

# Optimally Predictive 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**

*Complexity Sciences Center and Physics Department,  
University of California at Davis  
One Shields Avenue, Davis, CA 95616, USA.*

CHAOS@CSE.UCDAVIS.EDU  
CELLISON@CSE.UCDAVIS.EDU

**Editor:** TBD

## Abstract

Natural systems compute intrinsically and produce information. The organization of a stochastic dynamical system is reflected in the time series of observations made of the system and can be quantified by the *excess entropy* or *predictive information*—the mutual information between past and future. This information can be used to build models of varying complexity that capture the causal structure of the underlying system. Here we study two distinct cases of causal inference, which we call optimal causal filtering and optimal causal estimation. Optimal causal filtering corresponds to the ideal case in which infinite data are available. We show that, in the limit in which a model complexity constraint is relaxed, the filtering method finds the causal architecture of a stochastic dynamical system, known as the *causal state partition*. In that limit, it reconstructs exactly the system’s hidden, causal states. More generally, it finds a graded model-complexity hierarchy of approximations to the causal architecture. 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.

## 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 (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), is that discovering the hidden structure of such processes is key to successful modeling and prediction (Kantz and Schreiber, 2006).

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) (Tishby et al., 1999) together with a

complexity control method discussed by Still and Bialek (2004), which is necessary for learning from finite data.

Directly adapting IB to time series prediction results in *optimal causal filtering* (OCF).<sup>1</sup> Since OCF is in essence rate-distortion theory (Shannon, 1948), 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 by us (Still and Crutchfield, 2007).

Here 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 (Crutchfield and Young, 1989). 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.

## 2. 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  $\overrightarrow{X} = X_0X_1X_2\dots$  denotes the *future*.<sup>2</sup> 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).

Any process  $P(\vec{X})$  can be considered 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. Many authors have considered such questions. The review in Crutchfield and Feldman (2003), and references therein, gives an account 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 (Crutchfield and Young, 1989). The equivalence relation partitions the space  $\overleftarrow{\mathbf{X}}$  of histories and is

---

1. A more general approach is taken in Still (2009), where both predictive modeling and decision making are considered. The scenario discussed here is a special case. A further restriction to optimally predictive *linear* filters is discussed in Creuzig (2008).

2. 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).

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 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 (Crutchfield and Young, 1989; Crutchfield and Shalizi, 1999).

Any alternative model, called *rival*  $\mathcal{R}$  in Crutchfield and Shalizi (1999), 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  $W$  (Cover and Thomas, 2006). The quantity  $I[\overleftarrow{X}; \overrightarrow{X}]$  has been studied by several authors and given different names, such as excess entropy (Crutchfield and Packard, 1983), effective measure complexity (Grassberger, 1986), and predictive information (Bialek and Tishby, 1999), amongst others. For a review see Crutchfield and Feldman (2003) 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

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

The causal state partition has, out of all *equally* predictive partitions, called *prescient rivals*  $\widehat{\mathcal{R}}$ , the smallest entropy,  $C_\mu[\mathcal{R}] = H[\mathcal{R}]$  (Crutchfield and Young, 1989; Crutchfield and Shalizi, 1999):

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

known as the *statistical complexity*,  $C_\mu := H[\mathcal{S}]$ . This is the amount of information that the process communicates from the past to the future, by storing it in the present. Altogether, the causal states are *unique and minimal sufficient statistics* for prediction of the time series.

### 3. 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 (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 discarded. 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) argues for explicitly keeping the relevant information, defined as the mutual information that the data share with a desired relevant variable (Tishby et al., 1999). Now, 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.<sup>3</sup> 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  $\overleftarrow{\mathbf{X}}$  via a probabilistic assignment rule  $P(\mathcal{R}|\overleftarrow{x})$ . Second, require that this partition be maximally predictive. That is, it should maximize the information  $I[\mathcal{R}; \overrightarrow{X}]$  that the variables  $\mathcal{R}$  contain about the future  $\overrightarrow{X}$ . 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— $I[\overleftarrow{X}; \mathcal{R}]$ , the coding rate—measures the complexity, or bit cost, of the model. Intuitively, one wants to find the most predictive model at fixed complexity or, vice versa, the least complex model at fixed prediction accuracy.

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:

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

where the parameter  $\lambda$  controls the balance between prediction and model complexity. The optimization problem Eq. (7) has a family of solutions (Tishby et al., 1999), parametrized by the Lagrange multiplier  $\lambda$ , that gives the following optimal assignments of histories  $\overleftarrow{x}$  to states  $\rho \in \mathcal{R}$ :<sup>4</sup>

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

---

3. See, e.g., Bialek (2001).

4. The derivation follows Tishby et al. (1999).

with

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

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

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 (Arimoto, 1972; Blahut, 1972). 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., Rose et al. (1990). This is useful to guide intuition and, for example, has inspired *deterministic annealing* (Rose, 1998).

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

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

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

To see this, define the quantity  $D(\rho) = \mathcal{D} \left( P(\vec{X} | \overleftarrow{x}) || P(\vec{X} | \rho) \right) - \mathcal{D} \left( P(\vec{X} | \overleftarrow{x}) || P(\vec{X} | \eta(\overleftarrow{x})) \right)$ , which is positive, by definition of  $\eta(\overleftarrow{x})$ , Eq. (12). Now, write

$$P_{\text{opt}}(\eta(\overleftarrow{x}) | \overleftarrow{x}) = \left( 1 + \sum_{\rho \neq \eta(\overleftarrow{x})} \frac{P(\rho)}{P(\eta(\overleftarrow{x}))} \exp \left[ -\frac{D(\rho)}{\lambda} \right] \right)^{-1}. \quad (13)$$

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.

#### 4. 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

$$F_{\text{det}}[\mathcal{R}] = I[\mathcal{R}; \overleftarrow{X}] - \lambda H[\mathcal{R}]. \quad (14)$$

**Lemma 1** *The causal-state partition maximizes  $F_{\text{det}}[\widehat{\mathcal{R}}]$ .*

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

$$\begin{aligned}
F_{\text{det}}[\widehat{\mathcal{R}}] &= I[\mathcal{S}; \vec{X}] - \lambda H[\widehat{\mathcal{R}}] \\
&\leq I[\mathcal{S}; \vec{X}] - \lambda H[\mathcal{S}] \\
&= F_{\text{det}}[\mathcal{S}] .
\end{aligned} \tag{15}$$

■

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 Eq. (12) 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  $\vec{X}$  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

$$P(\vec{X} | \overleftarrow{x}) = P(\vec{X} | \epsilon(\overleftarrow{x})) . \tag{16}$$

We thus have,  $\forall \overleftarrow{x} \in \vec{X}$ ,

$$\mathcal{D}\left(P(\vec{X} | \overleftarrow{x}) || P(\vec{X} | \epsilon(\overleftarrow{x}))\right) = 0 , \tag{17}$$

and hence

$$\epsilon(\overleftarrow{x}) = \arg \min_{\rho} \mathcal{D}\left(P(\vec{X} | \overleftarrow{x}) || P(\vec{X} | \rho)\right) . \tag{18}$$

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

$$P_{\text{opt}}(\rho | \overleftarrow{x}) = \delta_{\rho, \epsilon(\overleftarrow{x})} . \tag{19}$$

■

Note that we have not restricted the size of the set,  $\mathcal{R}$ , of model states. Note also that the causal state partition is *unique* (Crutchfield and Shalizi, 1999). The Lemma establishes that 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. We now establish that this difference is controlled by the model size with proportionality constant  $\lambda$ .

**Corollary 1** *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*

$$\Delta F_{\text{det}}[\widehat{\mathcal{R}}] = \lambda \left( C_{\mu}[\widehat{\mathcal{R}}] - C_{\mu}[\mathcal{S}] \right) \geq 0 . \tag{20}$$

## Proof

$$\Delta F_{\text{det}}[\widehat{\mathcal{R}}] = F_{\text{det}}[\mathcal{S}] - F_{\text{det}}[\widehat{\mathcal{R}}] \quad (21)$$

$$= I[\mathcal{S}; \vec{X}] - I[\widehat{\mathcal{R}}; \vec{X}] - \lambda H[\mathcal{S}] + \lambda H[\widehat{\mathcal{R}}] \quad (22)$$

$$= \lambda \left( C_{\mu}[\widehat{\mathcal{R}}] - C_{\mu}[\mathcal{S}] \right) . \quad (23)$$

Moreover, Eq. (6) 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. This means that maximizing the predictive information alone, without the appropriate constraint on model complexity does not suffice to recover the causal state partition.

## 5. 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: For each value of  $\lambda$ , 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: 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<sup>5</sup> at each value of  $\lambda$ . To trace out the curves, a deterministic annealing (Rose, 1998) 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 sequences,  $\vec{x}^{t_p}$  and  $\vec{x}^{t_f}$ ; denote their lengths as  $t_p$  and  $t_f$ , respectively. The results established in the previous section hold for finite histories and finite futures.

### 5.1 Periodic limit cycle: A predictable process

We start with an example of an exactly periodic process, a limit cycle oscillation. These processes are not *causally compressible* (Still and Crutchfield, 2007), and their rate-distortion curve can be computed analytically—it lies on the diagonal (Still and Crutchfield, 2007). We demonstrate this here with a numerical example. Figure 1 shows how OCF works on a period-four process:  $(0011)^\infty$ . There are exactly two bits of predictive information to be captured about future words of length two (dotted horizontal line). This information

---

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

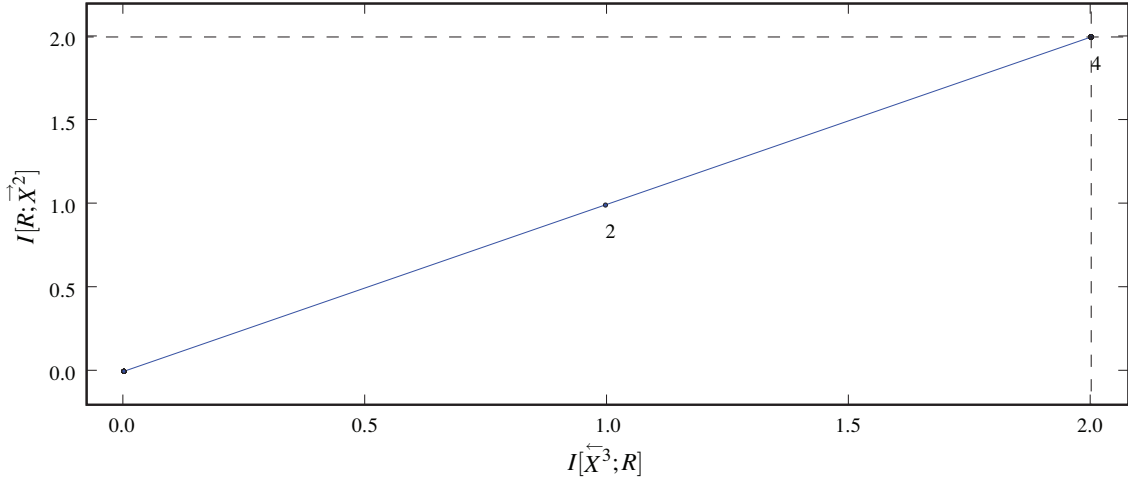


Figure 1: Prediction versus structure trade-off under OCF for the exactly predictable period-4 process:  $(0011)^\infty$ . Monitored in the information plane. The horizontal dashed line is the full predictive information ( $I[\overleftarrow{X}^3; \overrightarrow{X}^2] = 2$  bits) and the vertical dashed line is the block entropy ( $H[\overleftarrow{X}^3] = 2$  bits). Histories of length 3 were used, along with futures of length 2.

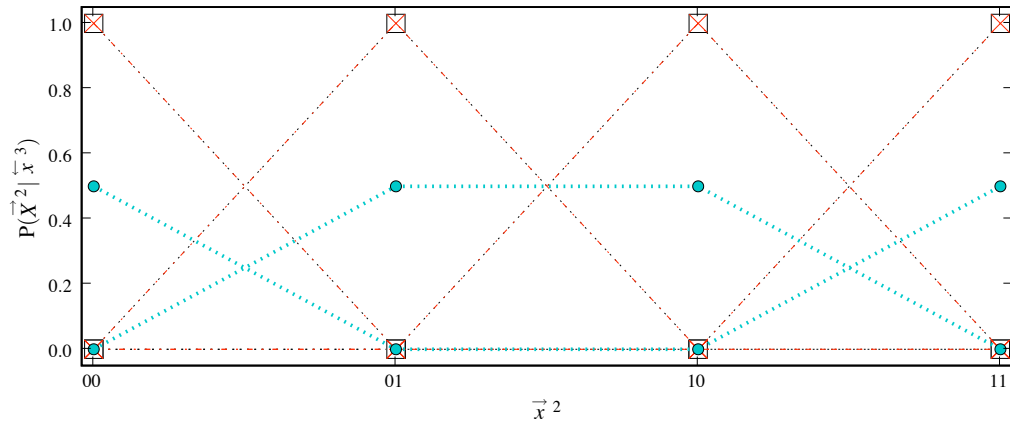


Figure 2: Morphs  $P(\overrightarrow{X}^2 | \cdot)$  for the period-4 process: the 2-state approximation (circles) compared to the  $\delta$ -function morphs for the 4 causal states (boxes). Histories of length 3 were used, along with futures of length 2 (crosses).

describes the phase of the period-four cycle. To capture those two bits, one needs exactly four underlying causal states and two bits (dotted vertical line).

The curve is the analog of the rate-distortion curve, displayed in the *information plane* (Tishby et al., 1999). The value of  $I[\mathcal{R}; \vec{X}^2]$ , evaluated at the optimal distribution, Eq. (8), is plotted versus  $I[\vec{X}^3; \mathcal{R}]$ , 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. As  $\lambda$  decreases, the next distinct point on the RD curve is that where half of the information is discarded. This comes exactly at the cost of one predictive bit. Finally, we find a model which 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 which is the worst case for possible beneficial trade-offs between prediction error and model complexity.

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. Crutchfield and Feldman (2003).

## 5.2 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., Crutchfield and Feldman (2003).

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, 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

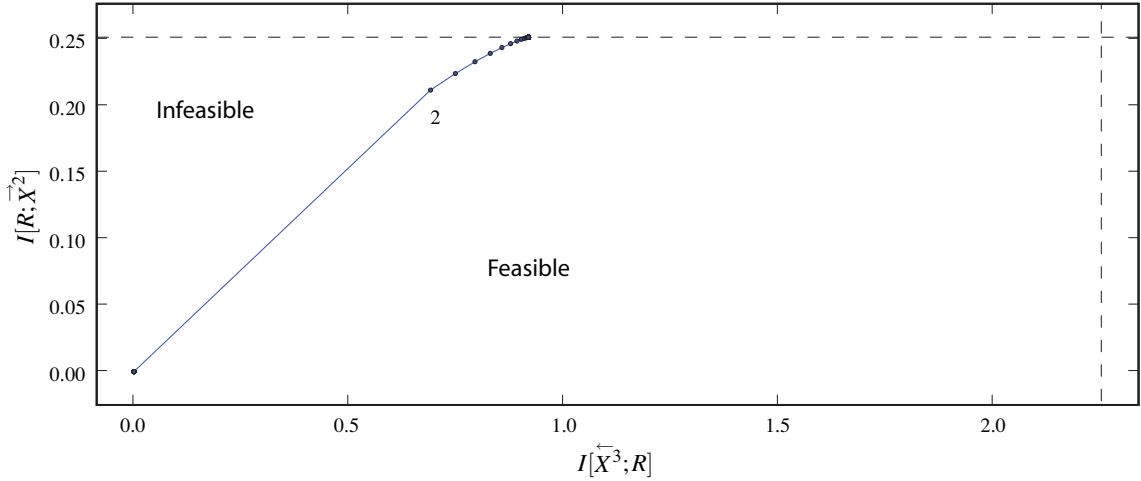


Figure 3: OCF’s behavior monitored in the information plane— $I[\mathcal{R}; \vec{X}^2]$  versus  $I[\overleftarrow{X}^3; \mathcal{R}]$ —for the Golden Mean process. Histories of length 3 were used, along with futures of length 2. The horizontal dashed line is the full predictive information  $I[\overleftarrow{X}^3; \vec{X}^2] = I[\sigma; \vec{X}^2] \approx 0.25$  bits which, as seen, is an upper bound on  $I[\mathcal{R}; \vec{X}^2]$ . Similarly, the vertical dashed line is the block entropy  $H[\overleftarrow{X}^3] \approx 2.25$  bits which is an upper bound on the retrodictive information  $I[\overleftarrow{X}^3; \mathcal{R}]$ . 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  $N_c$ .

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  $H[\overleftarrow{X}^3] \approx 2.25$  bits. The rest 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.

### 5.3 Even Process: Hidden Markov Chain

Now consider a hidden Markov process: the *Even Process* (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*. 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 sub-words which are themselves forbidden words. A system is *sofic* if its list of irreducible forbidden words is infinite. The

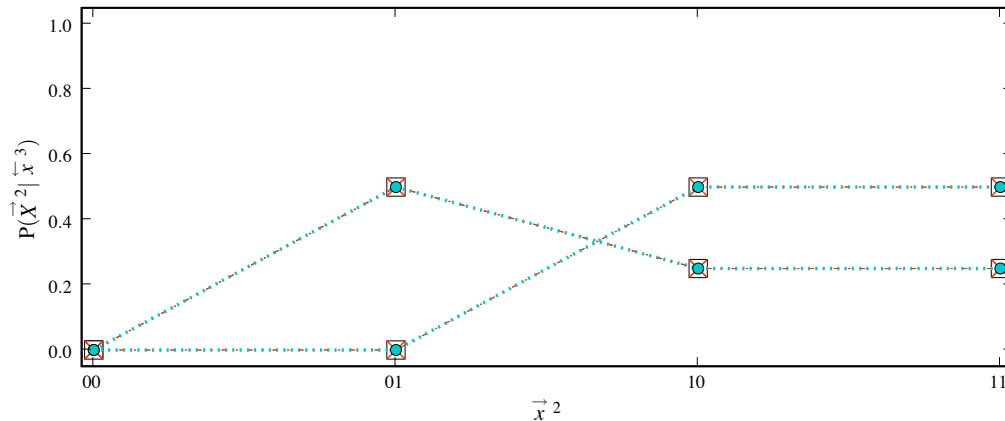


Figure 4: Future conditional probabilities  $P(\vec{X}^2 | \cdot)$  conditioned on causal states  $\sigma \in \mathcal{S}$  (boxes) and on the OCF reconstructed states  $\rho \in \mathcal{R}$  (circles) for the Golden Mean process. As an input to OCF, future conditional probabilities  $P(\vec{X}^2 |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

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 (Crutchfield and Feldman, 2003). 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 indicate the transitions 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.

#### 5.4 Random Random XOR: A structurally complex process

The previous examples have 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

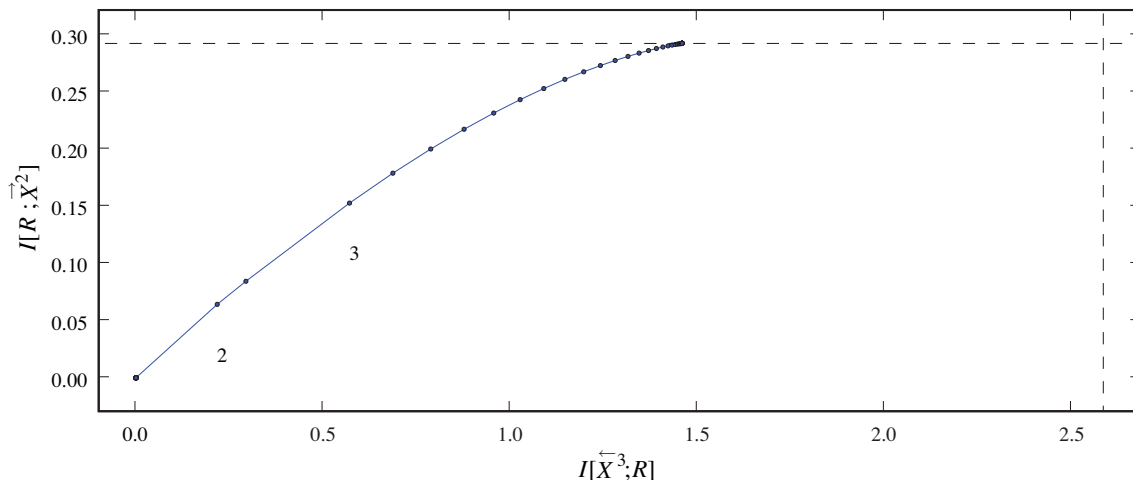


Figure 5: OCF’s behavior inferring the Even Process: monitored in the information plane— $I[\mathcal{R}; \vec{X}^2]$  versus  $I[\overleftarrow{X}^3; \mathcal{R}]$ . Histories of length 3 were used, along with futures of length 2. The horizontal dashed line is the full predictive information  $I[\overleftarrow{X}^3; \vec{X}^2] \approx 0.292$  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 retrodiction information  $I[\overleftarrow{X}^3; \mathcal{R}]$ .

terms of model size. Such processes are not causally compressible, a point that has been discussed in Still and Crutchfield (2007). 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. The Even Process, for example, is an information source that is fully causally compressible, meaning that the statistical complexity of the causal state partition,  $H[\mathcal{S}]$ , is smaller than the total available historical information—the entropy of the past,  $H[\overleftarrow{X}]$ .

Now we study a process that is causally compressible, but *not fully*, meaning that we need to keep all of the historical information to be maximally predictive, which is the same as stating that  $H[\mathcal{S}] = H[\overleftarrow{X}^L]$ . 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  $\lambda$  which 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 of which describe a complicated transient structure (Crutchfield and Feldman, 2003). As such it is a very

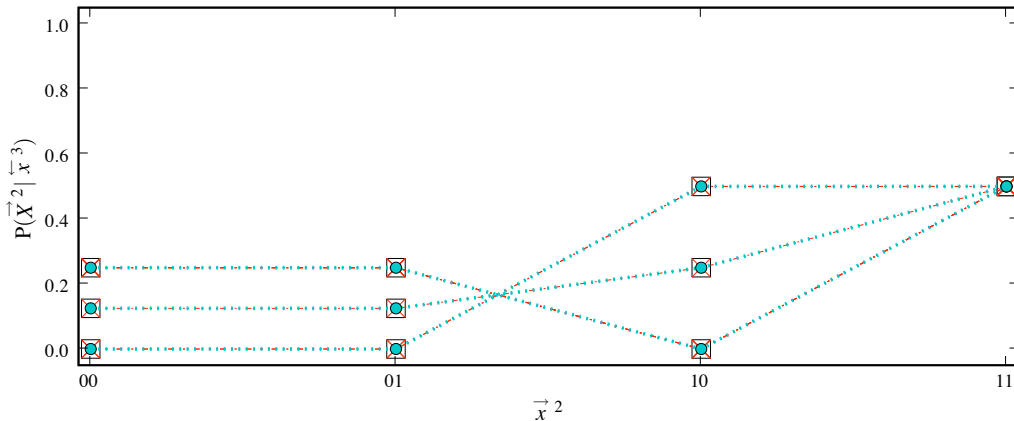


Figure 6: Future future conditional probabilities  $P(\bar{X}^2 | \cdot)$  conditioned on causal states  $\sigma \in \mathcal{S}$  (boxes) and on the OCF-reconstructed states  $\rho \in \mathcal{R}$  (circles) for the Even Process. As an input to OCF, future conditional probabilities  $P(\bar{X}^2 |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

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 (denoted) 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 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 statistical complexity of the causal state partition is equal to the full entropy of the past,  $C_\mu[\mathcal{S}] = H[\bar{X}^3]$ , so the process is not fully causally compressible, unlike the Even Process and the Golden Mean process. 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 conditional future probabilities of the (best) four-state model are shown in Fig. 8. They are mixtures of pairs of the eight causal states.

The rate-distortion curve informs the modeler about the (best possible) ratio of predictive power to model complexity:  $I[\mathcal{R}; \bar{X}] / I[\bar{X}; \mathcal{R}]$ . 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. Phase transitions occur to models with a larger number of effective states, as  $\lambda$  decreases (Rose, 1998).

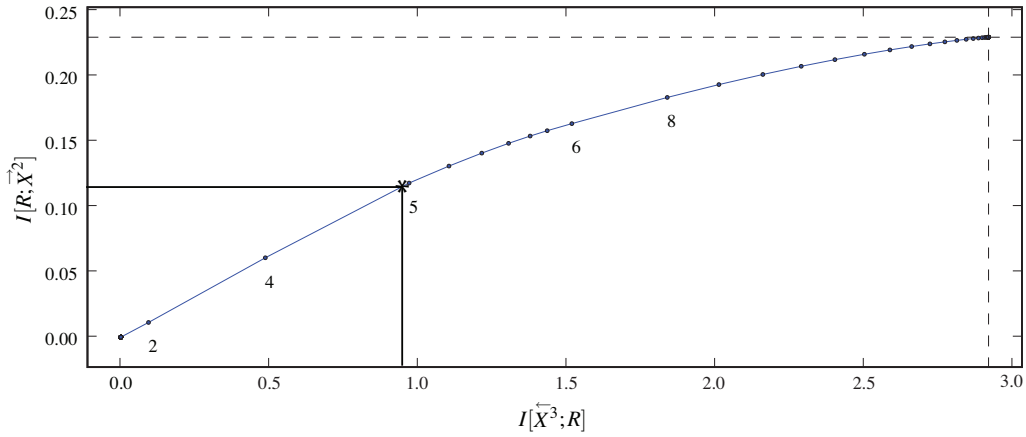


Figure 7: Prediction versus structure trade-off under OCF for the random-random XOR (RRXOR) process, as monitored in the information plane. As above, the horizontal dashed line is the predictive information ( $\approx 0.230$  bits) and the vertical dashed line is the block entropy ( $\approx 2.981$  bits). Histories of length 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 future conditional probabilities of which are depicted in Fig. 9.

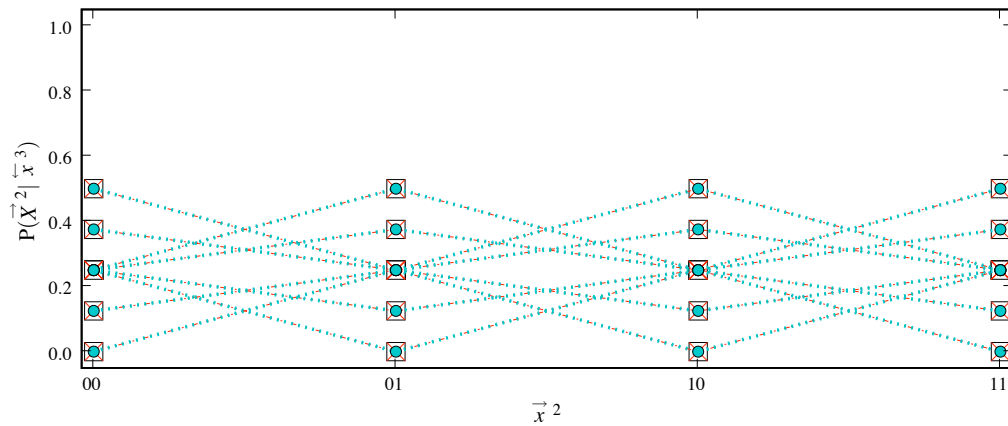


Figure 8: Future conditional probabilities  $P(\vec{X}^2 | \cdot)$  for the RRXOR process: the 8-state approximation (circles) finds the causal states (boxes). Histories of length 3 were used, along with futures of length 2.

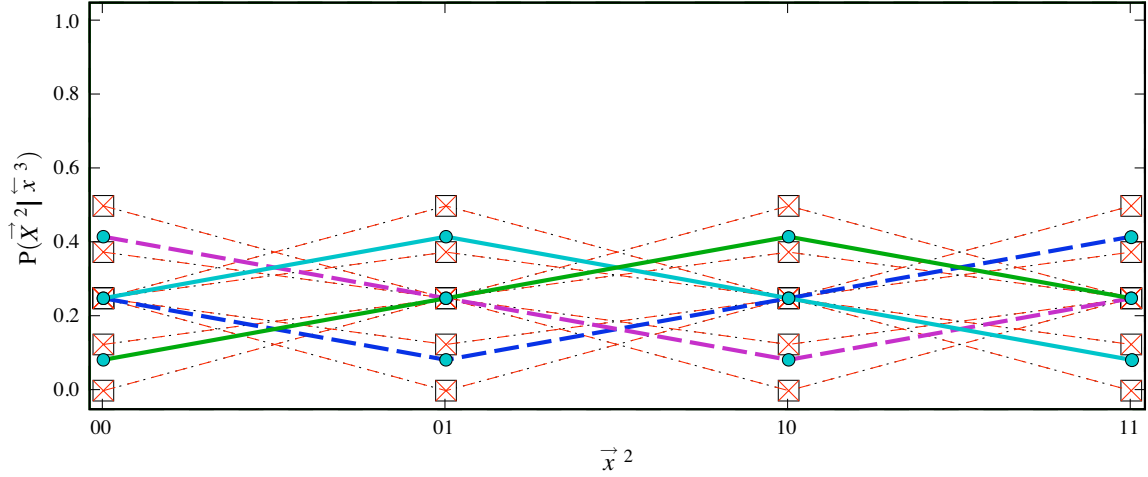


Figure 9: Morphs  $P(\vec{X}^2 | \cdot)$  for the RRXOR process: the 4-state approximation (circles and colored lines: state 1 - cyan/full, 2 - green/full, 3 - blue/dashed, 4 - purple/dashed) compared to causal states (boxes). Histories of length 3 were used, along with futures of length 2.

## 6. 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 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  $t_p$ ) 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:<sup>6</sup>  $\hat{I}[\vec{X}^{t_p}; \vec{X}^{t_f}] \geq I[\vec{X}^{t_p}; \vec{X}^{t_f}]$ . Evaluating the objective function at this estimate may lead 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 Still and Bialek (2004), 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  $\Delta(\vec{X}^{t_p}, \vec{X}^{t_f})$  is a small perturbation, then the error can be approximated by Still and Bialek (2004, Eq. (5.8)):

$$\mathcal{E}(N_c) = \frac{K - 1}{2 \ln(2)} \frac{N_c}{M}, \quad (24)$$

in the low temperature regime,  $\lambda \rightarrow 0$ .  $K$  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

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

mutual information is shared with the future, corrected by this error:

$$N_c^* := \arg \max_{N_c} \widehat{I}[\overleftarrow{X}^{t_p}; \overrightarrow{X}^{t_f}]_{\lambda \rightarrow 0}^{\text{corrected}}(N_c) , \quad (25)$$

with

$$\widehat{I}[\overleftarrow{X}^{t_p}; \overrightarrow{X}^{t_f}]_{\lambda \rightarrow 0}^{\text{corrected}}(N_c) = \widehat{I}[\overleftarrow{X}^{t_p}; \overrightarrow{X}^{t_f}]_{\lambda \rightarrow 0}(N_c) - \mathcal{E}(N_c) . \quad (26)$$

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

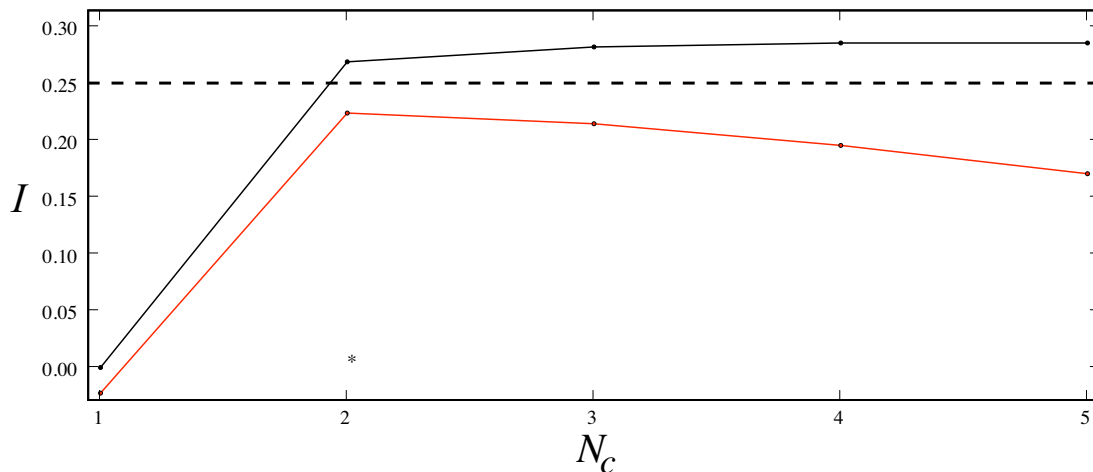


Figure 10: Information  $I$  captured about the future versus the number  $N_c$  of reconstructed states, with statistics estimated from length  $T = 100$  time series sample from the Golden Mean process. Upper line: plotted on the vertical axis is  $\widehat{I}[\mathcal{R}; \overrightarrow{X}^2]_{\lambda \rightarrow 0}$  (not corrected); lower line: plotted on the vertical axis is the quantity  $\widehat{I}[\mathcal{R}; \overrightarrow{X}^2]_{\lambda \rightarrow 0}^{\text{corrected}}$ , 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, for comparison. This value is not known to the algorithm, it is computed from the true process statistics. Histories of length 3 and futures of length 2 were used. The asterisk denotes the optimal number of effective states.

We illustrate OCE on the Golden Mean and Even Processes studied in Sec. 5. With the *correct* number of underlying states, those processes are fully causally compressible at a substantial compression. Figures 10 and 12 show the mutual information  $I[\mathcal{R}; \overrightarrow{X}^2]$  versus the number  $N_c$  of inferred states, with statistics estimated from time series of lengths  $T = 100$ . The graphs compare the mutual information  $\widehat{I}[\mathcal{R}; \overrightarrow{X}^2]_{\lambda \rightarrow 0}$  evaluated using the estimate  $\widehat{P}(\overrightarrow{X}^2; \overleftarrow{X}^3)$  (upper curve) to the corrected information  $\widehat{I}[\mathcal{R}; \overrightarrow{X}^2]_{\lambda \rightarrow 0}^{\text{corrected}}$  calculated by subtracting the approximated error Eq. (24) with  $K = 4$  and  $M = 96$  (lower curve).

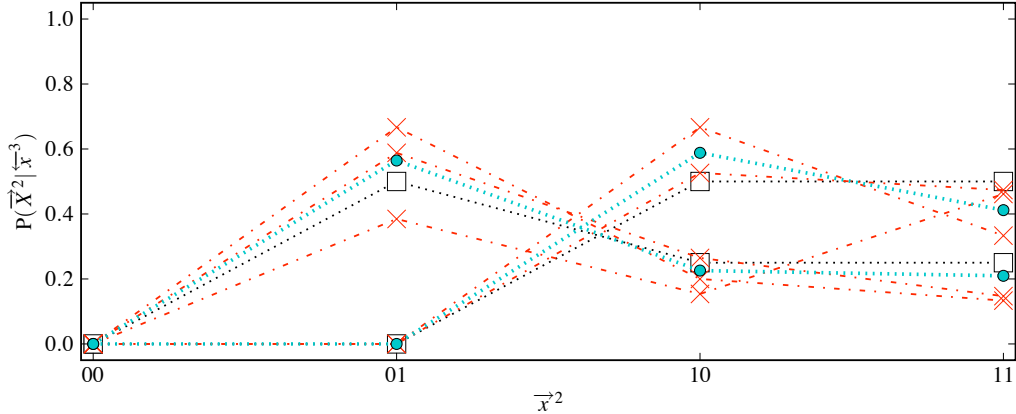


Figure 11: OCE’s best two-state approximated future conditional probabilities (circles) for the Golden Mean process. Compared to true (unknown) future conditional probabilities (squares). The OCE inputs are the estimates of  $\hat{P}(\vec{X}^2 | \vec{x}^3)$  (crosses).

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. 5) 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.

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 output 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, the OCE output approximates the correct future conditional probabilities. For the Even Process there is more spread in the estimated OCE output future conditional probabilities. Nonetheless, OCE reduced the fluctuations in its inputs and corrected in the direction of the true underlying future conditional probabilities.

## 7. 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-

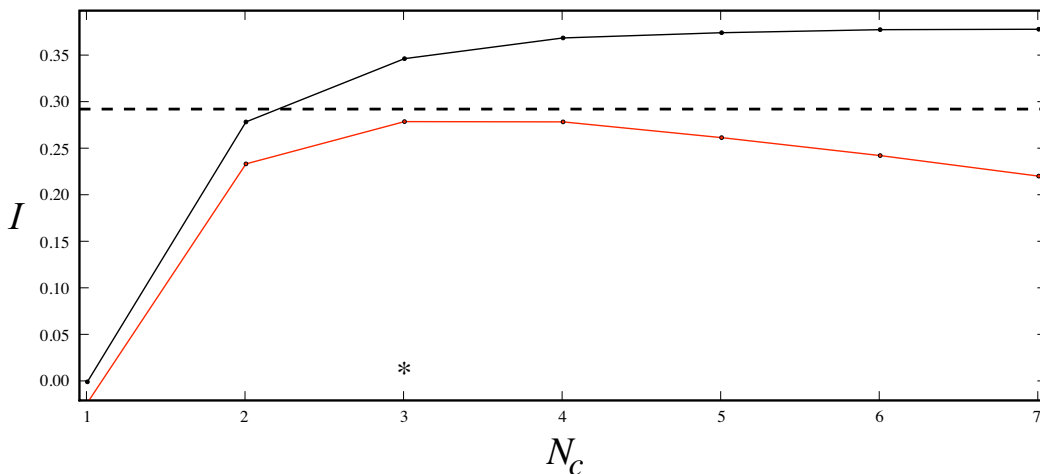


Figure 12: Information  $I$  captured about the future versus the number  $N_c$  of reconstructed states, with statistics estimated from length  $T = 100$  time series sample from the Even Process. Upper line:  $\hat{I}[\mathcal{R}; \vec{X}^2]_{\lambda \rightarrow 0}$ , not corrected; lower line:  $\hat{I}[\mathcal{R}; \vec{X}^2]_{\lambda \rightarrow 0}^{\text{corrected}}$ , 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 3 and futures of length 2 were used. The asterisk denotes the optimal number of effective states.

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.

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.

Our main result is that OCF reconstructs the causal state partition, a representation previously known from computational mechanics (Crutchfield and Shalizi, 1999). 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 (Tishby et al., 1999) recovers the unique minimal sufficient statistic 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 we have used, because we showed that this particular way of phrasing the causal inference problem results

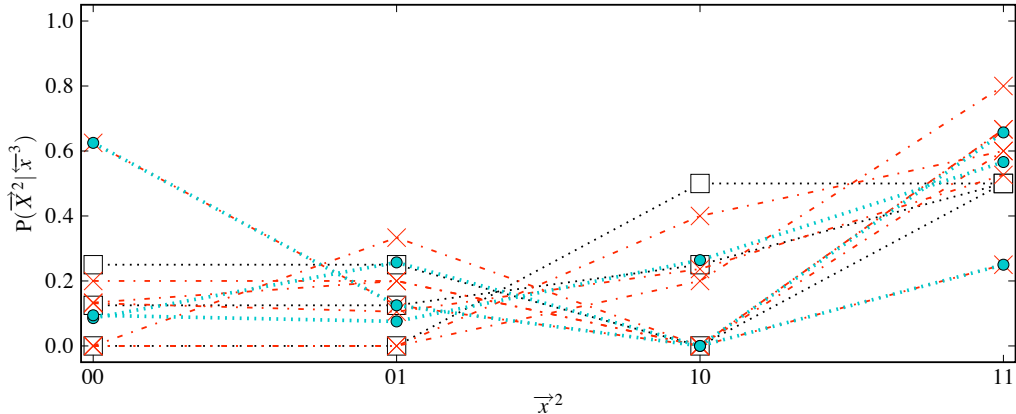


Figure 13: OCE’s best three-state approximated future conditional probabilities (circles) for the Even Process (d). Compared to true (unknown) future conditional probabilities (squares). The OCE inputs are the estimates of  $\hat{P}(\vec{X}^2 | \overleftarrow{x}^3)$  (crosses).

in a representation that is a sufficient statistic and minimal and reflects the structure of the underlying process that generated the data. It 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.

Additionally, the continuous trade-off allows us to go beyond the purely deterministic history-to-state assignments that computational mechanics introduced, 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.

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. CJE is supported by a Department of Education GAANN graduate fellowship. SS thanks W. Bialek, discussions with whom have contributed significantly to shaping some of the ideas expressed in this article, and thanks L. Bottou and I. Nemenmann for useful discussions.

## References

- S. Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Trans. Info. Th. IT-18*, pages 14–20, 1972.
- W. Bialek. Thinking about the brain. In *Physics of bio-molecules and cells; École d'été de physique théorique Les Houches Session LXXV*, pages 485–577. Springer-Verlag, 2001.
- W. Bialek and N. Tishby. Predictive information, 1999. URL [arXiv:cond-mat/9902341v1](https://arxiv.org/abs/cond-mat/9902341v1).
- R. E. Blahut. Computation of channel capacity and rate distortion function. *IEEE Trans. Info. Th. IT-18*, pages 460–473, 1972.
- T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, second edition, 2006.
- F. Creuzig. PhD thesis, Humboldt University, Berlin, Germany, 2008.
- 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 N. H. Packard. Symbolic dynamics of noisy chaos. *Physica D*, 7:201–223, 1983.
- J. P. Crutchfield and C. R. Shalizi. Thermodynamic depth of causal states: Objective complexity via minimal representations. *Phys. Rev. E*, 59(1):275–283, 1999.
- J. P. Crutchfield and K. Young. Inferring statistical complexity. *Phys. Rev. Lett.*, 63:105–108, 1989.
- P. Grassberger. Toward a quantitative theory of self-generated complexity. *Intl. J. Theo. Phys.*, 25(9):907–938, 1986.
- H. Kantz and T. Schreiber. *Nonlinear Time Series Analysis*. Cambridge University Press, Cambridge, UK, 2006.
- 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.
- K. Rose, E. Gurewitz, and G. C. Fox. Statistical Mechanics and Phase Transitions in CLustering. *Phys. Rev. Lett.*, 65(8):945–948, 1990.
- 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.
- S. Still. Information theoretic approach to interactive learning. *EPL*, 85, 2009. Earlier version (2007) available at <http://arxiv.org/abs/0709.1948>.
- 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? 2007. URL [arxiv.org:0708.0654](https://arxiv.org/abs/0708.0654)[physics.gen-ph].

- S. H. Strogatz. *Nonlinear Dynamics and Chaos: with applications to physics, biology, chemistry, and engineering*. Addison-Wesley, Reading, Massachusetts, 1994.
- 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.