Information Bottlenecks, Causal States, and Statistical Relevance Bases: How to Represent Relevant Information in Memoryless Transduction

Cosma Rohilla Shalizi[*] and James P. Crutchfield
Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501
Electronic address: {shalizi,chaos}@santafe.edu

(15 June 2000)

Discovering relevant, but possibly hidden, variables is a key step in constructing useful and predictive theories about the natural world. This brief note explains the connections between three approaches to this problem: the recently introduced information-bottleneck method, the computational mechanics approach to inferring optimal models, and Salmon's statistical relevance basis.

I. INTRODUCTION

Recently, Tishby, Pereira, and Bialek proposed a new method for finding concise representations of the information one set of variables contains about another [1]. This brief note explains the connections between this "information-bottleneck" method and existing mathematical frameworks and techniques. This comparison should enhance the value of research in this promising direction and clarify the relative uses of these techniques in applications.

In the interest of space, we assume readers are familiar with the notation of both [1] and [2].

II. THE INFORMATION-BOTTLENECK METHOD

In [1] the authors pose the following problem. Given a joint distribution over two random variables—the "input" X and the "output" Y, find an intermediate or "bottleneck" variable $\tilde{X}$ which is a (possibly stochastic) function of X such that $\tilde{X}$ is more compressed than X, but retains predictive information about Y. More exactly, they ask for a conditional distribution $\Pr(\tilde{x}|x)$ that minimizes the functional

$$\mathcal{F} = I[\tilde{X}; X] - \beta I[\tilde{X}; Y] \ , \qquad (1)$$

where $I[W, Z]$ is the mutual information between random variables W and Z [3] and β is a positive real number. Minimizing the first term represents the desire to find a compression of the original input data X; maximizing the second term represents the desire to retain the ability to predict Y.[1] The coefficient β governs the trade-off between these two goals: as $\beta \rightarrow 0$, we lose interest in prediction in favor of compression; whereas as $\beta \rightarrow \infty$, predictive ability becomes paramount.

Extending classical rate-distortion theory, the authors are not only able to state self-consistent equations that determine which distributions satisfy this variational problem, but give a convergent iterative procedure that finds one of these distributions. They do not address the rate of convergence.

III. CAUSAL STATES FOR TRANSDUCER-FUNCTIONALS

In [2] and earlier publications, we defined causal states for stationary stochastic processes as follows. Two histories $\overleftarrow{s}$ and $\overleftarrow{s}'$ belong to the same causal state $\mathcal{S}$ if and only if

$$\Pr(\overrightarrow{S}^L = s^L | \overleftarrow{S} = \overleftarrow{s}) = \Pr(\overrightarrow{S}^L = s^L | \overleftarrow{S} = \overleftarrow{s}') \ , \qquad (2)$$

for all s^L and for all L. That is, two histories belong to the same causal state if and only if they give the same conditional distribution for futures. In [2] we showed that the causal states defined by Eq. (2) possess two kinds of optimality. First, their ability to predict the future $\overrightarrow{S}$ is maximal. Second, they are the simplest such set of states.

Thus, we showed that the set $\boldsymbol{\mathcal{S}}$ of causal states is the solution to the following optimization problem. Given the joint distribution $\Pr(\overleftarrow{S}, \overrightarrow{S})$ over the past $\overleftarrow{S}$ and future $\overrightarrow{S}$, find the function (equivalently, partition) ϵ of $\overleftarrow{S}$ such that (i) the conditional entropy $H[\overrightarrow{S}^L | \epsilon(\overleftarrow{S})]$ is minimized for all L, and (ii) the entropy $H[\epsilon(\overleftarrow{S})]$ is minimized

[*] Permanent address: Physics Department, University of Wisconsin, Madison, WI 53706.

[1] Since $\tilde{X} = g(X, \Omega)$ for some auxiliary random variable Ω, a theorem of Shannon's assures us that $I[\tilde{X}; Y] \leq I[X; Y]$ and the transformation from X to $\tilde{X}$ cannot *increase* our ability to predict Y [4, App. 7].

among all functions $\hat{\eta}$ that satisfy condition (i).[2] The equivalence classes induced by ϵ are the causal states, and they are the unique[3] solution to the optimization problem. A moment's reflection shows that this optimization is equivalent to first maximizing the mutual information between the effective states and the futures and then minimizing the mutual information between the effective states and the histories. The first step maximizes the predictive ability of the effective states and the second selects the most concise set of states. Note that the opposite sequence of optimizations—first minimizing complexity and then maximizing predictability—is trivial, since it produces a single-state model that describes an IID sequence of random variables; e.g., a biased coin or die.

IV. MEMORYLESS TRANSDUCERS

As has been remarked earlier—e.g., [5] and [6]—the causal-state construction is not intrinsically limited to time series. Of particular interest here is the case of transducer-functionals: when one sequence of variables is a functional of another sequence, possibly a stochastic functional. In this case, one can construct causal states that (i) retain all predictive information about the output series, (ii) are deterministic functions of the prior causal state and the most recent value of the input series, and (iii) minimize the statistical complexity of (information stored in) the causal states. We present the theory of causal states for general transducers elsewhere.

In the case where the output depends on the *current* input alone—the case of memoryless transduction—the causal states assume a particularly simple form: two inputs x and x' belong to the same causal state if and only if

$$\Pr(Y = y|X = x) = \Pr(Y = y|X = x') \ , \qquad (3)$$

for all y.[4] In this case, it can be shown that

$$H[Y|\epsilon(X)] \leq H[Y|\eta(X)] \ , \qquad (4)$$

for any other partition of the inputs η and that, among all the rival partitions $\hat{\eta}$ minimizing the conditional entropy of the outputs (the prescient rivals of ϵ),

$$H[\epsilon(X)] \leq H[\hat{\eta}(X)] \ . \qquad (5)$$

This is to say, the causal states are the most compressed hidden variables. In the sense of [1], they are optimal bottleneck variables.

One concludes that these are precisely what should be delivered by the information-bottleneck method in the limit where $\beta \to \infty$. It is not immediately obvious that the iterative procedure of [1] is still valid in this limit. Nonetheless, that ϵ is the partition satisfying their original constraints is evident.

We note in passing that, as shown in [2], prescient rivals—those sets of states that retain all predictive information from the original inputs while compressing them—are sufficient statistics. Conversely, [1] states that, when sufficient statistics exist, then compression-with-prediction is possible.

V. THE STATISTICAL RELEVANCE BASIS

Before closing, we point out another solution to the problem of discovering concise and predictive hidden variables. In his books [7] and [8], Salmon put forward a construction, under the name of the "statistical relevance basis", that is identical in its essentials with that of causal states for memoryless transducers.[5] Owing to the rather different aims for which Salmon's construction was intended—explicating the notion of "causation" in the philosophy of science, no one seems to have proved its information-theoretic optimality properties nor even to have noted its connection to sufficient statistics. (Briefly: if a nontrivial sufficient partition of the input variables exists, then the relevance basis is the minimal sufficient partition. These proofs will appear elsewhere.)

VI. COMPARISON AND CONCLUSION

Recapitulating, the causal states for memoryless transduction coincide with the cells of the "bottleneck" partition in the limit $\beta \to \infty$. Moreover, both are identical with the statistical relevance basis of Salmon.

The construction of the causal states does not allow us to discard *any* predictive information about the output Y, even if this might allow for a substantial reduction in the statistical complexity. The bottleneck method, by contrast, generally throws away *some* predictive information. It trades one bit less statistical complexity for $1/\beta$ bits less predictive information. Of course, the linear trade-off and the particular value of the coefficient

[2] The states induced by the partition $\hat{\eta}$ are called the prescient rivals of the causal states induced by ϵ. The entropy $H[\epsilon(\overleftarrow{S})]$ is called the statistical complexity and measures the "size" of, or amount of information stored in, the causal states.

[3] More precisely, any other function satisfying conditions (i) and (ii) may differ from ϵ on at most a set of histories of measure zero.

[4] A somewhat more complicated construction is necessary when the transducer exhibits memory.

[5] Salmon's work only came to our attention in mid-1998, and thus it is not cited in our publications including and prior to [2].

β that controls it are *ad hoc* choices. Whether this is acceptable in applications would seem to depend on the goal. For example, if the goal is a practical "lossy" data-compression scheme, the bottleneck method recommends itself. However, if the goal is representing the intrinsic computation or causal structure of some natural process, causal states are better suited to the task.

ACKNOWLEDGMENTS

This work was partially supported by the SFI Computation, Dynamics, and Learning Program, by AFOSR via NSF grant PHY-9970158, and by DARPA under contract F30602-00-2-0583.

[1] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. Electronic pre-print, LANL archive, physics/0004057, 2000.

[2] James P. Crutchfield and Cosma Rohilla Shalizi. Thermodynamic depth of causal states: Objective complexity via minimal representations. *Physical Review E*, 59:275–283, 1999.

[3] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.

[4] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.

[5] James P. Crutchfield and Karl Young. Inferring statistical complexity. *Physical Review Letters*, 63:105–108, 1989.

[6] James P. Crutchfield. The calculi of emergence: Computation, dynamics, and induction. *Physica D*, 75:11–54, 1994. http://www.santafe.edu/projects/CompMech.

[7] Wesley C. Salmon. *Statistical Explanation and Statistical Relevance*. University of Pittsburgh Press, Pittsburgh, 1971. With contributions by Richard C. Jeffrey and James G. Greeno.

[8] Wesley C. Salmon. *Scientific Explanation and the Causal Structure of the World*. Princeton University Press, Princeton, 1984.