, along with futures of length 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 morphs of which are depicted in Fig. 7.

6.1 Random Random XOR: A structurally complex process

As an example, we use OCF to model 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 states and has a very large number of causal states (36 for semi-infinite past and future), most of which describe a complicated transient structure (Crutchfield and Feldman, 2003). As such it is a very structurally complex process that an analyst may wish to approximate with a smaller set of states.

Figure 5 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 (denoted λ) increases with model complexity. At a history length of 3 and future length of 2, the process has eight underlying causal states, which are found by OCF in the $\lambda \rightarrow 0$ limit. The corresponding future morphs are shown in Fig. 6.

However, the RRXOR process has a structure that does not allow for substantial compression. The statistical complexity of the causal-state partition is equal to the full entropy $C_\mu[\mathcal{S}] = H[\overleftarrow{X}^3]$. With half of the number of states (4), however, OCF reconstructs a model that is only 33% as large, while capturing 50% of the information about the future. The corresponding morphs of the (best) four-state model are shown in Fig. 6. They are mixtures of pairs of the eight causal states.

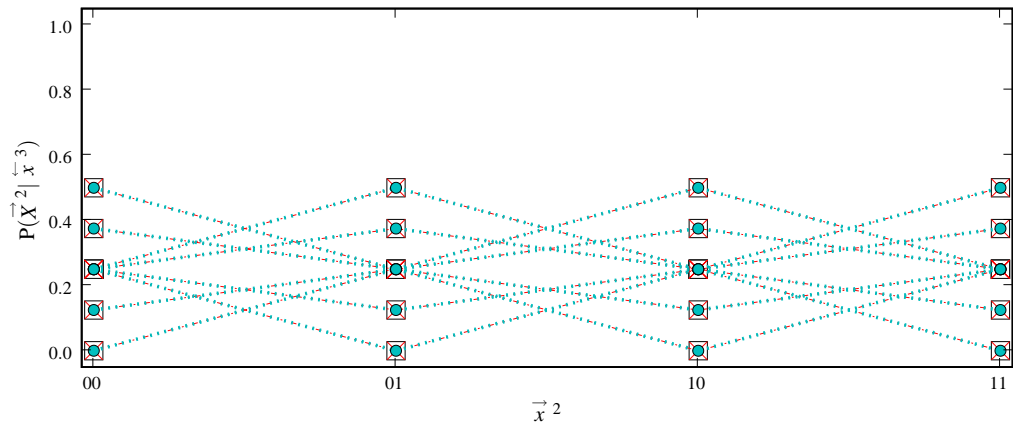


Figure 6: Morphs $P(\vec{X}^2 | \cdot)$ for the RRXOR process: the 8-state approximation (circles) finds the causal states (boxes). Histories of length 3 were used, along with futures of length 2.

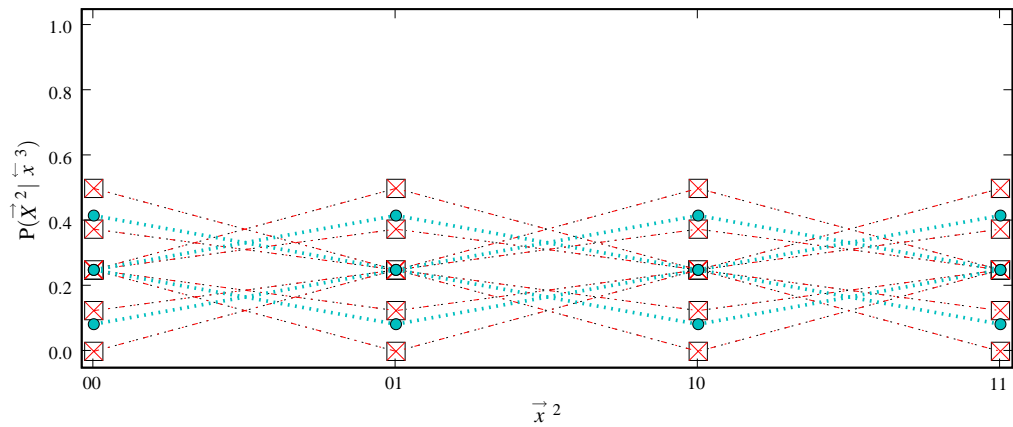


Figure 7: Morphs $P(\vec{X}^2 | \cdot)$ for the RRXOR process: the 4-state approximation (circles) compared to causal states (boxes). Histories of length 3 were used, along with futures of length 2.

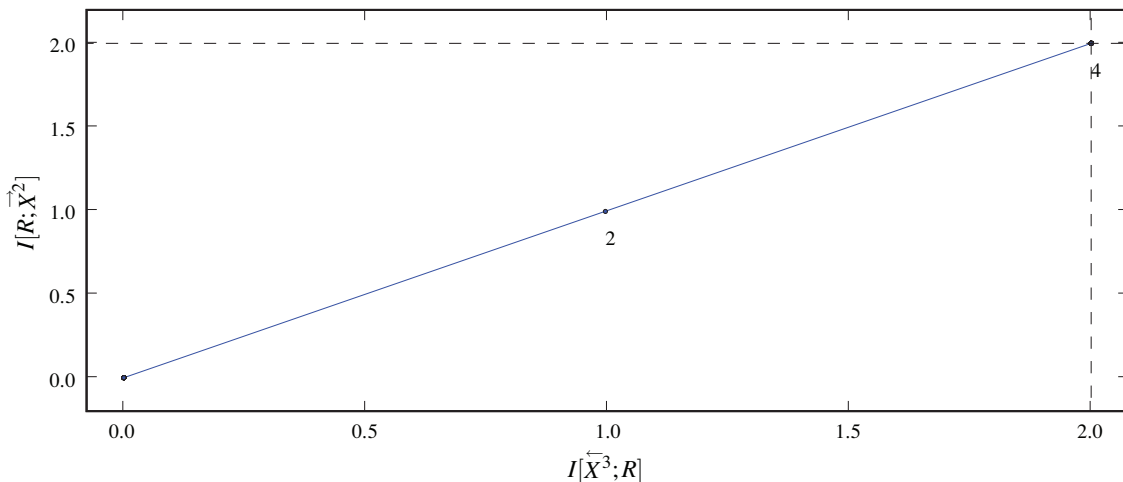


Figure 8: Structure versus prediction trade-off under OCF for the exactly predictable period-4 process: $(0011)^\infty$. Monitored in the information plane. As above, the horizontal dashed line is the full excess entropy (2 bits) and the vertical dashed line is the block entropy (2 bits). Histories of length 3 were used, along with futures of length 2.

The information curve informs us about the (best possible) ratio of predictive power to model complexity: $I[\mathcal{R}; \overrightarrow{X}]/I[\overleftarrow{X}; \mathcal{R}]$. 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 compared to the full causal-state model, then we can read off the curve that this can be achieved at 70% of the model complexity.

6.2 Periodic limit cycle: A predictable process

The RRXOR is a highly complex stochastic process. At the other end of the predictability spectrum are the exactly periodic processes produced by, for example, limit cycle oscillations. Figure 8 shows how OCF works on a period-four process: $(0011)^\infty$. There are exactly two bits of information to be captured about future words of length two. This information describes the phase of the period-four cycle. To capture those two bits, we need exactly four underlying causal states and two bits. However, if we compress to one bit (using two states), we can only capture one bit of excess entropy. The information curve falls onto the diagonal—a straight line which is the worst case for possible beneficial trade-offs between prediction error and model complexity. Processes of this type cause the algorithm to implement quenching, similar to what happens in Still et al. (2004). This interesting fact, and the linear interdependence, which can be showed for periodic processes, will be analyzed elsewhere.

In Fig. 9, we show the best two-state model compared to the full (exact) four-state model. Each of the two model future morphs captures zero probability events of odd

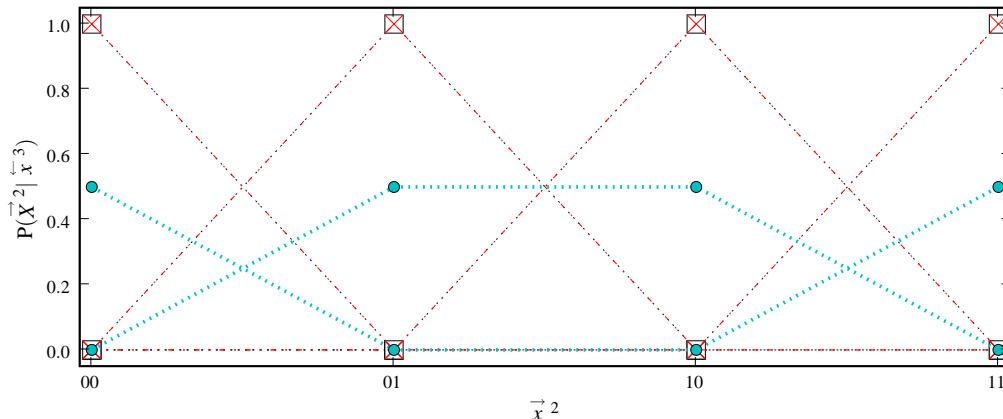


Figure 9: Morphs $P(\vec{X}^2 | \cdot)$ for the period-4 process: the 2-state approximation (circles) compared to the δ -function morphs for the 4 causal states (boxes). Histories of length 3 were used, along with futures of length 2 (crosses).

$\{01, 10\}$ or even $\{00, 11\}$ words, assigning equal probability to the even or odd words, respectively. This captures the fundamental determinism of the process: an odd word never follows an even word and vice versa. The overall result, though, illustrates how the actual long-range correlation in the completely predictable period-4 sequence is represented by a smaller *stochastic* model. In the four-state model the morphs are δ -functions. In the two-state approximate models, they are mixtures of those δ -functions. In this way, OCF converts structure to randomness when approximating.

7. Optimal Causal Estimation: Finite-data fluctuations

In real world applications, we do not know a process's underlying probability density, but instead we have to estimate it from the *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 t_p) and futures (of length t_f) via a histogram calculated using a sliding window. Altogether we have $M = T - (t_p + t_f - 1)$ observations.

The resulting estimate $\hat{P}(\vec{X}^{t_p}; \vec{X}^{t_f})$ will deviate from the true $P(\vec{X}^{t_p}; \vec{X}^{t_f})$ by $\Delta(\vec{X}^{t_p}, \vec{X}^{t_f})$. This leads to an overestimate of the mutual information:⁷ $\hat{I}[\vec{X}^{t_p}; \vec{X}^{t_f}] \geq I[\vec{X}^{t_p}; \vec{X}^{t_f}]$. Evaluating the objective function at this estimate may lead to model variations that are due to the sampling noise and not to the process's underlying structure; i.e., OCF may overfit. That is, the process may appear to have a larger number N_c of causal states than the true number.

7. All quantities denoted with a $\hat{\cdot}$ are evaluated at the estimate \hat{P} .

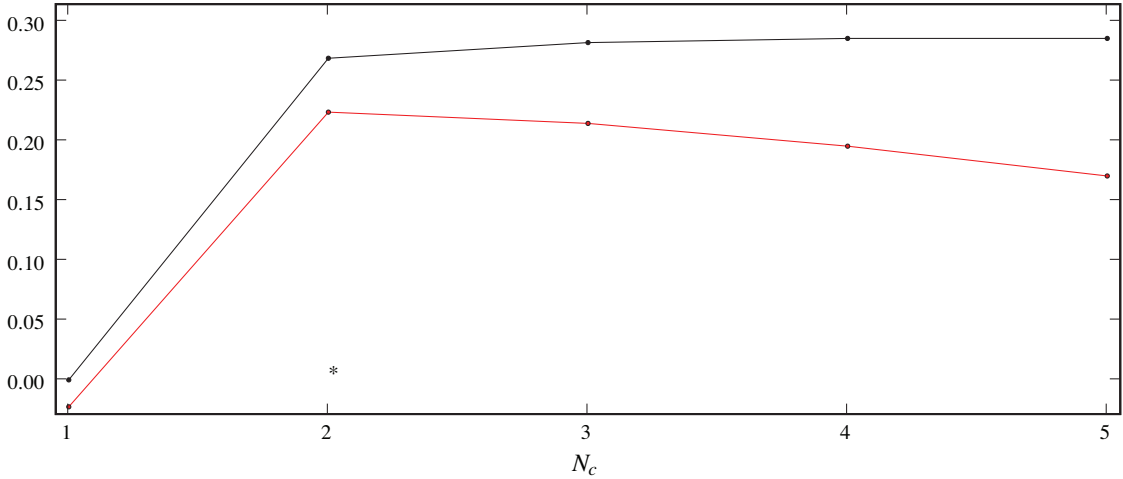


Figure 10: Information captured about the future versus the number N_c of reconstructed states, with statistics estimated from length $T = 100$ time series sample from the Golden Mean process. Upper line: $\widehat{I}[\mathcal{R}; \vec{X}^2]_{\lambda \rightarrow 0}$, not corrected; lower line: $\widehat{I}[\mathcal{R}; \vec{X}^2]_{\lambda \rightarrow 0}^{\text{corrected}}$, corrected for estimation error due to finite sample size. Histories of length 3 and futures of length 2 were used. The asterisk denotes the optimal number of effective states.

Following Still and Bialek (2004), we argue that this effect can be counteracted by subtracting from $\widehat{F}[\mathcal{R}]$ a model-complexity control term that approximates the error we make by calculating the estimate $\widehat{F}[\mathcal{R}]$ rather than the true $F[\mathcal{R}]$. If we are willing to assume that M is large enough, so that the deviation $\Delta(\vec{X}^{t_p}, \vec{X}^{t_f})$ is a small perturbation, then the error can be approximated by (Still and Bialek, 2004, Eq. 5.8):

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

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

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

with

$$\widehat{I}[\vec{X}^{t_p}; \vec{X}^{t_f}]_{\lambda \rightarrow 0}^{\text{corrected}}(N_c) = \left(\widehat{I}[\vec{X}^{t_p}; \vec{X}^{t_f}]_{\lambda \rightarrow 0}(N_c) - \mathcal{E}(N_c) \right). \quad (27)$$

With this correction we generalize OCF to a new procedure—*optimal causal estimation* (OCE)—that simultaneously accounts for the trade-off between structure, approximation, and statistical fluctuations.

We illustrate OCE on the Golden Mean and Even processes studied in Sec. 5. Figures 10 and 12 show the mutual information $I[\mathcal{R}; \vec{X}^2]$ versus the number N_c of inferred states,

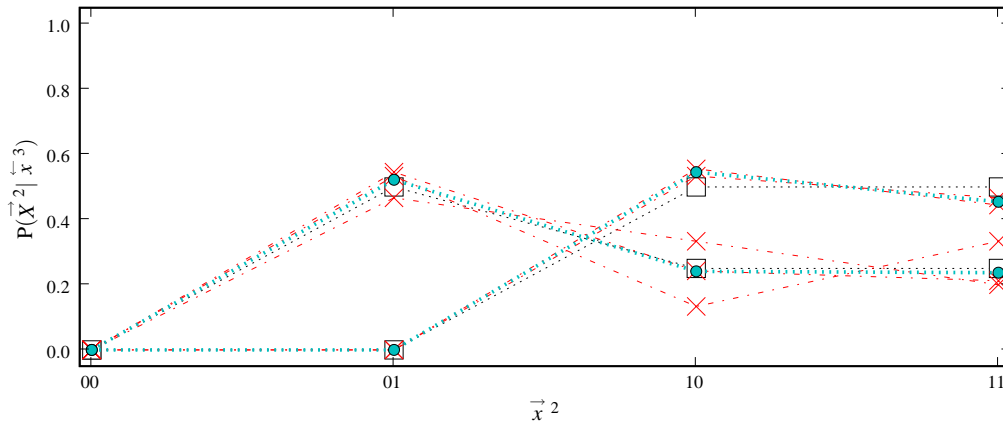


Figure 11: OCE’s best two-state approximated morphs (circles) estimated from a sample time series of length $T = 100$ from the Golden Mean process. Compared to true morphs (squares) from the ϵ -machine. The OCE inputs are the estimates of $\hat{P}(\vec{X}^2 | \overleftarrow{x}^3)$ (crosses).

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 $\hat{P}(\vec{X}^2; \overleftarrow{x}^3)$ (upper curve) to the corrected information $\hat{I}[\mathcal{R}; \vec{X}^2]_{\lambda \rightarrow 0}^{\text{corrected}}$ calculated by subtracting the approximated error Eq. (25) with $K = 4$ and $M = 96$ (lower curve).

We see that the corrected information curves peak at, and thereby select models with, two and three states, respectively. This corresponds with the true number of causal states, as we know from above (Sec. 5) for the two processes. Figures 11 and 13 show the OCE morphs corresponding to the (optimal) two- and three-state approximations, respectively. The input to OCE are the morphs given the histories $\hat{P}(\vec{X}^2 | \overleftarrow{x}^3)$ (crosses), which are estimated from the full historical information. Those morphs are corrupted by sampling errors due to the finite data set size and differ from the true morphs (squares).

Compare the OCE output morphs (circles) to the true morphs (squares), calculated with the knowledge of the ϵ -machine. (The latter, of course, is not available to the OCE algorithm.) In the case of the GM process, the OCE output approximates the correct morphs. For the Even process there is more spread in the estimated OCE output morphs. Nonetheless, OCE reduced the fluctuations in its inputs and corrected in the direction of the true underlying morphs.

8. 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

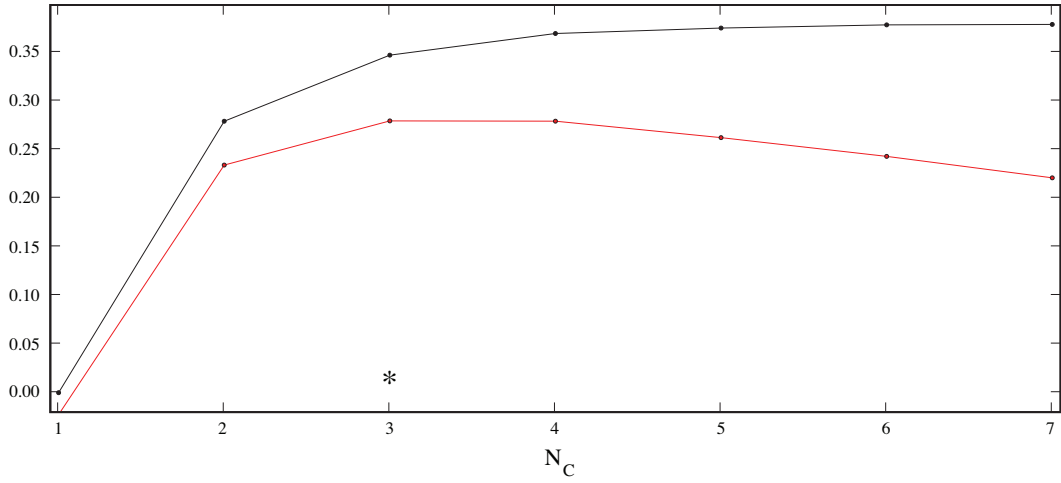


Figure 12: Information about the future versus the number N_c of reconstructed states, with statistics estimated from length $T = 100$ time series sample from the Even process. Upper line: $\hat{I}[\mathcal{R}; \vec{X}^2]_{\lambda \rightarrow 0}$, not corrected; lower line: $\hat{I}[\mathcal{R}; \vec{X}^2]_{\lambda \rightarrow 0}^{\text{corrected}}(N_c)$, corrected for estimation error due to finite sample size. The asterisk denotes the optimal number of effective states.

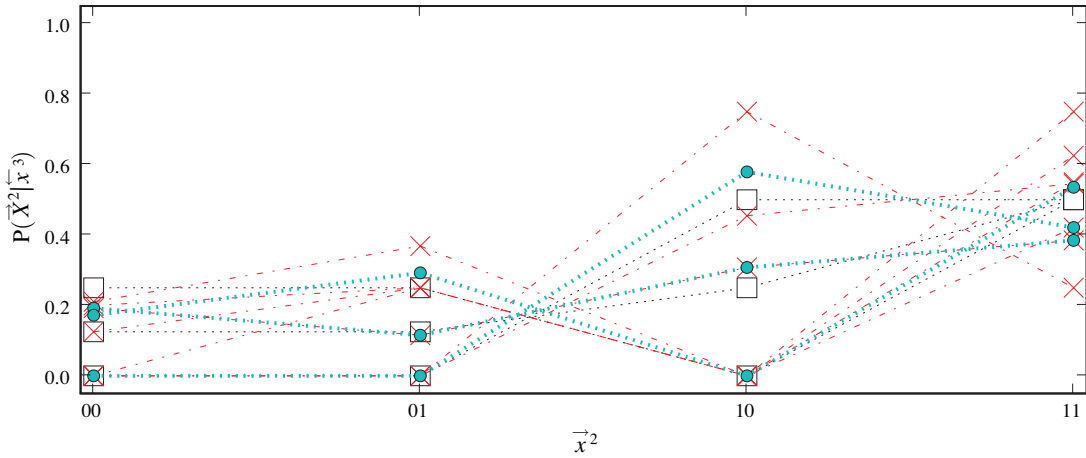


Figure 13: OCE's best three-state approximated morphs (circles) estimated from a sample time series of length $T = 100$ from the Even process. Compared to true morphs (squares) from the ϵ -machine. The OCE input are the estimates of $\hat{P}(\vec{X}^2 | \vec{x}^3)$ (crosses).

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 minimizes prediction error at minimal model complexity. The resulting principle of using small, predictive models follows from minimal prior knowledge that, in particular, makes little or no structural assumptions about a process’s architecture.

OCF stands in contrast with other approaches. Hidden Markov modeling, for example, assumes a set of states and an architecture. 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 set size (Rissanen, 1989), as happens even when modeling the ideal random process of a fair coin. OCF, however, finds the simplest (smallest) models.

The main result is that we proved that OCF reconstructs the causal-state representation of computational mechanics. This is important as it gives a structural meaning to the solutions of the optimization procedure specified by the causal inference objective function. We also gave quantitative comparisons to known cases.

We showed that 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 states. One benefit, not immediately apparent in the examples illustrating OCE, is that correct causal modeling using OCE employs rather small data sets. Why this is so will be addressed elsewhere.

Altogether, this allows us to go beyond plausibility arguments for the information-theoretic objective function. Rather, we showed that this particular way of (mathematically) phrasing the causal inference problem results in a representation that is a sufficient statistic and minimal. Moreover, it reflects the structure of the underlying process that generated the data and does so in a way that is meaningful and well grounded in physics and nonlinear dynamics—in particular, in computational mechanics. The optimal solutions to balancing prediction and model complexity take on meaning—they are the causal states of the ϵ -machine or coarse-grained mixtures of them. Additionally, the continuous trade-off allows us to go beyond the purely deterministic history-state assignments of computational mechanics, by giving a principled way of constructing stochastic approximations of the ideal causal states. The resulting approximated models can be substantially smaller and so will be useful in a number of applications.

ACKNOWLEDGMENTS

UCD and the Santa Fe Institute partially supported this work via the Network Dynamics Program funded by Intel Corporation. CJE is supported by a Department of Education GAANN graduate fellowship. S. Still thanks W. Bialek for discussions that shaped the ideas discussed in Sec. 3 and thanks I. Nemenman for helpful discussions.

References

S. Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory IT-18*, pages 14–20, 1972.

- N. Ay and J. P. Crutchfield. Reductions of hidden information sources. *J. Stat. Phys.*, 210 (3-4):659–684, 2005.
- P. Berge, Y. Pomeau, and C. Vidal. *Order within chaos*. Wiley, New York, 1986.
- W. Bialek, I. Nemenman, and N. Tishby. Predictability, complexity, and learning. *Neural Computation*, 13:2409–2463, 2001.
- R. E. Blahut. Computation of channel capacity and rate distortion function. *IEEE Transactions on Information Theory IT-18*, pages 460–473, 1972.
- M. Casdagli and S. Eubank, editors. *Nonlinear Modeling*, SFI Studies in the Sciences of Complexity, Reading, Massachusetts, 1992. Addison-Wesley.
- T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, second edition, 2006.
- J. P. Crutchfield. Semantics and thermodynamics. In M. Casdagli and S. Eubank, editors, *Nonlinear Modeling and Forecasting*, volume XII of *Santa Fe Institute Studies in the Sciences of Complexity*, pages 317 – 359, Reading, Massachusetts, 1992. Addison-Wesley.
- J. P. Crutchfield. The calculi of emergence: Computation, dynamics, and induction. *Physica D*, 75:11–54, 1994.
- J. P. Crutchfield and D. P. Feldman. Regularities unseen, randomness observed: Levels of entropy convergence. *CHAOS*, 13(1):25–54, 2003.
- J. P. Crutchfield and B. S. McNamara. Equations of motion from a data series. *Complex Systems*, 1:417 – 452, 1987.
- J. P. Crutchfield and N. H. Packard. Symbolic dynamics of noisy chaos. *Physica*, 7D:201 – 223, 1983.
- J. P. Crutchfield and C. R. Shalizi. Thermodynamic depth of causal states: Objective complexity via minimal representations. *Physical Review E*, 59(1):275–283, 1999.
- J. P. Crutchfield and K. Young. Inferring statistical complexity. *Phys. Rev. Lett.*, 63:105–108, 1989.
- J. Guckenheimer and P. Holmes. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*. Springer-Verlag, New York, 1983.
- H. Kantz and T. Schreiber. *Nonlinear Time Series Analysis*. Cambridge University Press, Cambridge, UK, second edition, 2006.
- E. Ott. *Chaos in Dynamical Systems*. Cambridge University Press, New York, 1993.
- N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw. Geometry from a time series. *Phys. Rev. Lett.*, 45:712, 1980.
- J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific, Singapore, 1989.

- K. Rose. Deterministic Annealing for Clustering, Compression, Classification, Regression, and Related Optimization Problems. *Proc. of the IEEE*, 86(11):2210–2239, 1998.
- C. E. Shannon. A mathematical theory of communication. *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.
- R. Shaw. *The Dripping Faucet as a Model Chaotic System*. Aerial Press, Santa Cruz, California, 1984.
- J. C. Sprott. *Chaos and Time-Series Analysis*. Oxford University Press, Oxford, UK, second edition, 2003.
- S. Still and W. Bialek. Active learning and optimal predictions. Technical Report UH-ICS-MLL-06-06, University of Hawaii, Manoa, Honolulu, USA, 2006.
- S. Still and W. Bialek. How many clusters? An information theoretic perspective. *Neural Computation*, 16(12):2483–2506, 2004.
- S. Still and J. P. Crutchfield. Structure or noise? Santa Fe Institute Working Paper 2007-08-020; arxiv.org physics.gen-ph/0708.0654, 2007.
- S. Still, W. Bialek, and L. Bottou. Geometric clustering using the information bottleneck method. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, 2004.
- S. H. Strogatz. *Nonlinear Dynamics and Chaos: with applications to physics, biology, chemistry, and engineering*. Addison-Wesley, Reading, Massachusetts, 1994.
- F. Takens. Detecting strange attractors in fluid turbulence. In D. A. Rand and L. S. Young, editors, *Symposium on Dynamical Systems and Turbulence*, volume 898, page 366, Berlin, 1981. Springer-Verlag.
- N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. In B. Hajek and R. S. Sreenivas, editors, *Proceedings of the 37th Annual Allerton Conference*, pages 368–377. University of Illinois, 1999. Available at <http://xxx.arXiv.cornell.edu/abs/physics/0004057>.
- B. Weiss. Subshifts of finite type and sofic systems. *Monastsh. Math.*, 77:462, 1973.
- K. Young and J. P. Crutchfield. Fluctuation spectroscopy. *Chaos, Solitons, and Fractals*, 4:5 – 39, 1994.