Pattern Discovery and Computational Mechanics

Cosma Rohilla Shalizi* and James P. Crutchfield
Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501
Electronic addresses: {shalizi,chaos}@santafe.edu
(27 January 2000)

Computational mechanics is a method for discovering, describing and quantifying patterns, using tools from statistical physics. It constructs optimal, minimal models of stochastic processes and their underlying causal structures. These models tell us about the intrinsic computation embedded within a process—how it stores and transforms information. Here we summarize the mathematics of computational mechanics, especially recent optimality and uniqueness results. We also expound the principles and motivations underlying computational mechanics, emphasizing its connections to the minimum description length principle, PAC theory, and other aspects of machine learning.

Santa Fe Institute Working Paper 00-01-008
Keywords: pattern discovery, machine learning, computational mechanics, information, complexity, ϵ-machines

Contents

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

I. INTRODUCTION

All students of machine learning are familiar with pattern recognition; in this paper we wish to introduce a new term for a related, relatively under-recognized concept, *pattern discovery*, and a way of tackling such problems, *computational mechanics*.

The term pattern *discovery* is meant to contrast with both pattern *recognition* and pattern *learning*. In pattern recognition, the goal of the system is to accurately assign inputs to pre-set categories. In most learning systems, the goal is to determine which of several pre-set categorization schemes is correct. (Naturally, the two tasks are closely connected (Vapnik, 1995).) In either case, the representations used have been, as it were, handed down from on high, due to choices external to the recognition and learning procedures.

In pattern *discovery*, however, the aim is to avoid, so far as possible, such *a priori* assumptions about what structures are relevant. This is, of course, an ancient problem, and one which has not been ignored in machine learning. While there are ingenious schemes for pattern discovery via trial and error, some even informed by empirical psychology (Holland et al., 1986), we believe that a more direct approach is not only possible but also illuminates the ideal results and the limitations of all pattern-discovery methods.

Computational mechanics originated in physics as a complementary approach to statistical mechanics for dealing with complex, organized systems (Crutchfield, 1994). In such systems the "forward" approach of statistical mechanics—deriving macroscopic properties from the interactions of microscopic components—is often intractable, though data can be had in abundance.[1] Computational mechanics follows an "inverse" strategy, extending the idea of extracting "geometry from a time series" (Packard et al., 1980). It builds the simplest model

[1] But see Chaikin and Lubensky (1995) and Cross and Hohenberg (1993) for organized systems where the "forward" approach works.

capable of capturing the patterns in the data—a representation of the causal structure of the hidden process which generated the observed behavior.[2] In a sense that will be made clear as we go on, this representation—the ϵ-machine—is the unique maximally efficient model of the observed data-generating process. The basic ideas of computational mechanics were introduced in Crutchfield and Young (1989). Since then they have been used to analyze dynamical systems, cellular automata, hidden Markov models, evolved spatial computation, stochastic resonance, globally coupled maps, and the dripping faucet experiment; see Shalizi and Crutchfield, (1999, Sec. 1) for references.

This paper is arranged as follows. First we examine some conceptual issues about pattern discovery and the way they are addressed by computational mechanics. We devote the bulk of this paper to a summary of the mathematical structure of computational mechanics, with particular attention to optimality and uniqueness theorems. Results are stated without proof; readers will find a full treatment in Shalizi and Crutchfield (1999). Then we discuss the ties between computational mechanics and several approaches to machine learning. Finally, we close by pointing out directions for future theoretical work.

II. CONCEPTUAL ISSUES

Any approach to handling patterns should, we claim, meet a number of criteria; the justifications for which are given in Shalizi and Crutchfield (1999) in detail. It should be at once

1. *Predictive*, i.e., the models it produces should allow us to predict the original process or system we are trying to understand; and, by that token, provide a compressed description of it;
2. *Computational*, showing how the process stores, transmits, and transforms information;
3. *Calculable*, analytically or by systematic approximation;
4. *Causal*, telling us how instances of the pattern are actually produced; and
5. *Naturally stochastic*, not merely tolerant of noise but explicitly formulated in terms of ensembles.

In any modeling approach, the two (related) problems are to devise a mapping from states of the world (or, more modestly, states of inputs) to states of the model, and to accurately and precisely predict future states of the world on the basis of the evolution of the model. (Cf. Holland et al. (1986) on "q-morphisms".) The key idea of computational mechanics is that the information required to do this is actually *in* the data, provided there is enough of it. In fact, if we go about it right, the key step is getting the mapping from data to model states right—equivalently, the problem is to decide which data-sets should be treated as equivalent and how data should be partitioned. Once we have the correct mapping of data into equivalence classes, accurate prediction is actually fairly simple. That the correct mapping should treat as equivalent all data-sets which leave us in the same degree of knowledge about the future has a certain intuitive plausibility, but also sounds hopelessly vague. In fact, we can specify such a partition in a precise, operational way, show that it is the best one to use, and determine it empirically. We call the function which induces that partition ϵ, and its equivalence classes *causal states*. In fact, the model we get from using such a partition — the ϵ-machine — meets all the criteria stated above. It is because the ϵ-machine shows, in a very direct way, how information is stored in the process, and how that stored information is transformed by new inputs and by the passage of time, that computational mechanics is about *computation*.

III. MATHEMATICAL DEVELOPMENT

A. Note on Information Theory

The bulk of the following development will be consumed with notions and results from information theory. We follow the standard definitions and notation of Cover and Thomas (1991), to which we refer readers unfamiliar with the theory. In particular, $H[X]$ is the entropy of the discrete random variable X, interpreted as the uncertainty in X, measured in bits.[3] $H[X|Y]$ is the entropy of X conditional on Y, and $I[X;Y]$ the mutual information between the two random variables.

B. Hidden Processes

We restrict ourselves to discrete-valued, discrete-time stationary stochastic processes. (See Sec. V A for discussion of these restrictions.) Intuitively, such processes are sequences of random variables S_i, the values of which are drawn from a countable set $\mathcal{A}$. We let i range over all the integers, and so get a bi-infinite sequence

[2] See Feldman and Crutchfield (1998) for an example of using both statistical and computational mechanics to analyze the same physical system.

[3] Here, and throughout, we follow the convention of using capital letters to denote random variables and lower-case letters their particular values.

$\overleftrightarrow{S} = \ldots S_{-1}S_0S_1\ldots$. In fact, we define a *process* in terms of the distribution of such sequences.

Given that $\overleftrightarrow{S}$ is well-defined, there are probability distributions for sequences of every finite length. Let $\overrightarrow{S}^L_t$ be the sequence of $S_t, S_{t+1}, \ldots, S_{t+L-1}$ of L random variables beginning at S_t. $\overrightarrow{S}^0_t \equiv \lambda$, the null sequence. Likewise, $\overleftarrow{S}^L_t$ denotes the sequence of L random variables going up to S_t, but not including it: $\overleftarrow{S}^L_t = \overrightarrow{S}^L_{t-L}$. Both $\overrightarrow{S}^L_t$ and $\overleftarrow{S}^L_t$ take values from $s^L \in \mathcal{A}^L$. Similarly, $\overrightarrow{S}_t$ and $\overleftarrow{S}_t$ are the semi-infinite sequences starting from and stopping at t and taking values $\overrightarrow{s}$ and $\overleftarrow{s}$, respectively.

Requiring the process S_i to be stationary means that

$$\mathrm{P}(\overrightarrow{S}^L_t = s^L) = \mathrm{P}(\overrightarrow{S}^L_0 = s^L) \ , \tag{1}$$

for all $t \in \mathbb{Z}$, $L \in \mathbb{Z}^+$, and all $s^L \in \mathcal{A}^L$. (A stationary process is one that is time-translation invariant.) Consequently, $\mathrm{P}(\overrightarrow{S}_t = \overrightarrow{s}) = \mathrm{P}(\overrightarrow{S}_0 = \overrightarrow{s})$ and $\mathrm{P}(\overleftarrow{S}_t = \overleftarrow{s}) = \mathrm{P}(\overleftarrow{S}_0 = \overleftarrow{s})$, and so the subscripts may be dropped.

C. Effective States

Our goal is to predict all or part of $\overrightarrow{S}$ using some function of some part of $\overleftarrow{S}$. We begin by taking the set $\overleftarrow{\mathbf{S}}$ of all pasts and partitioning it into mutually exclusive and jointly comprehensive subsets. That is, we make a class $\boldsymbol{\mathcal{R}}$ of subsets of pasts. (See Fig. 1.) Each $\rho \in \boldsymbol{\mathcal{R}}$ will be called a *state* or an *effective state*. When the current history $\overleftarrow{s}$ is included in the set ρ, we will say the process is in state ρ. Thus, there is a function from histories to effective states:

$$\eta : \overleftarrow{\mathbf{S}} \mapsto \boldsymbol{\mathcal{R}} \ . \tag{2}$$

An individual history $\overleftarrow{s} \in \overleftarrow{\mathbf{S}}$ maps to a specific state $\rho \in \boldsymbol{\mathcal{R}}$; the random variable $\overleftarrow{S}$ for the past maps to the random variable $\mathcal{R}$ for the effective states.

Any function defined on $\overleftarrow{\mathbf{S}}$ will serve to partition that set: we just assign to the same ρ all the histories $\overleftarrow{s}$ on which the function takes the same value. (Similarly, any equivalence relation on $\overleftarrow{\mathbf{S}}$ partitions it.) Each effective state has a well-defined conditional distribution of futures, though not necessarily a unique one. Specifying the effective state thus amounts to making a prediction about the process's future. In this way, the framework formally incorporates traditional methods of time-series analysis.

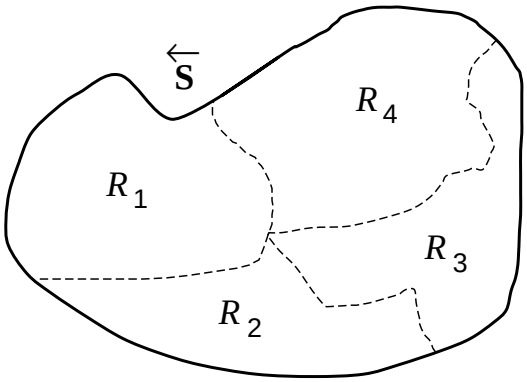

FIG. 1. A schematic picture of a partition of the set $\overleftarrow{\mathbf{S}}$ of all histories into some class of effective states: $\boldsymbol{\mathcal{R}} = \{\mathcal{R}_i : i = 1, 2, 3, 4\}$. Note that the $\mathcal{R}_i$ need not form compact sets; we simply draw them that way for clarity. One should have in mind Cantor sets or other more pathological structures.

D. Patterns in Ensembles

It will be convenient to have a way of talking about the uncertainty of the future. We do not want to use $H[\overrightarrow{S}]$, since that is infinite in general. Instead, we will work with $H[\overrightarrow{S}^L]$, the uncertainty of the next L symbols, treated as a function of L.

Definition 1 (Capturing a Pattern) $\boldsymbol{\mathcal{R}}$ *captures a pattern iff there exists an* L *such that*

$$H[\overrightarrow{S}^L|\mathcal{R}] < LH[S] \ . \tag{3}$$

$\boldsymbol{\mathcal{R}}$ captures a pattern when it tells us something about how the distinguishable parts of a process affect each other: $\boldsymbol{\mathcal{R}}$ exhibits their dependence. (We also speak of η as capturing a pattern.) The smaller $H[\overrightarrow{S}^L|\mathcal{R}]$, the stronger the pattern captured by $\boldsymbol{\mathcal{R}}$. Our first result bounds how strongly $\boldsymbol{\mathcal{R}}$ can capture a process's pattern.

Lemma 1 *For all* $\boldsymbol{\mathcal{R}}$ *and for all* $L \in \mathbb{Z}^+$,

$$H[\overrightarrow{S}^L|\mathcal{R}] \geq H[\overrightarrow{S}^L \mid \overleftarrow{S}] \ . \tag{4}$$

E. Minimality and Prediction

Let's invoke Occam's Razor: "It is vain to do with more what can be done with less". To use the razor, we have to fix what is to be "done" and what "more" and "less" mean. The job we want done is accurate prediction; i.e., to reduce the conditional entropies $H[\overrightarrow{S}^L|\mathcal{R}]$ as

far as possible, down to the bound set by Lemma 1. But we want to do this as simply as possible, with as few resources as possible. To meet both constraints—minimal uncertainty and minimal resources—we will need a measure of the second. Since $\mathrm{P}(\overleftarrow{S}=\overleftarrow{s})$ is well defined, it induces a probability distribution on the η-states, and we can set up the following measure of resources.

Definition 2 (Complexity of State Classes) *The statistical complexity of a class $\boldsymbol{\mathcal{R}}$ of states is*

$$\begin{aligned} C_\mu(\boldsymbol{\mathcal{R}}) &\equiv H[\mathcal{R}] \\ &= -\sum_{\rho\in\mathcal{R}} \mathrm{P}(\mathcal{R}=\rho)\log_2 \mathrm{P}(\mathcal{R}=\rho) \ , \end{aligned} \tag{5}$$

when the sum converges to a finite value.

$C_\mu(\boldsymbol{\mathcal{R}})$ is the average uncertainty in the process's current state $\mathcal{R}$. This is the same as the average amount of memory (in bits) that the process *appears* to retain about the past, given the chosen state class $\boldsymbol{\mathcal{R}}$. We wish to do with as little of this memory as possible. Our objective, then, is to find a state class which minimizes C_μ, subject to the constraint of maximally accurate prediction.

F. Causal States

Definition 3 (A Process's Causal States) *The* causal states *of a process are the members of the range of the function $\epsilon : \overleftarrow{\mathbf{S}} \mapsto 2^{\overleftarrow{\mathbf{S}}}$—the power set of $\overleftarrow{\mathbf{S}}$:*

$$\begin{aligned} \epsilon(\overleftarrow{s}) \equiv \{\overleftarrow{s}' | \mathrm{P}(\overrightarrow{S}=\overrightarrow{s} \ | \overleftarrow{S}=\overleftarrow{s}) = \mathrm{P}(\overrightarrow{S}=\overrightarrow{s} \mid \overleftarrow{S}=\overleftarrow{s}') \ , \\ \text{for all } \overrightarrow{s}\in\overrightarrow{S}, \overleftarrow{s}' \in \overleftarrow{S}\} \ , \end{aligned} \tag{6}$$

that maps from histories to classes of histories. We write the i^{th} causal state as $\mathcal{S}_i$ and the set of all causal states as $\boldsymbol{\mathcal{S}}$; the corresponding random variable is denoted $\mathcal{S}$ and its realization σ.

The cardinality of $\boldsymbol{\mathcal{S}}$ is unrestricted. $\boldsymbol{\mathcal{S}}$ can be finite, countably infinite, a continuum, a Cantor set, or something stranger still.[4]

We could equally well define an equivalence relation $\sim_\epsilon$ such that two histories are equivalent iff they have the same conditional distribution of futures, and define causal states as the equivalence classes of $\sim_\epsilon$. (In fact, this was the original approach of Crutchfield and Young (1989).) Either way, we break $\overleftarrow{\mathbf{S}}$ into parts that leave us in different conditions of ignorance about the future.

Each causal state $\mathcal{S}_i$ has a conditional distribution of futures, $\mathrm{P}(\overrightarrow{S} \ |\mathcal{S}_i)$. It follows directly from Def. 3 that no two states have the same distribution of futures; this is not true of effective states in general. Another immediate consequence of that definition is that

$$\mathrm{P}(\overrightarrow{S}=\overrightarrow{s} \ |\mathcal{S} = \epsilon(\overleftarrow{s})) = \mathrm{P}(\overrightarrow{S}=\overrightarrow{s} \mid \overleftarrow{S}=\overleftarrow{s}). \tag{7}$$

Again, this is not generally true of effective states.

G. Causal State-to-State Transitions

The causal state at any given time and the next value of the observed process together determine a new causal state (Lemma 2 below, which doesn't rely on the following). Thus, there is a natural relation of succession among the causal states.

Definition 4 (Causal Transitions) *The labeled transition probability $T_{ij}^{(s)}$ is the probability of making the transition from state $\mathcal{S}_i$ to state $\mathcal{S}_j$ while emitting the symbol $s \in \mathcal{A}$:*

$$T_{ij}^{(s)} \equiv \mathrm{P}(\mathcal{S}' = \mathcal{S}_j, \ \overrightarrow{S}^1 = s | \mathcal{S} = \mathcal{S}_i) \ , \tag{8}$$

where $\mathcal{S}$ is the current causal state and $\mathcal{S}'$ its successor on emitting s. We denote the set $\{T_{ij}^{(s)} : s \in \mathcal{A}\}$ by $\mathbf{T}$.

H. ϵ-Machines

Definition 5 (An ϵ-Machine Defined) *The* ϵ-machine *of a process is the ordered pair $\{\epsilon, \mathbf{T}\}$, where ϵ is the causal state function and $\mathbf{T}$ is set of the transition matrices for the states defined by ϵ.*

Lemma 2 (ϵ-Machines Are Deterministic) *For each $\mathcal{S}_i$ and $s \in \mathcal{A}$, $T_{ij}^{(s)} > 0$ only for that $\mathcal{S}_j$ for which $\epsilon(\overleftarrow{s}s) = \mathcal{S}_j$ iff $\epsilon(\overleftarrow{s}) = \mathcal{S}_i$, for all pasts $\overleftarrow{s}$.*

"Deterministic" is meant in the sense of automata-theory, not dynamics.

Lemma 3 (Causal States Are Independent) *The probability distributions over causal states at different times are conditionally independent.*

This indicates that the causal states, considered as a process, define a kind of Markov chain. We say "kind of" since the class of ϵ-machines is substantially richer than the one normally associated with Markov chains (Crutchfield, 1994; Upper, 1997).

Definition 6 (ϵ-Machine Reconstruction) ϵ-Machine reconstruction *is any procedure that given a process $\mathrm{P}(\overleftrightarrow{S})$, or an approximation of $\mathrm{P}(\overleftrightarrow{S})$, produces the process's ϵ-machine $\{\epsilon, \mathbf{T}\}$.*

[4] Examples of all of these are given in Crutchfield (1994) and Upper (1997).

Given a mathematical description of a process, one can often calculate analytically its ϵ-machine. (For example, see the computational mechanics analysis of statistical mechanical spin systems in Feldman and Crutchfield (1998).) There are also algorithms that reconstruct ϵ-machines from empirical estimates of $P(\overleftrightarrow{S})$. Those used in Crutchfield (1994), Crutchfield and Young (1989), Hanson (1993), and Perry and Binder (1999), operate in batch mode, taking the raw data as a whole and producing the ϵ-machine. Others could work on-line, taking in individual measurements and re-estimating the set of causal states and their transition probabilities.

I. Optimalities and Uniqueness

Theorem 1 (Causal States are Maximally Prescient) *For all $\boldsymbol{\mathcal{R}}$ and all $L \in \mathbb{Z}^+$,*

$$H[\overrightarrow{S}^L|\mathcal{R}] \geq H[\overrightarrow{S}^L|\mathcal{S}] = H[\overrightarrow{S}^L \mid \overleftarrow{S}] \ . \tag{9}$$

Causal states are as good at predicting the future—are as *prescient*—as complete histories. Since the causal states can be systematically approximated, we have shown that the upper bound on the strength of patterns (Def. 1 and Lemma 1) can in fact be reached.

All subsequent results concern rival states that are as prescient as the causal states. We call these *prescient rivals* and denote a class of them $\hat{\boldsymbol{\mathcal{R}}}$.

Definition 7 (Prescient Rivals) Prescient rivals $\hat{\boldsymbol{\mathcal{R}}}$ *are states that are as predictive as the causal states; viz., for all $L \in \mathbb{Z}^+$,*

$$H[\overrightarrow{S}^L|\hat{\mathcal{R}}] = H[\overrightarrow{S}^L|\mathcal{S}] \ . \tag{10}$$

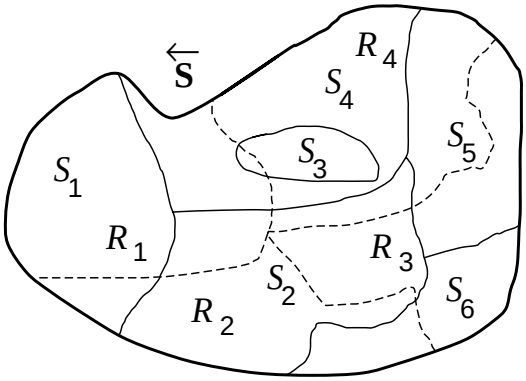

FIG. 2. An alternative class $\boldsymbol{\mathcal{R}}$ of states (delineated by dashed lines) that partition $\overleftarrow{\mathbf{S}}$ overlaid on the causal states $\boldsymbol{\mathcal{S}}$ (solid lines). Here, for example, $\mathcal{S}_2$ contains parts of $\mathcal{R}_1$, $\mathcal{R}_2$, $\mathcal{R}_3$ and $\mathcal{R}_4$. Note again that the $\mathcal{R}_i$ need not be compact nor simply connected, as drawn.

Lemma 4 (Refinement Lemma) *For all prescient rivals $\hat{\boldsymbol{\mathcal{R}}}$ and for each $\hat{\rho} \in \hat{\boldsymbol{\mathcal{R}}}$, there is a $\sigma \in \boldsymbol{\mathcal{S}}$ and a measure-0 subset $\hat{\rho}_0 \subset \hat{\rho}$, possibly empty, such that $\hat{\rho} \setminus \hat{\rho}_0 \subseteq \sigma$, where $\setminus$ is set subtraction.*

The lemma becomes more intuitive if we ignore for a moment the measure-0 set $\hat{\rho}_0$ of histories. It then says that any alternative partition $\hat{\boldsymbol{\mathcal{R}}}$ 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 contain parts of at least two causal states. Therefore, using $\hat{\mathcal{R}}_i$ to predict the future observables would lead to more uncertainty about $\overrightarrow{S}$ than using the causal states. (Compare Fig. 3 with Fig. 2.) Because the histories in $\hat{\rho}_0$ have zero probability, treating them the "wrong" way makes no discernible difference to predictions.

Theorem 2 (Causal States Are Minimal) *For all prescient rivals $\hat{\boldsymbol{\mathcal{R}}}$,*

$$C_\mu(\hat{\boldsymbol{\mathcal{R}}}) \geq C_\mu(\boldsymbol{\mathcal{S}}) \ . \tag{11}$$

If we were trying to predict, not the whole of $\overrightarrow{S}$ but some limited piece $\overrightarrow{S}^L$, the causal states might not be the simplest ones with full predictive power. For any value of L, however, the states constructed by analogy to the causal states—the "truncated causal states"—have maximal prescience and minimal C_μ.

The minimality theorem licenses the following definition.

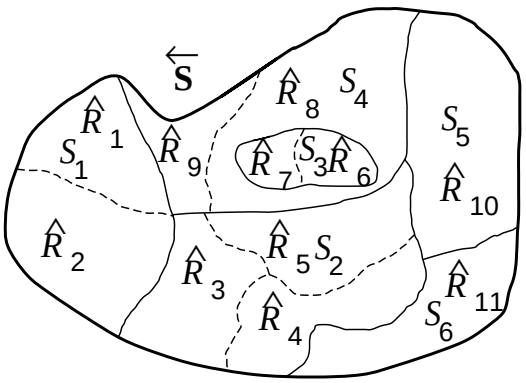

FIG. 3. A prescient rival partition $\hat{\boldsymbol{\mathcal{R}}}$ must be a refinement of the causal-state partition *almost everywhere*. Almost all of each $\hat{\mathcal{R}}_i$ must lie within some $\mathcal{S}_j$; the exceptions, if any, are a set of histories of measure 0. Here for instance $\mathcal{S}_2$ contains the positive-measure parts of $\hat{\mathcal{R}}_3$, $\hat{\mathcal{R}}_4$, and $\hat{\mathcal{R}}_5$. One of these rival states, say $\hat{\mathcal{R}}_3$, could have member-histories in any or all of the other causal states, if the total measure of these exceptional histories is zero.

Definition 8 (Statistical Complexity of a Process) *The statistical complexity $C_\mu(\mathcal{O})$ of a process $\mathcal{O}$ is that of its causal states: $C_\mu(\mathcal{O}) \equiv C_\mu(\boldsymbol{\mathcal{S}})$.*

Theorem 3 (Causal States Are Unique) *For all prescient rivals $\hat{\boldsymbol{\mathcal{R}}}$, if $C_\mu(\hat{\boldsymbol{\mathcal{R}}}) = C_\mu(\boldsymbol{\mathcal{S}})$, then there exists an invertible function between $\hat{\boldsymbol{\mathcal{R}}}$ and $\boldsymbol{\mathcal{S}}$ that almost always preserves equivalence of state: $\hat{\boldsymbol{\mathcal{R}}}$ and η are the same as $\boldsymbol{\mathcal{S}}$ and ϵ, respectively, except on a set of histories of measure 0.*

The remarks on Lemma 4 also apply to the ineliminable but immaterial measure-0 caveat here.

Theorem 4 (ϵ-Machines Are Minimally Stochastic) *For all prescient rivals $\hat{\boldsymbol{\mathcal{R}}}$,*

$$H[\hat{\mathcal{R}}'|\hat{\mathcal{R}}] \geq H[\mathcal{S}'|\mathcal{S}] \ , \tag{12}$$

where $\mathcal{S}'$ and $\hat{\mathcal{R}}'$ are the next causal state of the process and the next η-state, respectively.

Finally, we relate C_μ to an information theoretic quantity that is often used to measure complexity.

Definition 9 (Excess Entropy) *The* excess entropy $\mathbf{E}$ *of a process is the mutual information between its semi-infinite past and its semi-infinite future:*

$$\mathbf{E} \equiv I[\overrightarrow{S}; \overleftarrow{S}] \ . \tag{13}$$

Excess entropy is regularly re-introduced into the complexity-measure literature, as "predictive information", "stored information", "effective measure complexity", and so on (Shalizi & Crutchfield, 1999, Sec. VI). As these names indicate, it is tempting to see $\mathbf{E}$ as the amount of information stored in a process (which accounts for its popularity). According to the following theorem this temptation should be resisted.

Theorem 5 *The statistical complexity C_μ bounds the excess entropy $\mathbf{E}$:*

$$\mathbf{E} \leq C_\mu(\boldsymbol{\mathcal{S}}) \ , \tag{14}$$

with equality iff $H[\mathcal{S}|\overrightarrow{S}] = 0$.

$\mathbf{E}$ is thus only a lower bound on the true amount of information stored in the process, namely $C_\mu(\boldsymbol{\mathcal{S}})$.

IV. RELATIONS TO OTHER FIELDS

A. Computational and Statistical Learning Theory

The goal of computational learning theory (Kearns & Vazirani, 1994; Vapnik, 1995) is to identify algorithms that quickly, reliably, and simply lead to good representations of a target concept, usually taken to be a dichotomy of a feature or input space. Particular attention is paid to "probably approximately correct" (PAC) procedures (Valiant, 1984): those having a high probability of finding a close match to the target concept among members of a fixed representation class. The key word here is "fixed". While taking the representation class as a given is in line with implicit assumptions in most of mathematical statistics, it seems dubious when analyzing learning in the real world (Crutchfield, 1994; Boden, 1994).

In any case, such an assumption is clearly inappropriate if our goal is pattern discovery, and it was *not* made in the preceding development. While we plan to make every possible use of the results of computational learning theory in ϵ-machine reconstruction, we feel this theory is more properly a part of statistical inference and, particularly, of algorithmic parameter estimation, than of pattern discovery *per se*.

B. Formal Language Theory and Grammatical Inference

It is well known that formal languages can be classified into a hierarchy, the higher levels of which have strictly greater expressive power. The denizens of the lowest level of the hierarchy, the regular languages, correspond to finite-state machines and to hidden Markov models of finite dimension. In such cases, relatives of our minimality and uniqueness theorems are well known, and the construction of causal states is analogous to Nerode equivalence classing (Hopcroft & Ullman, 1979). Our theorems, however, are *not* restricted to this setting.

The problem of learning a language from observational data has been extensively studied by linguists and computer scientists. Unfortunately, good learning techniques exist only for the two lowest classes in the hierarchy, the regular and the context-free languages. (For a good account of these procedures see Charniak (1993).) Adapting this work to the reconstruction of ϵ-machines should be a useful area for future research.

C. The Minimum Description-Length Principles

Rissanen's *minimum description-length* (MDL) principle, best presented in his book (1989), is a way of picking the most concise generative model out of a chosen family of models that are all statistically consistent with given data. The MDL approach starts from Shannon's results on the connection between probability distributions and codes (Shannon & Weaver, 1963).

Suppose we choose a class $\mathcal{M}$ of models and are given data set x. The MDL principle tells us to use the model $\mathrm{M} \in \mathcal{M}$ that minimizes the sum of the length of the description of x given M, plus the length of description of

M given $\mathcal{M}$. The description length of x is taken to be $-\log \mathrm{P}(x|\mathrm{M})$. The description length of M may be regarded as either given by some coding scheme or, equivalently, by some distribution over the members of $\mathcal{M}$.

Though the MDL principle was one of the inspirations of computational mechanics, our approach to pattern discovery does not fit within Rissanen's framework. To mention only the most basic differences: We have no fixed class of models $\mathcal{M}$; we do not use encodings of rival models or prior distributions over them; and $C_\mu(\boldsymbol{\mathcal{R}})$ is not a description length.

D. Connectionist Models

Neural networks engaged in unsupervised learning (Becker, 1991) are often effectively doing pattern discovery. Certainly Hebb (1949) had this aim in mind when proposing his learning rule. While such networks certainly can discover regularities and covariations, they often represent them in ways baffling to humans. ϵ-Machines present the structures discovered by reconstruction in a clear and distinct way, but the learning dynamics are not (currently) as well understood as those of neural networks; see Sec. V B below.

V. CONCLUDING REMARKS

A. Limitations of the Current Results

We made some restrictive assumptions in our development above. Here we mention them in order of increasing severity and consider what may be done to lift them.

1. We know exact joint probabilities over sequence blocks of all lengths for a process. The cure for this is ϵ-machine reconstruction and in the next subsection we sketch work (underway) on a statistical theory of reconstruction.

2. The observed process takes on discrete values. This can probably be addressed with only a modest cost in increased mathematical subtlety, since the information-theoretic quantities we have used also exist for continuous variables. Many of our results appear to carry over to the continuous setting.

3. The process is discrete in time. This looks similarly solvable, since continuous-time stochastic process theory is moderately well developed. It may involve sophisticated probability theory or functional analysis, however.

4. The process is a pure time series; e.g., without spatial extent. There are already tricks to make spatially extended systems look like time series. Basically, one looks at all the paths through space-time, treating each one as if it were a time series. While this works well for data compression (Lempel & Ziv, 1986), it may not be satisfactory for capturing structure (Feldman, 1998).

5. The process is stationary. It's unclear how best to relax the assumption of stationarity. There are several straightforward ways of doing so, but it is unclear how much substantive content these extensions have. In any case, a systematic classification of non-stationary processes is (at best) in its infant stages.

B. Directions for Future Work

Two broad avenues for research present themselves.

First, we have the mathematics of ϵ-machines themselves. Assumption-lifting extensions have just been mentioned but there are many other ways to go. One which is especially interesting in the machine-learning context is the trade-off between prescience and complexity. For a given process there is a sequence of optimal machines connecting the one-state, zero-complexity machine with minimal prescience to the ϵ-machine. Each step on the path is the minimal machine for a certain degree of prescience; it would be very interesting to know what, if anything, we can say in general about the shape of this "prediction frontier".

Second, there is ϵ-machine reconstruction. As we remarked (p. 4), there are already several algorithms for reconstructing machines from data. What we need is knowledge of the *error statistics* (Mayo, 1996) of different reconstruction procedures, of the kinds of mistakes they make and the probabilities with which they make them. Ideally, we want to find "confidence regions" for the products of reconstruction: calculating the probabilities of different degrees of reconstruction error for a given volume of data or the amount of data needed to be confident of a fixed bound on the error. An analytical theory has been developed for the expected error in reconstructing certain kinds of processes (Crutchfield & Douglas, 1999). The results are encouraging enough that work is underway on a general theory of statistical inference for ϵ-machines—a theory analogous to what already exists in computational learning theory and grammatical inference.

Acknowledgments

This work was supported at the Santa Fe Institute under the Computation, Dynamics, and Inference Program via NSF grant IIS-9705830, the Novel Computation Program via a grant from AFOSR, and the Network Dynamics Program via support from Intel Corporation.

REFERENCES

Becker, S. (1991). Unsupervised learning procedures for neural networks. *International Journal of Neural Systems*, *2*, 17–33.

Boden, M. A. (1994). Precis of *The Creative Mind: Myths and Mechanisms*. *Behaviorial and Brain Sciences*, *17*, 519–531.

Chaikin, P. M., & Lubensky, T. C. (1995). *Principles of condensed matter physics*. Cambridge, England: Cambridge University Press.

Charniak, E. (1993). *Statistical language learning*. Language, Speech and Communication. Cambridge, Massachusetts: Bradford Books/MIT Press.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: John Wiley & Sons, Inc.

Cross, M. C., & Hohenberg, P. (1993). Pattern Formation Out of Equilibrium. *Reviews of Modern Physics*, *65*, 851–1112.

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

Crutchfield, J. P., & Douglas, C. (1999). Imagined complexity: Learning a random process. in preparation.

Crutchfield, J. P., & Young, K. (1989). Inferring statistical complexity. *Physical Review Letters*, *63*, 105–108.

Feldman, D. P. (1998). *Computational mechanics of classical spin systems*. Doctoral dissertation, University of California, Davis. Available on-line at http://hornacek.coa.edu/dave/Thesis/thesis.html.

Feldman, D. P., & Crutchfield, J. P. (1998). Discovering non-critical organization: Statistical mechanical, information theoretic, and computational views of patterns in simple one-dimensional spin systems. *Journal of Statistical Physics*, *submitted*. Santa Fe Institute Working Paper 98-04-026, http://www.santafe.edu/projects/CompMech/papers/ DNCO.html.

Hanson, J. E. (1993). *Computational mechanics of cellular automata*. Doctoral dissertation, University of California, Berkeley.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. New York: Wiley.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning, and discovery*. Computational Models of Cognition and Perception. Cambridge, Massachusetts: MIT Press.

Hopcroft, J. E., & Ullman, J. D. (1979). *Introduction to automata theory, languages, and computation*. Reading: Addison-Wesley. Second edition of *Formal Languages and Their Relation to Automata*, 1969.

Kearns, M. J., & Vazirani, U. V. (1994). *An introduction to computational learning theory*. Cambridge, Massachusetts: MIT Press.

Lempel, A., & Ziv, J. (1986). Compression of two-dimensional data. *IEEE Transactions in Information Theory*, *IT-32*, 2–8.

Mayo, D. (1996). *Error and the growth of experimental knowledge*. Science and Its Conceptual Foundations. Chicago: University of Chicago Press.

Packard, N. H., Crutchfield, J. P., Farmer, J. D., & Shaw, R. S. (1980). Geometry from a time series. *Physical Review Letters*, *45*, 712–716.

Perry, N., & Binder, P.-M. (1999). Finite statistical complexity for sofic systems. *Physical Review E*, *60*, 459–463.

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry*. Singapore: World Scientific.

Shalizi, C. R., & Crutchfield, J. P. (1999). Computational mechanics: Pattern and prediction, structure and simplicity. *Communications of Mathematical Physics*, *submitted*. SFI Working Paper 99-07-044, cond-mat/9907176.

Shannon, C. E., & Weaver, W. (1963). *The mathematical theory of communication*. Urbana, Illinois: University of Illinois Press.

Upper, D. R. (1997). *Theory and algorithms for hidden Markov models and generalized hidden markov models*. Doctoral dissertation, University of California, Berkeley.

Valiant, L. (1984). A theory of the learnable. *Communications of the Association for Computing Machinery*, *27*, 1134–1142.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Berlin: Springer-Verlag.