Thermodynamic depth of causal states: Objective complexity via minimal representations

James P. Crutchfield*
Physics Department, University of California, Berkeley, California 94720-7300
and the Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501

Cosma Rohilla Shalizi†
Physics Department, University of Wisconsin, Madison, Wisconsin 53706
and the Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501
(Received 17 August 1998)

Thermodynamic depth is an appealing but flawed structural complexity measure. It depends on a set of macroscopic states for a system, but neither its original introduction by Lloyd and Pagels nor any follow-up work has considered how to select these states. Depth, therefore, is at root arbitrary. Computational mechanics, an alternative approach to structural complexity, provides a definition for a system's minimal, necessary causal states and a procedure for finding them. We show that the rate of increase in thermodynamic depth, or *dive*, is the system's reverse-time Shannon entropy rate, and so depth only measures degrees of macroscopic randomness, not structure. To fix this, we redefine the depth in terms of the causal state representation—ϵ-machines—and show that this representation gives the minimum dive consistent with accurate prediction. Thus, ϵ-machines are optimally shallow. [S1063-651X(99)12401-2]

PACS number(s): 05.20.−y, 05.45.−a, 05.70.Ce

I. NATURAL COMPLEXITY

Dissipative dynamics, symmetry breaking, phase transitions, bifurcations, and pattern formation, acting over different temporal and spatial scales, at different levels and on different substrates, are presumably responsible for assembling and freezing in the wide diversity of structures observed in the natural world. Each of these processes has its more-or-less well-developed foundations. But where are the principles that define and describe their products? What is structure itself? Does each and every particular combination of forces lead to a different and unique class of natural structure, requiring its own vocabulary and theory? And, how do we detect that some new structure has emerged in the first place?

These and related questions about nature's complexity have engaged a large number of researchers for several decades now; for a sampling see, e.g., Refs. [1–5] and references therein. One focus has been on quantitative measures of the complexity of natural objects and of the processes that bring them into existence—measures that capture properties more interesting than mere randomness and disorder. Existing theory, such as is found in statistical mechanics, provides relatively well-understood measures of disorder in (say) temperature and thermodynamic entropy, and of the flow of energy that can do work in the various free energies. While many applications and problems remain, there is little pressing need for new conceptual approaches to randomness and energy transduction. However, when it comes to structure something is missing—something else must be invented and then added to physical theory to account for, work with, and quantify different kinds of structure.

One class of approaches to natural complexity is based on the theory of sequential discrete computation [6,7]—the theory of how sundry sorts of discrete-state devices process information at varying levels of sophistication. The resulting measures of complexity ultimately express structural properties in terms of universal Turing machines. Unfortunately, almost all interesting mathematical and quantitative questions about these measures of structure inherit the uncomputability associated with those all-powerful machines. More fundamentally, though, the idea that everything in the world is really a discrete-state computer strikes one as inadequate; at a minimum nature is parallel, continuous, spatially extended, noisy, and quantum mechanical.

Fortunately, in the thermodynamic depth of Lloyd and Pagels [8] we have a proposal for a noncomputation-theoretic, empirically calculable measure of the complexity of processes. One central motivation for defining the thermodynamic depth is that it is small both for regular and for random processes. Thus, one of its appealing features is that depth measures something other than randomness—a property already well-captured by both Kolmogorov-Chaitin complexity [9–11] and Shannon entropy rate [12–15].

In this paper we introduce the required background for thermodynamic depth [8] and for an alternative approach to natural complexity, called computational mechanics [16,17], that extends statistical mechanics to address issues of structure in a direct way. We review the definition of thermodynamic depth and apply it to several simple Markov processes, revealing several ambiguities. To remove them we redefine the depth in terms of a representation based on causal states, those states through which computational mechanics views the minimal structure of a system [16,17]. We then prove our main results on the predictive optimality and minimality of the causal state representation. Finally, we draw a number of conclusions about using thermodynamic

*Electronic address: chaos@santafe.edu
†Electronic address: shalizi@santafe.edu

depth as a measure of structural complexity in natural processes.

II. PROCESSES

Following Lloyd and Pagels, we focus on discrete-time processes and consider a given process as a joint probability distribution $\Pr(\ldots,X_{-1},X_0,X_1,\ldots)$ over random (''microscopic'') variables X_t at each time t that take values x_t in a continuous state space $\mathcal{X}$. In accord with experimental constraints, we assume that the process is not observed directly, but states are in fact measured via a finite-precision instrument. The result is that our description of the process is in all practicality a joint distribution over a chain $\overleftrightarrow{S}\equiv\cdots S_{-2}S_{-1}S_0S_1\cdots$ of discrete-valued random variables S_t that range over a finite set $\mathcal{A}$ of observed states. (Although our notation differs, this setup follows the account in Ref. [8], p. 194, of ''macroscopic,'' ''measured,'' or ''coarse-grained'' states as partitions of the underlying microscopic state space.)

We divide the chain into two semi-infinite halves by choosing a time t as the dividing point. Denote the past by

$$\overleftarrow{S}_t\equiv\cdots S_{t-3}S_{t-2}S_{t-1} \tag{1}$$

and the future by

$$\vec{S}_t\equiv S_tS_{t+1}S_{t+2}S_{t+3}\cdots. \tag{2}$$

We will assume that the observed process is described by a temporal shift-invariant measure μ on bi-infinite realizations $\cdots s_{-2}s_{-1}s_0s_1s_2\ldots,s_i\in\mathcal{A}$. The measure μ induces a family of distributions. Let $\Pr(s_t)$ denote the probability that at time t the random variable S_t takes on the particular value $s_t\in\mathcal{A}$ and $\Pr(s_{t+1},\cdots,s_{t+L})$ the joint probability over sequences of L consecutive measurements. Consistent with Ref. [8], we assume time-translation symmetry [18] and so $\Pr(s_{t+1},\ldots,s_{t+L})=\Pr(s_1,\ldots,s_L)$. We denote a sequence of L consecutive measurements by $S^L\equiv S_1\ldots S_L$; when looking to the future (past) the sequence is denoted $\vec{S}^L$ ($\overleftarrow{S}^L$). [In dropping the time index from Eqs. (1) and (2) we implicitly take $t=0$.] We shall follow the convention that a capital letter refers to a random variable, while a lowercase letter denotes a particular value of that variable. Hence, s^L will denote a particular measurement sequence of length L.

III. ENTROPY AND RANDOMNESS

The average uncertainty of an L sequence S^L is given by the Shannon entropy of the joint distribution $\Pr(S^L)$ [14]:

$$H[S^L]\equiv-\sum_{s^L\in\mathcal{A}^L}\Pr(s^L)\log_2\Pr(s^L). \tag{3}$$

Looking forward in time, the rate of increase of this uncertainty is defined by the entropy rate

$$\vec{h}_\mu\equiv\lim_{L\to\infty}\frac{H[\vec{S}^L]}{L}, \tag{4}$$

where μ denotes the above-mentioned measure. The quantity $\vec{h}_\mu$ measures the irreducible randomness in the generation of future behavior: the randomness that remains after the correlations over longer and longer futures are taken into account. The reverse-time entropy rate $\overleftarrow{h}_\mu$ is defined similarly in terms of $\overleftarrow{S}^L$ and measures historical randomness. Both can be expressed in terms of a conditional entropy: given knowledge of the measurement history, the uncertainty in the next measurement S_0 is

$$\vec{h}_\mu=H[S_0|\overleftarrow{S}]; \tag{5}$$

and similarly, given the future, we have

$$\overleftarrow{h}_\mu=H[S_{-1}|\vec{S}], \tag{6}$$

where the entropy of a random variable X conditioned on the value of another random variable Y is defined as $H[X|Y]\equiv H[X,Y]-H[Y]$.

IV. THERMODYNAMIC DEPTH

Lloyd and Pagels propose that the complexity of a macroscopic state $s\in\mathcal{A}$ is determined by the history that led to s. The motivation for this is that ''complexity must be a function of the *process*—the assembly routine—that brought the object into existence'' (emphasis theirs) ([8], p. 187); in particular, it is a ''measure of how hard it is to put something together'' ([8], p. 189). Starting from a distribution over macroscopic state sequences, one first finds the probability of length-L histories that end in state s:

$$\Pr(S_{-L+1},\ldots,S_{-1},S_0=s|s) \tag{7}$$

$$\equiv\frac{\Pr(S_{-L+1},\ldots,S_{-1},S_0=s)}{\Pr(s)}. \tag{8}$$

Then the thermodynamic L-depth $\mathcal{D}_L(s)$ of state s is defined by the conditional entropy

$$\mathcal{D}_L(s)\equiv H[S_{-L+1},\ldots,S_{-1},S_0=s|s]. \tag{9}$$

(From here on we ignore the distinction in Ref. [8] between ''depth'' and ''thermodynamic depth'' by, in effect, setting Boltzmann's constant to $1/\ln 2$.) Averaging over all such states gives one the L-depth $\mathcal{D}_L$ of the system as a whole:

$$\mathcal{D}_L\equiv\sum_{s\in\mathcal{A}}\Pr(s)\mathcal{D}_L(s), \tag{10}$$

or

$$\mathcal{D}_L=H[S_{-L+1},\ldots,S_{-1}|S_0], \tag{11}$$

where we have used the identity $H[X,Y|X]=H[Y|X]$. We define $\mathcal{D}_0=0$.

The backstage intuition motivating thermodynamic depth is the following: if there is little uncertainty about how to attain a macroscopic state and if trajectories are confined within narrow bounds, then the macroscopic state is easy to assemble. In this case the process leading to that state and generating those trajectories is simple and the state is shal-

low. If the historical uncertainty is large and if a wide range of historical alternatives has been excluded, then the process is complex and the macroscopic state is deep. ''The thermodynamic depth of a state b is proportional to the amount of information (in bits) needed to identify the trajectory that leads to b given the information that the system is in b'' ([8], p. 196).

Like all statistical complexity measures, thermodynamic depth has forsworn awarding high complexities to mere randomness. Reference [8] states that it vanishes for completely random processes, as well as for totally ordered ones ([8], pp. 187, 190, and 191). For systems satisfying the microcanonical assumption of statistical mechanics, Lloyd and Pagels ([8] pp. 190, 194, and 195) provide another expression for the depth, as the difference between a coarse-grained and a fine-grained thermodynamic entropy. Using this alternate expression, they argue that black holes ([8], p. 191), gases at thermal equilibrium ([8], p. 191), and salt crystals ([8], p. 191) are shallow and the self-assembly of protein complexes ([8], p. 196) is deep. While it is sometimes easier to evaluate the alternate expression than Eq. (11), it is strictly equivalent to the latter in the cases where the necessary (restrictive) conditions behind the former hold, so we shall confine ourselves to Eq. (11) in what follows.

The total depth, $\lim_{L\to\infty}\mathcal{D}_L$, of a process might as well be bottomless. Like L-depth, it depends on a baseline. That is, it depends on the time when we judge the process to have started and on the depth accumulated from the beginning of time until then. At best, these choices can be a bit tricky to figure out. Of greater physical significance, therefore, is the asymptotic rate v at which the depth increases, which we call *dive*:

$$v \equiv \lim_{L\to\infty}[\mathcal{D}_L - \mathcal{D}_{L-1}]. \tag{12}$$

The benefit of looking at a rate which is not considered in Ref. [8] is that v is independent of the origin of time and so allows one to more fairly compare processes by their rate of depth generation.

We now show that v is the reverse-time entropy rate. Recalling the definition of conditional entropy, $H[Y|X] = H[X,Y] - H[X]$, Eq. (12) becomes

$$v = \lim_{L\to\infty}\{H[S_{-L+1},\ldots,S_0] - H[S_0] - H[S_{-L+2},\ldots,S_0] + H[S_0]\} \tag{13}$$

$$= \lim_{L\to\infty}\{H[S_{-L+1},\ldots,S_0] - H[S_{-L+2},\ldots,S_0]\} \tag{14}$$

$$= \lim_{L\to\infty} H[S_{-L+1}|S_{-L+2},\ldots,S_0] \tag{15}$$

$$= H[S_{-L+1}|\vec{S}_{-L+2}] = H[S_{-1}|\vec{S}] \tag{16}$$

$$= \overleftarrow{h}_\mu, \tag{17}$$

where the next-to-last step follows from time-translation invariance.

For later use note that, since $H[Y] \geqslant H[Y|X]$, it follows from Eq. (16) and from translation invariance that

$$v \leqslant H[S_0]. \tag{18}$$

For stationary or asymptotically stationary processes, we have $H[S_{-L+2},\ldots,S_0] = H[S_{-L+1},\ldots,S_{-1}]$. Thus, starting from Eq. (14) we also conclude that

$$\lim_{L\to\infty}\{H[S_{-L+1},\ldots,S_0] - H[S_{-L+1},\ldots,S_{-1}]\} \tag{19}$$

$$= \lim_{L\to\infty} H[S_0|S_{-L+1},\ldots,S_{-1}] \tag{20}$$

$$= H[S_0|\overleftarrow{S}] \tag{21}$$

$$= \vec{h}_\mu. \tag{22}$$

From this we see that (i) the forward-time and reverse-time entropy rates are equal, $\overleftarrow{h}_\mu = \vec{h}_\mu$, and (ii) they are the same as the dive: $v = h_\mu$. (From here on we drop the time arrows and denote a process's entropy rate by h_μ.)

To summarize, we have shown that the Shannon entropy rate controls the average rate of increase in the thermodynamic depth and that the dive is invariant under time reversal. Recall that h_μ also controls the average rate of increase of Kolmogorov-Chaitin complexity [14]. These aspects of depth are not a surprise and are in accord with the claim that ''the average complexity of a state must be proportional to the Shannon entropy of the set of trajectories that experiment determines can lead to that state'' ([8], p. 190). From these elementary uses of information-theoretic identities, it is clear at this point that thermodynamic depth measures nothing other than the macroscopic randomness generated by a system.

V. SOMETHING ROTTEN IN THE STATES

The analysis of the preceding section leaves us with a puzzle: How is it that Lloyd and Pagels can state—e.g., on each of the first six pages of Ref. [8]—that depth discounts for disorder and so captures something other than randomness?

The problem, we claim, lies in their choice of states. In the illustrative examples in Ref. [8] macroscopic states are selected that support the desired properties of depth. That is, the results and interpretations do not follow from a direct application of the given definition of thermodynamic depth alone; biases external to the definition are invoked.

Moreover, employing an appropriate set of macroscopic states is crucial for obtaining a well-defined depth, since by judiciously redefining them one can give the depth any value from 0 on up. To see this, remember that the depth is the conditional entropy of a sequence of states. If there is only one state, the depth vanishes. If we make spurious macroscopic distinctions—e.g., acting as though one state was really n degenerate, equiprobable states—we add a contribution to the dive that is proportional to log$n$. And, we can keep doing this until the depth is as large as we like (cf. discussion in Ref. [8] leading up to the example on page 191).

The states of whichever dynamical system underlies the observed process are, at least, unambiguous candidates for use in the calculation of depth, but have an unfortunate habit of being unknown, redundant, or excessively fine grained. Lloyd and Pagels considered this problem by implication, discussing why, in some particular cases, certain choices of state are better than others. They explain, for instance, on page 191 of Ref. [8] how an unfortunate choice of measurements can make even systems in thermodynamic equilibrium quite deep. But they neither presented a procedure for picking sets of states nor gave general criteria for ranking possible alternative selections. This lack has not been remedied by follow-up work on thermodynamic depth, though commentary at that time by Landauer ([19], p. 307) raised related concerns.

Assuming one wants to use thermodynamic depth to measure complexity, Occam's razor [20] advises us to pick the simplest representation we can—in this case, whichever selection of states gives the smallest depth; cf. Ref. [8], p. 193. But this can always be trivially achieved by lumping everything into one state, as just noted, which gives a vanishing depth. More confusingly there are even cases, as we'll see a bit later, where such lumping is entirely appropriate.

Nor can the problem of state choice be reduced to that of coarse graining the space of observables; as done in Ref. [8], pp. 194 and 195 and elsewhere, for example in Refs. [21] and [22]. While this space can be readily represented by a finite alphabet, as done above—indeed, digital measuring devices so represent it without even asking permission—the problem is that the connection between what we measure and the underlying process is often obscure to the point of total darkness. (The definitions of ''measurement'' for Hamiltonian and quantum mechanical systems in Ref. [8] shed no light on this point.) It is certainly not desirable to conflate a process's complexity with the complexity of whatever apparatus connects the process to the variables we happen to have seized upon as handles.

One helpful step in developing any measure of complexity is that it be calculated on simple illustrative examples that can be thoroughly and unambiguously analyzed. We now proceed to do this for a series of examples—all of them based on Markov chains, if only to guarantee that nothing especially tricky or esoteric is at issue. In fact, we can interpret each example as a type of one-dimensional spin-1 statistical mechanical system; cf. Ref. [23]. (We emphasize that our results in other sections are not restricted to this class of Markov processes.)

The hidden Markov models we analyze contain a set of ''internal'' states, belonging to a finite alphabet $\mathcal{X}$, which are not directly observable. At each time step, there is some probability of moving from the current state to any other, while ''emitting'' an observable symbol drawn from another alphabet $\mathcal{A}$. We denote the probability of going from internal state i to internal state j while emitting the measurement value s as $T_{ij}^{(s)}$. These models thus generate a pair of linked stochastic processes, one over the internal states and the other over the observable values, and only the latter is directly detectable. Nonhidden Markov models are those where these two processes are one and the same: where $\mathcal{A}=\mathcal{X}$ and $T_{ij}^{(s)}=0$ unless $s=j$.

Consider first a nonhidden system of three states $\mathcal{X}=\mathcal{A}=\{\mathbf{A},\mathbf{B},\mathbf{C}\}$, each of which can go to any other, including itself, with equal probability; see Fig. 1. Here, according to the prescription of Lloyd and Pagels, $\mathcal{D}_L=L\log_2 3$, the total depth is infinite, and the dive is exactly equal to the entropy rate of the observable sequences, i.e., $v=h_\mu^{\mathcal{A}}=\log_2 3$ bits per step. The sequences generated are completely random, but neither the depth nor dive vanish.

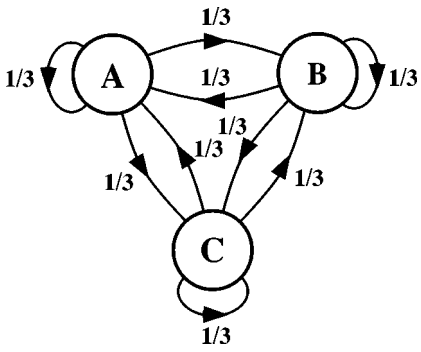

FIG. 1. A simple Markov chain that generates random sequences—**BBAC**$\cdots$—with finite dive ($v=\log_2 3$) and so infinite total depth ($\mathcal{D}_L\to_{L\to\infty}\infty$).

Next, we hide the internal states $\mathcal{X}$ from observation, but at each time step a measuring instrument emits one of two observable symbols $s\in\mathcal{A}=\{0,1\}$, as in Fig. 2. In this way we recover a simple version of the micromacroscopic distinction of Ref. [8]. The transition matrices $T_{ij}^{(s)}$ are, in this case,

$$T^{(0)}=\begin{bmatrix} 1/2 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1/2 \end{bmatrix} \tag{23}$$

and

$$T^{(1)}=\begin{bmatrix} 0 & 1/2 & 0 \\ 0 & 0 & 1/2 \\ 1/2 & 0 & 0 \end{bmatrix}. \tag{24}$$

That is, each state either loops back on itself, emitting $s=0$, or goes to the next state in the chain, emitting $s=1$, with equal probability. The dive, i.e., the entropy rate $h_\mu^{\mathcal{A}}$ of the observables, is $v=1$ bit per step. The entropy rate $h_\mu^{\mathcal{X}}$ of the

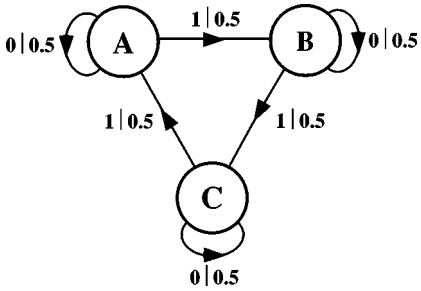

FIG. 2. A simple hidden Markov model that generates strings with finite dive ($v=h_\mu^{\mathcal{A}}=1$ bit per step) and infinite long-run depth. The edge notation $s|p$ denotes that a transition is to be taken with probability p, emitting measurement value s.

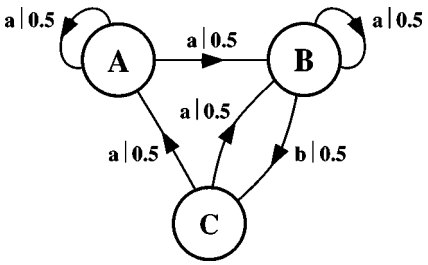

FIG. 3. A hidden Markov model of the logistic map symbolic dynamics observed with a nongenerating partition.

internal states is also 1 bit per step, since, given the current state, there are two possible, equiprobable successors. Moreover, while the system is a quite adequate source of random sequences, macroscopic states $s \in \mathcal{A}$, as well as the three hidden states, continue to deepen at the rate of 1 bit per step.

Note that by inserting additional states between **A** and **B**, which are equally likely to either loop back to themselves on $s=0$ or go to the next state in the chain on $s=1$, it is easy to go from Fig. 2 to ''Rube Goldberg'' automata. These are representations with elaborated sets of states with exactly the same observable process and properties (i.e., with the same $\Pr(\ldots,s_{-1},s_0,s_1,\ldots)$, where $s_t \in \{0,1\}$), but with increasing internal-state structure. Thus, there are inherent ambiguities in using inappropriately baroque sets of states when describing the structural properties of a process; ambiguities that must be addressed somehow.

Finally, consider the symbolic dynamics of the logistic map of the unit interval: $x_{t+1}=f(x_t)=4x_t(1-x_t)$. Here the microscopic state space is continuous: $x_t \in \mathcal{X}=[0,1]$, but we observe x_t with a binary-valued instrument $\mathcal{A}=\{\text{''}a\text{''} \sim [0,\hat{x}), \text{''}b\text{''} \sim [\hat{x},1]\}$, where $\hat{x}$ is the largest preimage of 1/2. When $x_t \in [0,\hat{x})$ the instrument emits $s=a$ and when $x_t \in [\hat{x},1]$ it emits $s=b$. This ''nongenerating'' partition of $\mathcal{X}$ leads to the three hidden states that are coarse grainings of $\mathcal{X}$: $\mathbf{A} \sim [0,1-\hat{x})$, $\mathbf{B} \sim [1-\hat{x},\hat{x})$, and $\mathbf{C} \sim [\hat{x},1]$. Recalling that we can calculate the invariant distribution $\Pr(x)$, the resulting stochastic finite-state model of the symbolic dynamics process is shown in Fig. 3. (See Refs. [17] and [24] for more discussion of this example.)

The transition matrices for this process are

$$T^{(b)}=\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1/2 \\ 0 & 0 & 0 \end{bmatrix} \tag{25}$$

and

$$T^{(a)}=\begin{bmatrix} 1/2 & 1/2 & 0 \\ 0 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \end{bmatrix}. \tag{26}$$

The entropy rate $h_\mu^{\mathcal{X}}$, measured over the states, is 1 bit per step, but the dive $v=h_\mu^{\mathcal{A}}$ (of the observables) is *lower*: $v \approx 0.811$ bits per step. The states, in other words, are actually worse—less predictable, deeper, and more demanding of memory (in a sense made precise presently)—than the surface phenomena (sequences over $\mathcal{A}$) they are supposed to explain. (Refs. [17] and [25] discuss this curious phenomenon. A detailed mathematical analysis is found in Ref. [24].)

This example illustrates the measurement dependency of both randomness and complexity. In contrast with the binary instrument just used, if the logistic map is observed with a generating partition, for which infinite *a*-*b* sequences are in correspondence with the microscopic states $x_t \in [0,1]$, there is only a single internal state. In this case, the internal state entropy rate $h_\mu^{\mathcal{X}}$ is zero and the entropy rate of the observed symbol sequences is $h_\mu^{\mathcal{A}}=1$ bit per symbol. It turns out that this is the correct description of the logistic map dynamics; see Ref. [26] for an elementary exposition.

Readers will have already noticed, and been troubled by, the fact that all our examples are simple sources of random strings, but have steep dives. According to the definition, they are deep, complex processes, despite the explicit statement of Lloyd and Pagels that depth is small or vanishes for random processes.

VI. CAUSAL STATES AND ϵ-MACHINES

On the one hand, what these examples make clear is that we generally will not find macroscopic states appropriate to measuring a process's statistical complexity just by translating observables (via coarse graining) into a finite alphabet. On the other hand, especially in experimental work, we often have no source of information other than the sequence of finite-precision discrete-valued observables. There is a fundamental difficulty here. Moreover, part of the attraction of thermodynamic depth, compared to (say) Kolmogorov-Chaitin complexity [9,10] and logical depth [27], was its claimed calculability from empirical data.

There is at least one release from these ambiguities: it is found in the use of *causal states*, as they are conceived of by *computational mechanics*—an extension of statistical mechanics that explicitly accounts for a process's structure [16,23]. From the viewpoint of an observer, the idea is that two trajectories leave one in the same causal state if they leave one equally knowledgeable as to the future. More formally, a causal state $\mathcal{S}$ is an equivalence class over histories $\overleftarrow{s}$ of observed states, such that all the sequences in the causal state give the same conditional distribution for the semi-infinite future $\overrightarrow{s}$:

$$\epsilon(\overleftarrow{s})=\{\overleftarrow{s}' \,|\, \forall \overrightarrow{s} \;\; \Pr(\overrightarrow{s}|\overleftarrow{s}')=\Pr(\overrightarrow{s}|\overleftarrow{s})\}. \tag{27}$$

The causal-state equivalence classes form a partition of the set $\overleftarrow{\mathbf{S}}$ of all histories; see Fig. 4. Thus defined, $\epsilon(\overleftarrow{s})$ is a function from a history $\overleftarrow{s}$ to a set of histories, which are the causal states $\boldsymbol{\mathcal{S}}_i$, $i=0,1,2,3,\ldots$. We denote the set $\{\mathcal{S}_i\}$ of all causal states by $\boldsymbol{\mathcal{S}}$. It is convenient sometimes to have a function taking one from a history $\overleftarrow{s}$ to the label i of its equivalence class and, in a slight abuse of notation, we will also call this $\epsilon(\overleftarrow{s})$.

Since we need *some* choice of state if we are to apply depth at all, and if we are not to consign it to the growing collection of subjective complexity measures (see Ref. [4]), we might as well select a process's causal states. What is

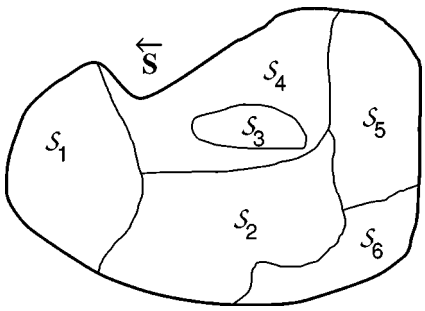

FIG. 4. A schematic representation of partitioning the set $\overleftarrow{\mathbf{S}}$ of all histories into causal states $\mathcal{S}_i$. Within each causal state all the individual histories $\overleftarrow{s}$ have the same conditional distribution $\Pr(\overrightarrow{S}|\overleftarrow{s})$ for future observables. Note that the $\mathcal{S}_i$ need not form compact sets; we have simply drawn them that way here for clarity.

notable, though, is that, while causal states were not designed with this end in mind, they minimize dive.

The representation of a process consisting of the causal states and their transitions is known as an ϵ-*machine*. In the simplest setting, an ϵ-machine is a Markov chain over a finite number of causal states and so can be compactly described by a labeled transition matrix $T_{ij}^{(s)}$, notationally similar to that for the examples above. This matrix can be calculated (analytically or empirically) from the distribution of observed sequences, a procedure called ϵ-machine *reconstruction*.

An ϵ-machine lets us calculate the probability of different sequences of observables. It also leads to an invariant probability distribution $\Pr(\mathcal{S})$ over the causal states. The resulting complexity measure for a process is the *statistical complexity* C_μ that is defined simply as the Shannon entropy of that distribution [16]: $C_\mu = H[\mathcal{S}]$. C_μ measures the average amount of historical information stored in the current state. Our results in Sec. VII are not, however, restricted to cases where the ϵ-machine is finite Markovian, merely to ones where there is a probability measure over the causal states.

A process's thermodynamic depth, and thus its dive, are defined with reference to its macroscopic states, whatever we take those to be. Due to the ambiguities that follow from a prosaic interpretation of depth's definition we propose to redefine depth, and by implication the dive, solely in terms of a process's causal states. The first result of taking the ''macroscopic'' states to be the causal states is that the dive is the entropy rate of the ϵ-machine's internal-state process: $v \equiv h_\mu^{\mathcal{X}}$, where $\mathcal{X}=\mathcal{S}$. The second result is that by Eq. (18) $v \leqslant C_\mu$. In fact, $v < C_\mu$, if there is any mutual information in the observed sequences $\overleftrightarrow{S}$, by Eq. (106) in Ref. [23].

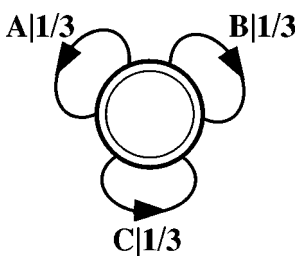

FIG. 5. The ϵ-machine for the unhidden Markov model of Fig. 1. The internal entropy rate $h_\mu^{\mathcal{X}}$ and the statistical complexity C_μ vanish since there is a single causal state.

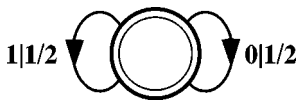

FIG. 6. The ϵ-machine for the hidden Markov model of Fig. 2. The internal entropy rate $h_\mu^{\mathcal{X}}$ and the statistical complexity C_μ again vanish since there is a single causal state.

Causal-state equivalence-classing guarantees that the ϵ-machine is as small as it can be and still be an accurate predictor of future observed sequences; see Sec. VII B below. This makes ϵ-machines for both highly ordered and highly random sequences very simple: a high degree of randomness means that many distinct sequences of observables leave one equally uncertain about the future and, consequently, those sequences all leave the system in the same causal state. In this way one *derives* the desired ''boundary conditions'' for statistical complexity measures—low both for ordered and for random processes—from the underlying principle of optimal prediction; that is, from Eq. (27).

These properties of causal states suffice to rescue the complexity analysis of the examples from the confusions of the last section. The first (Fig. 1) corresponds to an ϵ-machine with a single causal state $\mathcal{S}_0$ that returns to itself on three separate, equally probable symbols $\mathcal{A}=\{\mathbf{A},\mathbf{B},\mathbf{C}\}$ (see Fig. 5). The entropy rate $h_\mu^{\mathcal{A}}$ of the observed sequences is (as always) preserved under the change of representation to causal states, but the entropy rate $h_\mu^{\mathcal{S}}$ of the causal state process itself, i.e., the now-redefined dive v, is, like the statistical complexity, zero.

A similar fate awaits our second example (Fig. 2). Under causal-state equivalence-classing, the three alleged states collapse into one, yielding an ideal coin-tossing machine with a single state and two transitions (see Fig. 6). Here again the statistical complexity and the new dive vanish. Defining depth in terms of a process's causal states leads us, in both examples, to recover the intuitively correct notion that these sources of purely random sequences are neither structurally complex nor store much information about their history.

In our final example (Fig. 3), the future conditional distribution of observables depends only on how long it has been since the last b, leading to a countable infinity of causal states (see Fig. 7). It turns out that the new dive and the statistical complexity can be analytically calculated; one finds $v \approx 0.677\,867$ bits per measurement and $C_\mu \approx 2.711\,47$

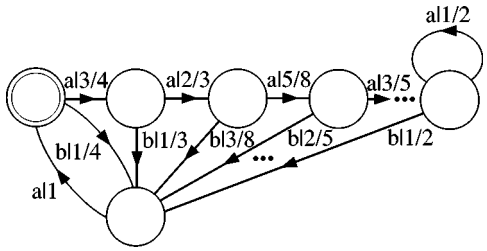

FIG. 7. The ϵ-machine for the hidden Markov model of Fig. 3 has a countable infinity of causal states. The internal entropy rate $h_\mu^{\mathcal{X}}$ and statistical complexity C_μ are both positive, indicating that this is an intrinsically more complex process than the other two examples.

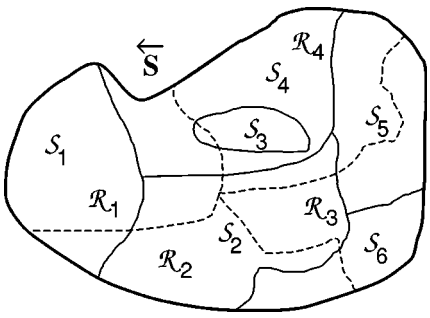

FIG. 8. An alternative set $\{\mathcal{R}_i\}$ of states that partition $\overleftarrow{\mathbf{S}}$ overlaid on the causal states. (The $\mathcal{R}_i$ are delineated by dashed lines.) The collection of all such alternative partitions form Occam's ''pool.'' Note again that the $\mathcal{R}_i$ need not be compact.

bits of historical memory are stored by the process [17,24]. It is a more complex process than the other two examples.

One of the desired properties of thermodynamic depth was that it accounted for the history of the ''assembly process'' ([8], pp. 187–189 and *passim*). We should emphasize that by definition causal states account for a form of historical memory, though in an importantly different way. Causal states measure the amount of historical information stored in a system.

VII. OPTIMAL SHALLOWNESS OF ϵ-MACHINES

Working with the ϵ-machine representation forces one to distinguish between (1) sequences over coarse-grained observables $\mathcal{A}$, (2) sequences over causal states $\boldsymbol{\mathcal{S}}$, and (3) sequences over transitions, the labeled edges $\{(i,j,s): T_{ij}^{(s)} > 0\}$. There is a many-to-one relation between edge sequences and causal-state sequences and also between edge sequences and observable sequences. But, as we saw when we defined the causal states as equivalence classes, Eq. (27), there is a function that takes a history to a causal state: namely, $\epsilon: \overleftarrow{\mathbf{S}} \mapsto \boldsymbol{\mathcal{S}}$. One consequence is that one can specify all of the relevant historical information by noting which of the causal states the process is in, rather than recounting a possibly infinite amount of information from the history $\overleftarrow{S}$ that led to the current state. That is, causal states provide a compression of a process's history.

These distinctions and the historical compression are good motivations for deciding which type of state to use for a process. But these alone are not enough, so let us consider alternatives to causal states, namely, another set $\boldsymbol{\mathcal{R}}$ of states, call them $\mathcal{R}$-states, that are determinable from observed sequences and that, like causal states, partition $\overleftarrow{\mathbf{S}}$; see Fig. 8. We assume that these rivals to the ϵ-machine are, like the ϵ-machine itself, restricted to using only the past history of observables in their predictions, without any other hints.

As one ranges over alternative choices of state—swimming around in Occam's pool of possible partitions—we will show that the ϵ-machine has a threefold optimality: (i) no set of $\mathcal{R}$-states is more informative about future observables than the causal states; of those choices of states that are as predictive as the causal states, none has (ii) a smaller statistical complexity nor (iii) a smaller entropy rate over the internal states. We conclude that none of the alternatives, if used to calculate the depth, would give us a shallower dive than the causal states. We will prove these in order.

A. Nothing forecasts better than an ϵ-machine

Call the sequence of observables up to the present time $\overleftarrow{S}$, the random variable that is the next observable S, and the random variable that is the whole sequence of future observables $\vec{S}$. Recall that the function $\epsilon: \overleftarrow{\mathbf{S}} \mapsto \boldsymbol{\mathcal{S}}$ returns the causal state the ϵ-machine is in after observing $\overleftarrow{S}$ and define the function $\eta: \overleftarrow{\mathbf{S}} \mapsto \boldsymbol{\mathcal{R}}$ similarly for the $\mathcal{R}$-states. We measure the forecasting ability of a set of states by $H[\vec{S}|\mathcal{R}]$ [28], the uncertainty that remains in the future observables once we know the current state. That is, the better the set of states is at forecasting—the more prescient it is—the smaller this uncertainty. From Eq. (27) it follows that

$$\Pr(\vec{S}|\epsilon(\overleftarrow{S})) = \Pr(\vec{S}|\overleftarrow{S}), \tag{28}$$

and so

$$H[\vec{S}|\epsilon(\overleftarrow{S})] = H[\vec{S}|\overleftarrow{S}]. \tag{29}$$

Since, for any random variables X and Y and function f,

$$H[Y|f(X)] \geqslant H[Y|X], \tag{30}$$

it follows that

$$H[\vec{S}|\eta(\overleftarrow{S})] \geqslant H[\vec{S}|\overleftarrow{S}] \tag{31}$$

$$= H[\vec{S}|\epsilon(\overleftarrow{S})] \tag{32}$$

and so

$$H[\vec{S}|\mathcal{R}] \geqslant H[\vec{S}|\mathcal{S}]. \tag{33}$$

Thus, no alternative set $\boldsymbol{\mathcal{R}}$ of states sees the future better than the causal states.

In what follows, we will put a hat over the name of any rival set of states that is as predictive as the causal states, i.e., we refer to states in $\hat{\boldsymbol{\mathcal{R}}}$ if and only if $H[\vec{S}|\hat{\mathcal{R}}] = H[\vec{S}|\mathcal{S}]$.

B. Nothing as prescient as an ϵ-machine is simpler

Suppose we have a set $\hat{\boldsymbol{\mathcal{R}}}$ of states for which $H[\vec{S}|\hat{\mathcal{R}}] = H[\vec{S}|\mathcal{S}]$. Then, because the causal states are equivalence classes with respect to future conditional probabilities, the $\hat{\mathcal{R}}$-states must be refinements of these classes. That is, rather than the situation depicted in Fig. 8, we have the $\hat{\mathcal{R}}$-partitioning shown in Fig. 9. Otherwise at least one $\hat{\mathcal{R}}_i$, considered as a set, would have to include histories that belonged to at least two distinct causal states. Such mixing of causal states can only increase the uncertainty about the future sequence $\vec{S}$ of observables. That is, for every $\hat{\mathcal{R}}_i$ there is a $\mathcal{S}_j$ such that $\hat{\mathcal{R}}_i \subseteq \mathcal{S}_j$ and so every causal state is a union of $\hat{\mathcal{R}}$-states.

The result is that the causal state is a function of the $\hat{\mathcal{R}}$-state: $\mathcal{S} = g(\hat{\mathcal{R}})$. Thus,

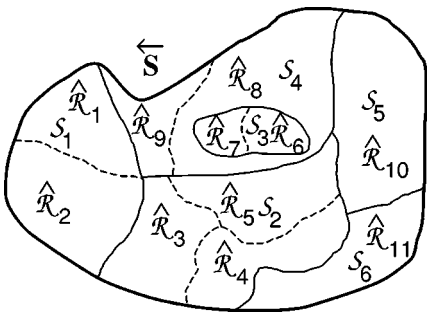

FIG. 9. Any alternative partition that is as prescient as the causal states must be a refinement of the causal-state partition. That is, each $\hat{\mathcal{R}}_i$ must be a (possibly improper) subset of some $\mathcal{S}_j$. Otherwise, at least one $\hat{\mathcal{R}}_i$ would have to contain parts of at least two causal states. And so using this $\hat{\mathcal{R}}_i$ to predict the future observables leads to more uncertainty about $\vec{S}$ than using the causal states.

$$H[\mathcal{S}] = H[g(\hat{\mathcal{R}})] \leqslant H[\hat{\mathcal{R}}]. \tag{34}$$

But $H[\mathcal{S}]$ is C_μ, the statistical complexity of the ϵ-machine, whereas $H[\hat{\mathcal{R}}]$ is the statistical complexity of the η-machine—the set of $\hat{\mathcal{R}}$ states and their transitions. Thus, of the optimally predictive alternative representations the ϵ-machine is the smallest, as measured by C_μ.

An argument exactly parallel to the one in the preceding subsection shows, when applied to the equally prescient alternatives, that

$$H[\vec{S}|\hat{\mathcal{R}}] = H[\vec{S}|\mathcal{S}] \Rightarrow H[\vec{S}^L|\hat{\mathcal{R}}] = H[\vec{S}^L|\mathcal{S}], \tag{35}$$

for $L = 1,2,\ldots$ (the opposite implication is not true, however). Thus, the causal states are also at least as informative about the next (single) observable S as any rival and, for that matter, about any finite subsequence $\vec{S}^L$ of the future. However, in the general case of the previous paragraphs it is necessary to consider the whole semi-infinite future because, potentially, coarser partitions can match these finite-L predictive powers. If, for instance, two histories have the same distribution for S, but different distributions over the whole future, they belong to different causal states. An $\hat{\mathcal{R}}$-state that combined those two causal states, however, would enjoy the same ability to predict S and its η-machine would have a smaller statistical complexity.

C. Nothing as prescient as an ϵ-machine has a smaller dive

We will now show that the ϵ-machine's dive ($v = h_\mu^{\mathcal{S}}$) is at least as small as that of any equally prescient alternative. This also turns on the fact that such $\hat{\mathcal{R}}$-states are refinements of the causal states. The ϵ-machine is deterministic in the sense of automata theory [23]; that is, the present state $\mathcal{S}$ and the next observable S together fix the next state $\mathcal{S}'$, and so $H[\mathcal{S}'|S,\mathcal{S}] = 0$. Thus, we have

$$H[S|\mathcal{S}] = H[\mathcal{S}',S|\mathcal{S}]. \tag{36}$$

The $\hat{\mathcal{R}}$-machine, however, is not necessarily deterministic in this sense, but all entropies are non-negative, so $H[\hat{\mathcal{R}}'|S,\hat{\mathcal{R}}] \geqslant 0$. Since we are considering alternatives with the same predictive power as the ϵ-machine, i.e., alternatives for which $H[\vec{S}|\mathcal{S}] = H[\vec{S}|\hat{\mathcal{R}}]$, then we have $H[S|\mathcal{S}] = H[S|\hat{\mathcal{R}}]$. On the one hand,

$$H[\hat{\mathcal{R}}',S|\hat{\mathcal{R}}] = H[S|\hat{\mathcal{R}}] + H[\hat{\mathcal{R}}'|S,\hat{\mathcal{R}}] \tag{37}$$

$$\geqslant H[S|\hat{\mathcal{R}}] \tag{38}$$

$$= H[S|\mathcal{S}] \tag{39}$$

$$= H[\mathcal{S}',S|\mathcal{S}] \tag{40}$$

$$= H[\mathcal{S}'|\mathcal{S}] + H[S|\mathcal{S}',\mathcal{S}]. \tag{41}$$

On the other hand,

$$H[\hat{\mathcal{R}}',S|\hat{\mathcal{R}}] = H[\hat{\mathcal{R}}'|\hat{\mathcal{R}}] + H[S|\hat{\mathcal{R}}',\hat{\mathcal{R}}], \tag{42}$$

as well, so

$$H[\hat{\mathcal{R}}'|\hat{\mathcal{R}}] + H[S|\hat{\mathcal{R}}',\hat{\mathcal{R}}] \geqslant H[\mathcal{S}'|\mathcal{S}] + H[S|\mathcal{S}',\mathcal{S}], \tag{43}$$

or

$$H[\hat{\mathcal{R}}'|\hat{\mathcal{R}}] - H[\mathcal{S}'|\mathcal{S}] \geqslant H[S|\mathcal{S}',\mathcal{S}] - H[S|\hat{\mathcal{R}}',\hat{\mathcal{R}}]. \tag{44}$$

Since a causal state is a function of an $\hat{\mathcal{R}}$-state, the transition pair $(\mathcal{S}',\mathcal{S})$ is a function of the transition pair $(\hat{\mathcal{R}}',\hat{\mathcal{R}})$, implying that $H[S|\mathcal{S}',\mathcal{S}] \geqslant H[S|\hat{\mathcal{R}}',\hat{\mathcal{R}}]$. Thus, the right-hand side of Eq. (44) is non-negative and this implies that

$$H[\hat{\mathcal{R}}'|\hat{\mathcal{R}}] \geqslant H[\mathcal{S}'|\mathcal{S}], \tag{45}$$

which is the desired result; namely, $v^{\hat{\mathcal{R}}} \geqslant v^{\mathcal{S}}$. That is, nothing that predicts as well as the ϵ-machine has a smaller dive than the ϵ-machine does.

VIII. CONCLUSION

If one prefers processes over static descriptions and dislikes pretending every natural thing is a digital computer, thermodynamic depth seemed to be an attractive complexity measure: ''one of the remarkably few thrusts in this area that is not conspicuously vacuous,'' in the words of Landauer [19]. Since total depth most likely shares the incalculability, though not the formal uncomputability, of Kolmogorov-Chaitin complexity and logical depth, it is not, at face value, physically significant. Dive, the rate at which depth increases, is both calculable and significant. We showed that dive is the reverse-time Shannon entropy rate of the stochastic process over the macroscopic states one takes the system to be in. With nothing else said or added, however, depth typically measures historical randomness; as do Kolmogorov-Chaitin complexity and the Shannon entropy rate.

Unfortunately, Ref. [8], which introduced depth, gave no clue as to how macroscopic states are to be selected; though it strongly suggested this is simply a matter of coarse-graining the space of microscopic states; cf. [8], pp. 194-195. As we have shown, this approach produces manifestly ambiguous results.

By way of fixing depth, we highlighted the key role of the choice of macroscopic states. The causal states of computational mechanics do not suffer from the defects and ill-definedness that led to trouble with other sorts of states. The procedure that identifies them, ϵ-machine reconstruction, also gives us a way to calculate depth and dive. We removed depth's ambiguities and recovered its claimed features by redefining it in terms of the causal states.

We then gave our main results, showing that no alternative set of states to the causal states contains more information about the future of observables. Moreover, unless an alternative throws some of that information away it cannot have a smaller statistical complexity or a lower dive. Thus, ϵ-machines are optimally shallow.

ACKNOWLEDGMENTS

We thank David Feldman and Deirdre des Jardins for helpful discussions. We also appreciate comments on the manuscript by Dave Feldman, Jon Fetter, and Nigel Snoad. This work was supported at U.C. Berkeley by ONR Grant No. N00014-95-1-0524, at the Santa Fe Institute by ONR Grant No. N00014-95-1-0975, and by Sandia National Laboratory Contract No. AU-4978.

[1] P. Grassberger, Int. J. Theor. Phys. **25**, 907 (1986).
[2] K. Lindgren and M. G. Nordahl, Complex Syst. **2**, 409 (1988).
[3] B. Wackerbauer, A. Witt, H. Atmanspacher, J. Kurths, and H. Scheingraber, Chaos Solitons Fractals **4**, 133 (1994).
[4] D. P. Feldman and J. P. Crutchfield, Phys. Lett. A **238**, 244 (1998).
[5] R. Badii and A. Politi, *Complexity: Hierarchical Structures and Scaling in Physics* (Cambridge University Press, Cambridge, 1997).
[6] J. E. Hopcroft and J. D. Ullman, *Introduction to Automata Theory, Languages, and Computation* (Addison-Wesley, Reading, Massachusetts, 1979).
[7] C. H. Papadimitriou, *Computational Complexity* (Addison-Wesley, Reading, Massachusetts, 1994).
[8] S. Lloyd and H. Pagels, Ann. Phys. **188**, 186 (1988).
[9] A. N. Kolmogorov, Probl. Inf. Transm. **1**, 1 (1965).
[10] G. Chaitin, J. Assoc. Comput. Mach. **13**, 145 (1966).
[11] M. Li and P. M. B. Vitanyi, *An Introduction to Kolmogorov Complexity and its Applications* (Springer-Verlag, New York, 1993).
[12] C. E. Shannon, Bell System Tech. J., **27**, 379 (1948), in *The Mathematical Theory of Communication*, edited by C. E. Shannon and W. Weaver (University of Illinois Press, Urbana Champaign, 1963).
[13] A. N. Kolmogorov, Dokl. Akad. Nauk. SSSR **124**, 754 (1959).
[14] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).
[15] A. A. Brudno, Trans. Moscow Math. Soc. **44**, 127 (1983).
[16] J. P. Crutchfield and K. Young, Phys. Rev. Lett. **63**, 105 (1989).
[17] J. P. Crutchfield, Physica D **75**, 11 (1994).
[18] For the dissipative, Hamiltonian, and quantum systems considered, Ref. [8], pp. 194 and 195, assumes, moreover, that the probabilities over the microscopic states are uniform with respect to Lebesgue measure and that the probabilities of sequences over the coarse-grained state space are time invariant.
[19] R. Landauer, Nature (London) **336**, 306 (1988).
[20] William of Ockham, *Philosophical Writings: A Selection, Translated, with an Introduction, by Philotheus Boehner, O.F.M., Late Professor of Philosophy, The Franciscan Institute* (Bobbs-Merrill, Indianapolis, 1964).
[21] Y. C. Zhang, J. Phys. I **1**, 971 (1991).
[22] H. C. Fogedby, J. Stat. Phys. **69**, 411 (1992).
[23] D. P. Feldman and J. P. Crutchfield, Santa Fe Institute Working Paper 98-04-026, http://www.santafe.edu/projects/CompMech/papers/DNCO.html
[24] D. R. Upper, Ph.D. thesis, University of California, Berkeley, 1997.
[25] J. P. Crutchfield, Phys. Lett. A **171**, 52 (1992).
[26] R. L. Devaney, *A First Course in Chaotic Dynamical Systems: Theory and Experiment* (Addison-Wesley, Reading, Massachusetts, 1992).
[27] C. H. Bennett, Found. Phys. **16**, 585 (1986).
[28] This quantity can be infinite. In this case, we should use the set of $H[\vec{S}^L|\mathcal{R}]$ for all L. Optional predictors are those that simultaneously minimize these entropies for all values of L. The resulting proofs are essentially similar, but slightly more involved.