

Structure or Noise?

Susanne Still^{1,*} and James P. Crutchfield^{2,†}

¹*Information and Computer Sciences, University of Hawaii at Manoa, Honolulu, HI 96822*

²*Center for Computational Science & Engineering and Physics Department,
University of California Davis, One Shields Avenue, Davis, CA 95616*

(Dated: August 7, 2007)

We show how theory building can naturally distinguish between regularity and randomness. Starting from basic modeling principles, using rate distortion theory and computational mechanics we argue for a general information-theoretic objective function that embodies a trade-off between a model's complexity and its predictive power. The family of solutions derived from this principle corresponds to a hierarchy of models. At each level of complexity, they achieve maximal predictive power, identifying a process's exact causal organization in the limit of optimal prediction. Examples show how theory building can profit from analyzing a process's *causal compressibility*, which is reflected in the optimal models' rate-distortion curve.

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

Progress in science is often driven by discovering novel patterns where they were not recognized previously. Historically, physics has relied on the creative mind of the theorist to articulate mathematical models that capture nature's patterns in physical principles and laws. Pattern discovery is now a primary concern in many disciplines, with the last decade witnessing a new era in collecting truly vast data sets. Examples include contemporary experiments in particle physics and astronomical sky surveys, but range to automated language translation, genomics, and web social organization. The volume of data far exceeds what any human can analyze directly by hand. Therefore, a principled understanding of model making is critical. It is necessary for developing procedures that can guide automated theory building in data-rich domains.

A key realization is that any theory delineates mechanism from randomness by deciding what part of an observed phenomenon is due to the underlying process's structure and what is irrelevant. Irrelevant parts are considered noise and typically modeled probabilistically. Notably, the distinction between structure and noise is often made implicitly. Successful theory building, however, depends centrally on deciding where that demarcation should lie.

Even if given all the microscopic data produced by a box of gas molecules (say, via large-scale molecular dynamic simulation), the physicist focuses on a few concise macroscopic properties—such as equations of state, thermodynamic phase, and symmetry group—while discarding a vast amount of microscopic information. Implicitly, the evolution of microstates is considered irrelevant. The consequence, of course, is that exact prediction of individual molecular trajectories fails. The extremely large prediction error, however, is greatly outweighed by the com-

pactness of the resulting theory which leads to a gain in scientific understanding [1, pp. 357-358]. This example illustrates that the implicit trade-off between structure and noise in theory building is equivalent to a trade-off between model complexity and prediction error.

The trade-off between assigning a causal mechanism to the occurrence of an event or explaining the event as being merely random has a long history. Perhaps surprisingly given that history, how one implements the trade-off is still a very active topic. Nonlinear time series analysis [2–4], for example, attempts to account for long-range correlations produced by dynamical systems—correlations not adequately modeled by typical assumptions, such as linearity and independent, identically distributed data. Success in this endeavor requires directly addressing the notion of structure and pattern [2, 5]. Careful examination of the essential goals of prediction has led to a principled definition of structure that captures a dynamical system's causal organization [6–8].

This approach views a process $P(\vec{X}, \vec{X})$ as a communication channel [21]: it transmits information from the *past* $\vec{X} = \dots X_{-3}X_{-2}X_{-1}$ to the *future* $\vec{X} = X_0X_1X_2\dots$ by storing it in the present. For the purpose of forecasting the future, two different pasts are equivalent if they result in the same prediction. In general this prediction is probabilistic, given by the *morph* $P(\vec{X} | \vec{x})$, which is the distribution of futures given a particular past \vec{x} . Computational mechanics [6] introduced an equivalence relation $\vec{x} \sim \vec{x}'$ that groups all histories which give rise to the same morph: $\epsilon(\vec{x}) = \{\vec{x}' : \Pr(\vec{X} | \vec{x}) = \Pr(\vec{X} | \vec{x}')\}$. The resulting causal model, the *causal-state partition* $\mathcal{S} = P(\vec{X}, \vec{X}) / \sim$, partitions the space \vec{X} of pasts.

Causal states have the Markovian property that they *shield* past and future [8]:

$$P(\vec{X}, \vec{X} | \mathcal{S}) = P(\vec{X} | \mathcal{S})P(\vec{X} | \mathcal{S}). \quad (1)$$

This is related to the fact that the causal-state partition is optimally predictive. Causal shielding is equivalent to

*Electronic address: sstill@hawaii.edu

†Electronic address: chaos@cse.ucdavis.edu

$P(\vec{X} | \vec{X}, \mathcal{S}) = P(\vec{X} | \mathcal{S})$. When the past is known, then the future distribution is not altered by the history space partitioning:

$$P(\vec{X} | \vec{X}, \mathcal{R}) = P(\vec{X} | \vec{X}), \quad (2)$$

This is true for *all* partitions \mathcal{R} with states \mathcal{R} . For the causal state partition it means that $P(\vec{X} | \mathcal{S}) = P(\vec{X} | \vec{X})$. Therefore, causal shielding is equivalent to the fact [8] that the causal states capture *all* of the information $I[\vec{X}; \vec{X}]$ that is shared between past and future: $I[\mathcal{S}; \vec{X}] = I[\vec{X}; \vec{X}]$. This is the process's *excess entropy* or *predictive information* [9, and references therein]. Out of all optimally predictive models $\hat{\mathcal{R}}$ for which $I[\hat{\mathcal{R}}; \vec{X}] = I[\vec{X}; \vec{X}]$ the causal-state partition has the smallest *statistical complexity* [8], $C_\mu := H(\mathcal{S}) \leq H[\hat{\mathcal{R}}]$. This is the minimal amount of information that must be stored in order to communicate all of the excess entropy from the past to the future. Combined, these properties mean that the causal-state partition is the basis against which modeling should be compared, since it captures all of a process's predictive information at maximum efficiency.

There are many scenarios, however, in which one does not need to or explicitly does not want to capture *all* of the predictive information, as in the case of the thermodynamic description of a box of gas molecules. Is there a way to systematically approximate the causal states?

We frame this problem in terms of communicating a model over a channel with limited capacity. Rate-distortion theory [10] provides a principled way to find a lossy compression of an information source such that the resulting code is minimal at fixed fidelity to the original signal. The compressed representation, denote it \mathcal{R} , is generally specified by a probabilistic map $P(\mathcal{R} | \vec{x})$ from the input message, here the past \vec{x} , to code words, here the model's states \mathcal{R} with values $\rho \in \mathcal{R}$. This map specifies a model, and the *coding rate* $I[\vec{X}; \mathcal{R}]$ measures its complexity. The latter is related to a model's statistical complexity [22], since $I[\vec{X}; \mathcal{R}] = H[\mathcal{R}] - H[\mathcal{R} | \vec{X}]$. The causal states $\sigma \in \mathcal{S}$ induce a deterministic partition of $\vec{\mathbf{X}}$ [8], as one can show that $P(\sigma | \vec{x}) = \delta_{\sigma, \epsilon(\vec{x})}$.

We can now use rate-distortion theory to back away from the best (causal state) representation towards less complex models by controlling the coding rate. Small models are distinguished from more complex models by the fact that they can be transmitted more concisely. However, they are also associated with a larger error. Rate-distortion theory quantifies this loss by a *distortion function* $d(\vec{x}; \rho)$. The coding rate is then minimized [11] over the assignments $P(\mathcal{R} | \vec{X})$ at fixed average distortion $D[\vec{X}; \mathcal{R}] = \left\langle d(\vec{x}; \rho) \right\rangle_{P(\vec{x}, \rho)}$.

In building predictive models, the loss should be measured by how much the resulting models deviate from accurate prediction. Causal shielding, Eq. (1), says the

goal for approximate models should be to come as close to decorrelating past and future as possible, minimizing the excess entropy *conditioned on the model states* \mathcal{R}

$$I[\vec{X}; \vec{X} | \mathcal{R}] = \left\langle \left\langle \log \left[\frac{P(\vec{x}, \vec{x} | \rho)}{P(\vec{x} | \rho)P(\vec{x} | \rho)} \right] \right\rangle_{P(\vec{x} | \vec{x})} \right\rangle_{P(\vec{x}, \rho)}. \quad (3)$$

This corresponds to using the relative entropy $\mathcal{D}(P(\vec{x} | \vec{x}) || P(\vec{x} | \rho))$ between the past-conditioned morphs and those induced by the model state ρ as the distortion function:

$$\begin{aligned} d(\vec{x}; \rho) &= \left\langle \log \left[\frac{P(\vec{x}, \vec{x} | \rho)}{P(\vec{x} | \rho)P(\vec{x} | \rho)} \right] \right\rangle_{P(\vec{x} | \vec{x})} \\ &= \mathcal{D}(P(\vec{x} | \vec{x}) || P(\vec{x} | \rho)). \end{aligned} \quad (4)$$

(We used Eq. (2) to obtain Eqs. (3) and (4).) Altogether, then, we must solve the constrained optimization problem

$$\min_{P(\mathcal{R} | \vec{X})} \left(I[\vec{X}; \mathcal{R}] + \beta I[\vec{X}; \vec{X} | \mathcal{R}] \right), \quad (5)$$

where the Lagrange multiplier β controls the trade-off between model complexity and prediction error, between structure and noise.

The conditional excess entropy Eq. (3) is the difference between the process's excess entropy and the information $I[\mathcal{R}; \vec{X}]$ that the model states contain about the future: $I[\vec{X}; \vec{X} | \mathcal{R}] = I[\vec{X}; \vec{X}] - I[\mathcal{R}; \vec{X}]$, due to Eq. (2). The excess entropy $I[\vec{X}; \vec{X}]$ is a property of the process, however, and not dependent on the model. Therefore, the optimization problem in Eq. (5) is equivalent to that analyzed in Ref. [12] and maps onto the information bottleneck method [13], with the solution

$$P_{\text{opt}}(\rho | \vec{x}) = \frac{P(\rho)}{Z(\vec{x}, \beta)} e^{-\beta E(\rho, \vec{x})}, \quad (6)$$

where

$$E(\rho, \vec{x}) = \mathcal{D}(P(\vec{X} | \vec{x}) || P(\vec{X} | \rho)), \quad (7)$$

$$P(\vec{X} | \rho) = \frac{1}{P(\rho)} \sum_{\vec{x} \in \vec{\mathbf{X}}} P(\vec{X} | \vec{x}) P(\rho | \vec{x}) P(\vec{x}), \quad \text{and} \quad (8)$$

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

Eqs. (6)-(9) must be solved self-consistently, and this can be done numerically [13].

Eq. (6) specifies a family of models parametrized by β with the form of Gibbs distributions. Within an analogy to statistical mechanics [14], β corresponds to the inverse temperature, E is the energy, and $Z = \left\langle e^{-\beta E(\rho, \vec{x})} \right\rangle_{P(\rho)}$ the partition function. Past observations are microstates

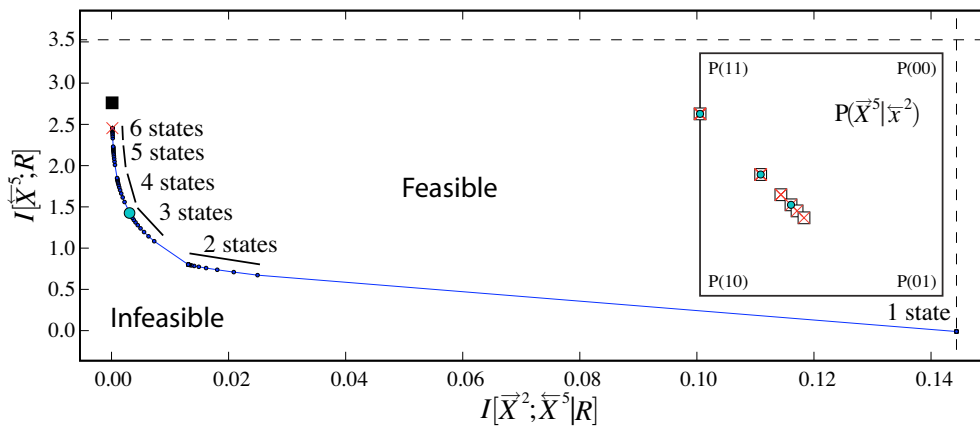


FIG. 1: Trading structure off against noise using optimal causal inference (OCI): Rate-distortion curve for the SNS, coding rate $I[\bar{X}^5; \mathcal{R}]$ versus distortion $I[\bar{X}^2; \bar{X}^5 | \mathcal{R}]$. Dashed lines mark maximum values: past entropy $H[\bar{X}]$ and excess entropy $I[\bar{X}; \bar{X}]$, respectively. The causal-state limit for infinite sequences is shown in the upper left (solid box). (Inset) SNS future morphs $P(\bar{X}^2 | \bar{x}^5)$: OCI six-state reconstruction (crosses), true causal states (boxes), and three-state approximation (circles).

which are summarized by macrostates—the model states $\rho \in \mathcal{R}$.

One of our main results is that these optimal solutions retrieve the causal-state partition in the limit $\beta \rightarrow \infty$, which emphasizes prediction; Ref. [12] gives the proof. This means one finds what we argued above is the goal of predictive modeling. Remember that the model complexity C_μ of the causal-state partition is minimal among the optimal predictors and so not necessarily equal to the maximum value of the coding rate $I[\bar{X}; \mathcal{R}] \leq H[\bar{X}]$. Since, in effect, we map from the original process to the best causal-state approximation, we refer to the method as *optimal causal inference* (OCI) [12].

The causal-state model captures *all* of the predictive information. However, we can find less complex models if we allow for a larger distortion, accepting less predictive power. We study the nature of the trade-off embodied in Eq. (5) by evaluating the objective function at the optimum for each value of β . The shape of the resulting *rate-distortion curve* characterizes a process’s *causal compressibility* via the interdependence between $I[\bar{X}; \mathcal{R}]$ and $I[\bar{X}^2; \bar{X}^5 | \mathcal{R}]$. Since the variation of the objective function in Eq. (5) vanishes at the optimum, the curve’s slope is $\delta I[\bar{X}; \mathcal{R}] / \delta D[\bar{X}^2; \bar{X}^5 | \mathcal{R}] = -\beta$. For a given process the rate-distortion curve determines what predictability the best model at a fixed complexity can achieve and, vice versa, how small a model can be made at fixed predictability. Below the curve, lie *infeasible* causal compression codes; above are *feasible* larger models that are no more predictive than those directly on the curve. In short, the rate-distortion curve determines how to *optimally* trade structure for noise.

As an example, consider the *simple nondeterministic source* (SNS)—a two-state hidden Markov model that specifies a binary information source with nontrivial statistical structure, including infinite-range correlations and an infinite number of causal states; for its definition

see [7]. The SNS’s rate-distortion curve, calculated for pasts of length 5 and futures of length 2 is shown in Fig. 1. We computed the curve using deterministic annealing, following [14] with an annealing rate of 1.1.

The finite causal-state model is recovered by OCI (cross in upper left). The *causal-state limit*, which is calculated analytically for *infinite* pasts and futures (solid box) gives a reference point. The curve drops rapidly away from the finite causal-state model with six effective states, indicating that there is little predictive cost in using significantly smaller models with successively fewer effective states. The curve then levels out below three states: smaller models incur a substantial increase in distortion (loss in predictability) while little is gained in terms of compression. Quantitatively, specifying the best four-state model (at $I[\bar{X}; \mathcal{R}] = 1.92$ bits) leads to 0.5% distortion, capturing 99.5% the SNS’s excess entropy. The distortion increases to 2% for three states (1.43 bits), 9% for two states (0.81 bits), and 100% for a single state (0 bits). Overall, the three-state model lies near a knee in the rate-distortion curve and this suggests that it is a good compromise between model complexity and predictability.

The inset in Fig. 1 shows the reconstructed morphs for the optimal three-state and six-state models. The six-state model (crosses) reconstructs the true causal-state morphs (boxes), calculated from analytically known finite-sequence causal states. The figure illustrates why the three-state model (circles) is a good compromise: two of the three-state model’s morphs capture the two more distinct SNS morphs, and its third morph summarizes the remaining, less different, SNS morphs.

While the SNS process has intricate causal structure and nontrivial causal compressibility properties, other frequently studied processes do not. Two classes of process are of particular interest. On one extreme of randomness are the independent, identically distributed (i.i.d.) processes, such as the biased coin—by definition

a completely random and unstructured source. For i.i.d. processes the rate-distortion curve collapses to a single point at $(0, 0)$, indicating that they are wholly unpredictable and causally incompressible. This is easily seen by noting first that for i.i.d. processes the excess entropy $I[\overleftarrow{X}; \overrightarrow{X}]$ vanishes, since $P(\overrightarrow{x} | \overleftarrow{x}) = P(\overrightarrow{x})$. Therefore, $I[\overleftarrow{X}; \overrightarrow{X} | \mathcal{R}] \leq I[\overleftarrow{X}; \overrightarrow{X}] = 0$ vanishes, too. Second, the energy function $E(\rho, \overleftarrow{x})$ in the optimal assignments, Eq. (7), vanishes, since $P(\overrightarrow{x} | \rho) = \left\langle P(\overrightarrow{x} | \overleftarrow{x}) \right\rangle_{P(\overleftarrow{X} | \rho)} = P(\overrightarrow{x})$. The optimal assignment given by Eq. (6) is therefore the uniform distribution and $I[\overleftarrow{X}; \mathcal{R}]|_{P_{\text{opt}}(\rho | \overleftarrow{x})} = 0$.

At the other extreme are processes for which

$$P(\overrightarrow{x} | \overleftarrow{x}) = \delta_{\overrightarrow{x}, f(\overleftarrow{x})}, \quad (10)$$

where f is *invertible*, such as periodic processes. These processes have a rate distortion curve that is a straight line, the negative diagonal. Note that $P(\overrightarrow{x} | \rho) = P(f^{-1}(\overleftarrow{x}) | \rho) = P(\overleftarrow{x} | \rho)$, and therefore $I[\overleftarrow{X}; \overrightarrow{X} | \mathcal{R}] = I[\overleftarrow{X}; \overrightarrow{X}] - I[\overleftarrow{X}; \mathcal{R}]$. The variational principle now reads $\delta(1 - \beta)I[\overleftarrow{X}; \mathcal{R}] = 0$, which implies that $\beta = 1$. The diagonal runs from $[0, \mathbf{H}]$ to $[\mathbf{H}, 0]$, where $\mathbf{H} = I[\overleftarrow{X}; \overrightarrow{X}] = H[\overleftarrow{X}]$, for these processes, due to Eq. (10) and invertibility. This diagonal represents the worst possible case for causal compression: at each level, specifying the future to one bit higher accuracy costs us exactly one bit in model complexity. Processes in this class are thus

not causally compressible. To be causally compressible, a process's rate-distortion curve must lie below the diagonal, and the more concave the curve, the more causally compressible is the process. An extremely causally compressible process can be predicted to high accuracy with a model that can be encoded at a very low model cost.

As in statistical mechanics, we have assumed so far that the distribution $P(\overleftarrow{X}, \overrightarrow{X})$ is known. In real-world applications, though, it must be estimated from the available (finite) time series data. Intuitively, the limited data size sets a bound on how much we can consider to be structure without over-fitting. Using the results in Ref. [15], optimal causal inference can be adapted to correct for the effects of finite data as well and so to not over-fit [12]. This connects our approach to statistical inference, where model complexity control is designed to avoid over-fitting due to finite-sample fluctuations (cf., e.g. [16–20]).

We have argued here that even with complete knowledge of all the data statistics, abstraction and simplification are still necessary because they serve the purpose of scientific discovery. Simple and intuitive principles of inference, information theory, and computational mechanics led us to propose studying a process's causal compressibility as a systematic method for balancing structure and noise in theory building. The result is a hierarchy of models that optimally trade-off structure and noise at each level of approximation.

We thank Chris Ellison, on a GANN fellowship, for programming. The Santa Fe Institute and CCSE Network Dynamics Programs funded by Intel Corporation supported this work.

-
- [1] J. P. Crutchfield. Semantics and thermodynamics. In Casdagli and Eubank [2], pages 317–359.
- [2] M. Casdagli and S. Eubank, editors. *Nonlinear Modeling*, SFI Studies in the Sciences of Complexity, Reading, Massachusetts, 1992. Addison-Wesley.
- [3] J. C. Sprott. *Chaos and Time-Series Analysis*. Oxford University Press, Oxford, UK, second edition, 2003.
- [4] H. Kantz and T. Schreiber. *Nonlinear Time Series Analysis*. Cambridge University Press, Cambridge, UK, second edition, 2006.
- [5] J. P. Crutchfield and B. S. McNamara. *Complex Systems*, 1:417 – 452, 1987.
- [6] J. P. Crutchfield and K. Young. *Phys. Rev. Let.*, 63:105–108, 1989.
- [7] J. P. Crutchfield. *Physica D*, 75:11–54, 1994.
- [8] J. P. Crutchfield and C. R. Shalizi. *Phys. Rev. E*, 59(1):275–283, 1999.
- [9] J. P. Crutchfield and D. P. Feldman. *CHAOS*, 13(1):25–54, 2003.
- [10] C. E. Shannon. *Bell Sys. Tech. J.*, 27, 1948.
- [11] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, second edition, 2006.
- [12] S. Still, J. P. Crutchfield, and C. J. Ellison. 2007. in preparation.
- [13] N. Tishby, F. Pereira, and W. Bialek. In B. Hajek and R. S. Sreenivas, editors, *Proc. 37th Allerton Conference*, pages 368–377. University of Illinois, 1999.
- [14] K. Rose. *Proc. IEEE*, 86(11):2210–2239, 1998.
- [15] S. Still and W. Bialek. *Neural Computation*, 16(12):2483–2506, 2004.
- [16] C. Wallace and D. Boulton. *Comput. J.*, 11:185, 1968.
- [17] H. Akaike. *Ann. Inst. Statist. Math.*, 29A:9, 1977.
- [18] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific, Singapore, 1989.
- [19] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.
- [20] D. MacKay. Cambridge University Press, Cambridge, 2003.
- [21] We follow the notation and definitions of Refs. [8, 11].
- [22] For deterministic partitions the statistical complexity and the coding rate are equal, since $H[\mathcal{R} | \overleftarrow{X}] = 0$. However, in the nondeterministic case the probabilistic map also reflects some of the model's complexity, and the coding rate $I[\overleftarrow{X}; \mathcal{R}]$ captures this. Consider the extreme of uniform assignments: $P(\mathcal{R} | \overleftarrow{x}) = 1/c$, for any given \overleftarrow{x} . In this case, even if there are many states (large statistical complexity $H[\mathcal{R}] = \log_2(c)$), they are indistinguishable and so the model effectively has only one state. Therefore, its complexity vanishes. The coding rate $I[\overleftarrow{X}; \mathcal{R}] = 0$, since $H[\mathcal{R} | \overleftarrow{X}] = H[\mathcal{R}]$.