

INFERRING THE DYNAMIC, QUANTIFYING PHYSICAL COMPLEXITY*

James P. Crutchfield
Physics Department
University of California
Berkeley, California 94720 USA
chaos@gojira.berkeley.edu

SUMMARY

Through its formalization of inductive inference, computational learning theory provides a foundation for the inverse problem of chaotic data analysis: inferring the deterministic equations of motion underlying observed random behavior in physical systems. Integrating the geometric and statistical techniques of dynamical systems with learning theory provides a framework for consistently, although not absolutely, distinguishing between deterministic chaos and extrinsic fluctuations at a given level of computational resources. Two approaches to the inverse problem, estimating symbolic equations of motion and reconstructing minimal automata from chaotic data series, are reviewed from this point of view. With an inferred model dynamic the dynamical entropies and dimensions can be estimated. More interestingly, its structural properties give a measure of the intrinsic computational complexity of the underlying process.

INTRODUCTION

During the last decade or so a number of mechanisms have been investigated by which physical processes can generate complex behavior. The most widely studied of these, deterministic *chaos*, while *globally* stable to perturbations produces noisy behavior by exponential *local* amplification of microscopic fluctuations.¹ The central problem in applying the theory of dynamical systems to the sciences is identifying the deterministic, and possibly chaotic, component in a set of observations and distinguishing it from behavior due to ever-present measurement uncertainty, extrinsic noise, and uncontrolled degrees of freedom. This is the inverse problem in nonlinear dynamics: inferring the deterministic equations of motion, if any, underlying observed random behavior in physical systems.

Over a similar historical interval, computational learning theory has formalized a range of learning paradigms for inductive inference. In this it provides a language and a collection of complexity theoretic methods appropriate to the inverse problem. When integrated with the geometric and statistical techniques of dynamical systems theory, the result is

* Portions of this essay were distributed as "Learning the Dynamic", an extended abstract submitted 15 April 1989 to the Conference on Computational Learning Theory to be held 31 July - 2 August 1989, University of California, Santa Cruz. This work was supported by ONR contract N00014-86-K-0154.

a framework for consistently distinguishing between deterministic chaotic behavior and extrinsic information sources to which it is coupled.

This essay gives a brief overview of two complementary approaches to the inverse problem. The first estimates symbolic equations of motion using Bayesian statistical inference² and the second uses techniques from stochastic grammatical inference to reconstruct minimal computational models of chaotic behavior.^{3,4,5} They have two common goals. The first is to quantify the observed complexity in a data stream. The second is to capture the underlying dynamics in a form that can be related to first principles and used in forecasting and numerical modeling of the behavior. The following discussion emphasizes their similarities as problems in learning theory. It concludes by indicating where learning theory can contribute to a number of existing problems and remarks on the limitations of existing proposals for quantifying physical complexity. The hope is not only that dynamical systems will benefit from this synthesis but that with real world, quantitative problems computational learning theory will be made more generally accessible.

INDUCTIVE INFERENCE FORMALISM

Given the inverse problem of chaotic data analysis just outlined, this section reviews the appropriate tools from inductive inference and attempts to justify the choices made along the way from a scientific and dynamical systems viewpoint.

The types of problems of interest to physicists include the onset and structure of various forms of fluid turbulence and the occurrence of noise in superconducting Josephson junctions and other nonlinear solid state devices; just to mention a very few out of hundreds of applications in physics, chemistry, and biology. These systems generate very complex behavior and no one expects to exactly describe their detailed behavior. Nonlinear dynamicists are (or should be) satisfied with with a theory that *systematically* approximates some phenomenon. This means that with a given level of experimental limitations \dot{I} † and fixed data reduction (computational) resources C , observations can be explained quantitatively to within an error level δ . Furthermore, the error must vanish exponentially fast with increasing \dot{I} and C .

This is a form of *identification in the limit*⁶ with the additional requirement of estimating the *rate* of convergence of the inference method on an ensemble of data sets. Here, the inductive process continues indefinitely and its asymptotic behavior is used as the success criterion. The identification error is to decrease on larger and more accurate example sets.

The inverse problem can be considered also as a sequence inference problem since the example presentations are ordered by time. Typically, there are two problems in sequence inference. The first is to infer a model, such as the graph of the dynamic; this is the identification problem. The second is to forecast future observations knowing past ones from the sequence; this is the prediction problem. The formal restrictions that distinguish these subproblems in general sequence inference are not appropriate here. For the inverse problem identification and prediction are the same; the identified model can be used for forecasting.

To specify the inverse problem as a problem in inductive inference several components need to be defined. An inference method M is a computable process implemented (say) on a Turing machine. The latter is augmented to read in example data and output hypothesized models. To formalize this we must specify a success criterion C and a data representation D for the examples. The space of inference methods is then the set of all triplets $(M,$

† Specified in terms of sampling resolution and frequency as a information acquisition rate.

C, D). The basic components of this inductive inference problem are a rule space, an hypothesis space, an example set, an inference method, and its success criterion. These are defined as follows.

1. The *rule space*: This is the set of all mappings $\vec{F} : M \rightarrow M$ from the state space M into itself.
2. The *hypothesis space*: This is the space $H = \left\{ D : D = (\vec{F}, \vec{P}, M) \right\}$ of noisy discrete time dynamical systems

$$\vec{x}_{n+1} = \vec{F}(\vec{x}_n) + \vec{\xi}_n$$

where \vec{F} is a real vector-valued function, the *dynamic*; $\vec{\xi}_n$ is a realization of the real vector-valued random process \vec{P} ; and $\vec{x}_n \in M$, the manifold of states or *state space*.

3. The *example set*: The data with which the inference method works is obtained from a single realization $\{\vec{x}_n : n = 0, 1, 2, \dots, N - 1\}$ or *orbit* of the noisy dynamical system by way of a measurement partition $P_\epsilon = \{e_i : e_i \subset M\}$. By coarse graining the state space into size ϵ cells, the partition maps the sequence of states along an orbit into a sequence of symbols, $s \in \mathbf{A}$ in the measurement *alphabet* $\mathbf{A} = \{i : i = 0, \dots, k - 1; k = f(\epsilon^{-1})\}$. These symbols label the partition elements in which the orbit is found at the time of measurement. The measurement sequence obtained this way will be referred to as the *measurement language* L_{D, P_ϵ} . A set of length n subsequences is derived from a sliding window of length n onto the single long sequence. The length n measurement sequences $s^n = \{s_0 \cdots s_{n-1} : s_i \in \mathbf{A}\}$ are called *n-cylinders* in reference to the bundle of orbits that lead to the same set of n measurements.⁷ The examples will be restricted to only those measurement sequences that are observed. That is, the inference method will use only *positive* presentations. An *admissible presentation* P of the rule, then, will be a set $P = \{s^n : i = 0, 1, \dots, N - n\}$ of n -cylinders that are observed and that are ordered by time. The latter is guaranteed by the sliding window construction of the individual cylinders from the single long measurement sequence. In fact, we could have taken for a presentation a sequence of examples derived from orbits with different initial conditions rather than a sequence from a single orbit. In any case, the initial conditions \vec{x}_0 might contain transients. In the following, though, the orbit is assumed to be a *typical* orbit governed by the asymptotic probability measure on the attractor.
4. The class of *inference methods*: An inference method M takes an admissible presentation P of the measurement language and produces a model D of a noisy dynamical system: $D = M(P)$.
5. The *success criterion*: The most appropriate criterion is that developed by Wharton⁸ for approximate language identification. This requires a measure of goodness of fit of the inferred models to the measurement language. For this, we need a metric on the space \mathbf{L} of measurement languages. The weight $w(s^n)$ of an n -cylinder is the number of occurrences of that particular sequence in the presentation. The weights are effectively normalized $\sum_{s^n \in P} w(s^n) = \|P\| = N - n$. The distance between two languages $L_1, L_2 \in \mathbf{L}$ is given by the metric on \mathbf{L}

$$d(L_1, L_2) = \sum_{s \in L_1 \Delta L_2} w(s)$$

where $L_1 \Delta L_2$ is the symmetric difference of the two languages. Using this metric, an inference method M δ -identifies a measurement language L if and only if M converges to a grammar G associated with the hypothesized D such that

$$d(L(G), L) < \delta \cdot N$$

It is an encouraging result that if one tolerates finite error ($\delta > 0$) then the class of all languages is δ -identifiable in the limit from positive presentations by a method that infers only grammars for finite languages.⁸ Requiring that the error vanish in the limit leads to a more restricted form of identification. *Convergent* identification in the limit occurs if

$$\lim_{n \rightarrow \infty} d(L^n(G), L^n) = 0$$

where G is the grammar associated with the inferred D , L^n is the measurement language of observed n -cylinders, and $L^n(G)$ is the set of n -cylinders generated by G . This is the identification criterion of interest in the following.

There are a number of other contexts in which this type of identification is problematic. For example, with the further constraint, central to deductions concerning properties of the underlying noisy dynamical system, that the inferred grammar be minimal the resulting computational problem is NP-Hard.^{9,10} As will be discussed shortly, numerical results indicate that convergent identification often works, but proofs for the general case are not available. This is to be expected due to the very complex behavior arbitrary noisy dynamical systems are capable of producing. Nonconvergence can even be taken as a useful measure of the effective computational structure of the system, as noted later on. For the case of interest here there are, thus, several open problems concerning the identifiability of noisy dynamical systems. In this brief overview we can only hope to sketch the basic problem and encourage further investigation.

SYMBOLIC EQUATIONS OF MOTION

With this introduction to inductive inference for noisy dynamical systems, this section outlines the inverse problem of estimating symbolic equations of motion. This approach to modeling is closest to conventional statistical time series analysis. The basic statistical assumption for learning a dynamical system from data is that the observed time series is the result of a deterministic system in contact with a fluctuation source. The equations of motion for a continuous time process, for example, are given by a stochastic differential equation

$$\dot{\vec{v}} = \vec{F}(\vec{v}) + \vec{\xi}(t), \quad \vec{v}(0) \in M$$

where the second term is a random driving force.

Using a fine measurement partition P_ϵ ($\epsilon \ll 1$),¹¹ the goal of estimating equations of motion is to produce a symbolic representation of the dynamic \vec{F} and an estimate of the extrinsic noise level due to the fluctuating force. If the deterministic behavior is stably periodic then this problem essentially reduces to conventional linear prediction theory, as originated by Wiener and Kolmogorov. The case of general interest comes in not restricting \vec{F} to periodic behavior and considering the possibility that the deterministic dynamic itself produces complex behavior. The general question is then, given a noisy data stream $\vec{v}(t)$ or a function of it, how to infer that some portion of the noise is due to the fluctuating force and how much is due to the deterministic chaos.

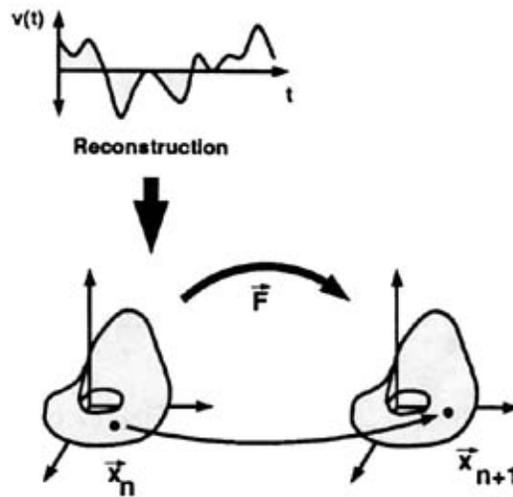


Figure 1 Estimating symbolic equations of motion: The first step is to reconstruct from a time series an effective state space in which to represent the behavior.¹² In this step the continuous time series is discretely sampled in time and in value. The dynamic is the mapping \bar{F} that takes the state space into itself; or equivalently, takes a single state at one time into its successor. This is the object of attention for the second step: a statistical fit then produces the coefficients of some expansion of \bar{F} using a chosen function basis.

The procedure for estimating the symbolic form of the dynamic applies Bayesian statistical inference to estimate nonlinear models from reconstructed chaotic data series.² (Figure 1 illustrates the overall procedure.) When approached from this point of view, the inverse problem reduces to statistical quadrature: estimating the symbolic equations of motion reduces to curve fitting in a space of dimension twice that of the reconstructed state space. An *ad hoc* choice of a function basis (e.g. polynomial or Fourier functions, or local splines) is made at this point. An unsupervised learning procedure then searches the space of consistent models using an optimality criterion that trades off forecasting error against model complexity. The residual fit error then yields an estimate of the extrinsic noise level. Once a model has been estimated it can be used in an interactive simulation interpreter to reproduce behavior in the same chaotic class as the data.

The choice of a finite function basis is a coordinatization of the hypothesis space H : the coordinates are fit parameters. The finite-dimensional reduction of H this affords determines the model space \mathcal{M} . The observed data is a realization of some “true” distribution $p_{D'}$ on H . A model D is a good estimate then if the differences between p_D , the distribution specified by D , and $p_{D'}$ is minimized. One natural measure of the minimization is the model entropy

$$I(D, D') = \int_H dm p_D(m) \log \frac{p_D(m)}{p_{D'}(m)}$$

While formally appropriate to the task at hand, there are two problems with this. First, the integral is over the infinite-dimensional function space H . Second, we do not know the “true” distribution $p_{D'}$. Both of these problems can be addressed by pulling back the reconstructed data distribution on H to the finite-dimensional model space \mathcal{M} . In this we identify errors in the fit parameter estimation with the reconstructed distribution in H of fluctuation-induced deviations from the “true” dynamic. (This is an approximation and is not to say these are necessarily the same thing.) Both this error and the dimension of the model space play a role in finding an optimal model. By trading off these two components of the model entropy, we would like to maximize the information in the presentation P that is captured or “explained” by the estimated model. In the model space, then, we

approximate p_D by the probability $p_{\mathcal{M}}(m|P)$ conditioned on the presentation. The model entropy is then

$$I(D, D') \approx \int_{\mathcal{M}} dm p_{\mathcal{M}}(m|P) \log p_{\mathcal{M}}(m|P)$$

Using Bayes's theorem, though,

$$p_{\mathcal{M}}(m|P) = \frac{p_{\mathcal{M}}(P|m) p_{\mathcal{M}}(m)}{p_{\mathcal{M}}(P)}$$

where $p_{\mathcal{M}}(P) = \int_{\mathcal{M}} dm p_{\mathcal{M}}(P|m) p_{\mathcal{M}}(m)$ is a normalizing constant that can be dropped without affecting the minimization. $p_{\mathcal{M}}(m)$ is the probability of a particular model $m \in \mathcal{M}$ and $p_{\mathcal{M}}(P|m)$ the probability that m has produced the data P . Putting this all together, gives the following two-term approximation

$$I(D, D') \propto \int_{\mathcal{M}} dm p_{\mathcal{M}}(P|m) \log p_{\mathcal{M}}(P|m) + \int_{\mathcal{M}} dm p_{\mathcal{M}}(m) \log p_{\mathcal{M}}(m)$$

The first term on the right-hand side is a measure of the amount of information in the observations not explained by the model. In the case of Gaussian error distribution, this can be again approximated by the fit error $-\log \sigma$, where σ is the one-step prediction error variance and the data range has been normalized. The second term is an informational measure of the complexity of the consistent models. This term is dominated by the number of model parameters k . With these further, somewhat extreme approximations, the Bayesian version of Akaike's model identification criterion is recovered

$$I(D, D') \propto -\log \sigma + \frac{k \log N}{N}$$

where N is the number of examples in the presentation.¹³

There are two broad classes of equations of motion inference methods: global and atlas. The latter uses coordinate charts on M over which splines are fit to the graph of the dynamic. The former uses a function basis that is applied over the entire state space manifold M . The class of dynamics that can be learned with global equations of motion is clearly a subset of those inferable using atlas equations of motion. The latter, unfortunately, indicates very little about any underlying functional simplicity of the dynamic. The identification of new *laws* requires the use of both methods and so a generalized success criterion favoring global equations of motion when they are the most compact representation.

There are two problems with estimating symbolic equations of motion. The first concerns the identification of generalized *states* in the data stream. Presumably the most parsimonious use of the given data and the simplest model require detecting patterns that contain the most information and that are optimal for forecasting. This is indirectly addressed in the above procedure during the state space reconstruction step; which was given short shrift. The second problem derives from the lack of an absolute measure of a model's complexity. Here, as in all statistical curve fitting, the size of a model is measured relative to an *ad hoc* function basis. There is no general method to compare model complexities across function bases. This is essential to the search for the smallest representation for reasons of efficiency and for inferring that a given structure in the estimated model is a property of the original physical process.

STATISTICAL MECHANICS OF ϵ -MACHINES

Both of these problems are addressed by the second approach to the inverse problem which seeks to quantify the intrinsic computation performed by a physical process.³ The goal is to reconstruct a minimal and unique automaton or grammar that recognizes the measurement language to within some approximation. The automata are referred to as ϵ -machines, with ϵ indicating a generic level of approximation. A reconstructed machine in principle could be based on any computation model. The Chomsky hierarchy provides a graded set useful for systematically distinguishing more from less powerful computational models. In practice an ϵ -machine is either a deterministic finite automaton (DFA) or, if a finite DFA is inconsistent, a minimal pushdown automaton (PDA). Reconstructing approximations of higher level machines from data is understandably fraught with difficulty. The computational approach reduces the inverse problem to the learning problem of stochastic grammatical inference.

Probabilistic structure in the measurement language is taken into account via a statistical mechanical formalism. One result of this is a connection between the structure of the inferred machine and traditional dynamical systems measures of the unpredictability. Additionally, this leads to a new invariant, the *complexity*, for dynamical systems based on the Rényi entropy that measures the amount of information contained in the inferred machine states. This quantity reflects the computational difficulty inherent in modeling nonlinear dynamical systems. Both repetitive and very unpredictable behavior have simple descriptions and so low complexity.^{3,5}

An ϵ -machine is described by a labeled, directed graph, or *l-digraph*, that consists of a set of vertices \mathbf{V} and a set of edges \mathbf{E} connecting them. Its statistical structure is given by a parametrized *probabilistic connection matrix*

$$T_\alpha = \{t_{ij}\} = \sum_{s \in \mathbf{A}} T_\alpha^{(s)}$$

that is the sum over each symbol of the state transition matrices

$$T_\alpha^{(s)} = \{t_{ij;s}^\alpha\}, \quad t_{ij;s}^\alpha = p(v_i | v_j; s)$$

for the vertices $\mathbf{V} = \{v_i\}$. The entries are the conditional probabilities of making a transition from state i to state j on symbol s .

For the topological case ($\alpha = 0$) the unique and minimal l-digraph and the associated connection matrix can be reconstructed from a data stream using a variant of standard grammatical inference applied to a prefix tree of unique measurement sequences.^{3,5,14} For arbitrary α , probabilistic structure of the data is translated into estimates of the transition probabilities. The resulting machine in this case is only an approximation at each window length. The inferred l-digraph states, called *morphs*, represent historical templates that are optimal for forecasting.

The α -order total Rényi *entropy*, or *free information*, of a reconstructed ϵ -machine is given by

$$H_\alpha(n) = (1 - \alpha)^{-1} \log Z_\alpha(n)$$

where the *partition function* is

$$Z_\alpha(n) = \sum_{s^n \in \{s^n\}} e^{\alpha \log p(s^n)}$$

with the probabilities $p(s^n)$ defined on the observed n -cylinders $\{s^n\}$. The parameter α has two interpretations, both of interest in the present context. From the physical point of view, it plays the role of an inverse temperature in statistical mechanics which emphasizes different invariant subsets in orbit space. From the point of view of Bayesian inference α is a Lagrange multiplier specifying a maximum entropy distribution consistent with the observed cylinder probabilities.

The Rényi *specific entropy*, i.e. entropy per unit time, is approximated for n -cylinders by

$$h_\alpha(n) = n^{-1} H_\alpha(n)$$

The *metric entropy*, a measure of predictability from dynamical systems theory, is then

$$h_\mu = \lim_{n \rightarrow \infty} h_1(n)$$

The α -order *graph complexity* is defined as

$$C_\alpha = (1 - \alpha)^{-1} \log \sum_{v \in \mathbf{V}} p_v^\alpha$$

where the probabilities p_v are defined on the machine states. It measures an ϵ -machine's information capacity in terms of the amount of information stored in the morphs. The entropies and complexities are dual in the sense that the former is determined by the maximum eigenvalue λ_α of T_α ,

$$h_\alpha = \log_2 \lambda_\alpha$$

and the latter by the associated (left) eigenvector $\vec{p}_\alpha = \{p_v^\alpha : v \in \mathbf{V}\}$ that gives the asymptotic vertex probabilities. The specific entropy is also given directly in terms of the transition probabilities

$$h_\alpha = \sum_{v \in \mathbf{V}} \frac{p_v^\alpha}{1 - \alpha} \log \sum_{\substack{v' \in \mathbf{V} \\ s \in \mathbf{A}}} t_{vv';s}^\alpha$$

A complexity based on the asymptotic edge probabilities $\vec{p}_e = \{p_e : e \in \mathbf{E}\}$ can also be defined

$$C_\alpha^e = (1 - \alpha)^{-1} \log \sum_{e \in \mathbf{E}} p_e^\alpha$$

Not much is gained, however, since

$$C_1^e = C_1 + h_1$$

and so there are only two independent quantities for a finite ϵ -machine.

The graph complexity is also a measure of the informational fluctuations in the data. These fluctuations are most readily quantified by the total excess (α -)entropy for L -cylinders^{7,15,16}

$$F_\alpha(L) = H_\alpha(L) - Lh_\alpha = \sum_{n=1}^L [H_\alpha(n) - H_\alpha(n-1) - h_\alpha]$$

This measures the deviation of finite cylinder statistics from asymptotic. It can also be interpreted as the mutual information between the future L -cylinders and the infinite past. In the thermodynamic limit

$$C_\alpha \underset{L \rightarrow \infty}{\propto} F_\alpha(L)$$

Loosely speaking, the graph complexity measures the amount of mathematical work necessary to produce a deviation from uniform statistics. From this limit, it is seen to be the mutual information between the infinite past and the infinite future.

A measure of convergent identification in the limit can now be defined. Since the complexity is a measure of the size of the current hypothesized (DFA) model at the given level of approximation, it can be used as a diagnostic of convergence. Holding the measurement partition P_ϵ fixed, that is assuming it produces adequately representative symbols,[‡] identification with longer L -cylinders is the primary concern. The identification method then converges with increasing cylinder length if the rate of change of the complexity vanishes. That is, if

$$c_\alpha = \lim_{L \rightarrow \infty} \frac{2^{C_\alpha(L)}}{L}$$

vanishes then the noisy dynamical system has been identified. If this is not the case, then c_α , a measure of the rate of divergence of the model size, is yet another measure of complexity, but at a higher level of computation.

ϵ -machines have been reconstructed from hundreds of data sets from prototype chaotic systems. The implementation is relatively fast and has been used to infer machines with several hundred states. This constructive approach to complexity has proved itself useful in elucidating the phase transition structure at the onset of chaos in systems with a control parameter.³ Although developed in the context of reconstruction from a data series, the underlying theory provides an analytic approach to calculating entropies and complexities for a number of dynamical systems.⁴ One noteworthy result is a universal description of period-doubling cascade transitions to chaos that depends only on the intrinsic computation and information processing capacity of the dynamical systems in the bifurcation sequence, and is independent of explicit nonlinearity controls.

Another application, that relies heavily on the inductive inference formalism, is a method for inferring the direction of time in a data stream and for measuring quantitatively the degree of irreversibility.¹⁷ With this, computational learning theory sheds new light on the long-standing paradox of irreversible macroscopic processes that are described microscopically by reversible dynamics.

Finally, we note that ϵ -machine reconstruction provides a data compression technique that gives an efficient method for encoding a chaotic data stream, transmitting the compressed form, and uniquely decoding it.¹⁸ Technically, this is referred to as data compaction; data compression allows for some error in encoding so that exact reconstruction of the original sequence is not always possible. (The chaotic data compaction method can be modified to allow for compression at a specified fidelity.) The compaction ratio r , defined as the ratio of the number output bits from the encoder to the number input, is given by

$$r = C_\alpha^e - C_\alpha$$

where log base the number of symbols is used to give a normalized ratio. This form of the data compaction ratio makes clear the dependence on branching structure in the reconstructed machine. To the extent that there is branching in the machine state transitions, bits must be passed to the output from the input. To the extent there is determinism in the transitions, the machine captures it and there is no need to output symbols on those transitions.

[‡] That is, assuming it is a *generating* partition.⁷

The goal of estimating optimal models, as reflected in minimizing (say) the model entropy for selecting symbolic equations of motion, is most generally expressed by Rissanen's minimum description length (MDL) criterion.¹⁹ This is easily explained for ϵ -machine reconstruction by recalling the preceding comments on chaotic data compaction. Briefly, the MDL criterion says to choose that computational model m that reduces the size of the model m and the length $l(P|m)$ of the data P encoded with that model. That is, the description length

$$l(P) = C_0(m) + l(P|m)$$

should be minimized. Since the reconstruction method produces the minimal and unique ϵ -machine for a complete presentation, a search procedure and so the MDL criterion are unnecessary.

CONCLUDING REMARKS

The last decade has seen tremendous progress in modeling complex natural phenomena. This has largely been accomplished through the solution of isolated problems within dynamical systems theory and its applications. These include

1. Extensive studies of nonlinear phenomenology: chaotic attractors, complex basin structures, phase transition description of the onset of chaos;
2. Data analysis techniques: reconstruction of state space from single time series, optimal coordinates, estimating equations of motion;
3. Statistical characterizations of noisy behavior: estimating the embedding and attractor dimensions as well as information production rates, a thermodynamic description of invariant measures on attractors and their orbits; and
4. Estimation of the intrinsic computational information processing within physical processes governed by noisy dynamical systems.

While some of this is unified within abstract dynamical systems theory, the practical application has been plagued by inconsistencies and unsystematic development that has produced a patchwork of theories garnered from engineering, statistics, and physics.

It is not unreasonable to explain the distance between theory and practice and the latter's difficulties as stemming from a lack of formalizing the overall goal of the enterprise: detecting and modeling deterministic structure in noisy data. This is exactly where learning theory can contribute since one of its mandates is to formalize learning paradigms. The preceding discussion has attempted to outline, however schematically, the learning paradigms appropriate to chaotic data analysis. A number of existing problems can be more clearly articulated in this context. There is much to be gained by the application of learning theory to dynamical systems.

This is all well and good for dynamical systems and the sciences it serves, but what about learning theory itself? It seems likely that it too will benefit by access to and the appreciation of the well-defined and wide-ranging class of learning problems provided by dynamical systems. One example to look forward to is a quantitative investigation of learning a parametrized class of dynamical systems that go from periodic to chaotic behavior in which the basic computational problem goes from P to NP. This is already implicit in the tension between the typical polynomial speed of machine reconstruction and the complexity theoretic results that minimal consistent DFA inference is NP-complete. The observation is that it is only noisy dynamical systems at complexity phase transitions which are hard to learn.

More generally, the type of learning problems found in dynamical systems, only some of which have been presented above, are of a different character than the symbolic AI problem traditional in learning theory. They are more akin to problems in computational geometry and in neural networks with the additional feature that dynamical systems generate a probability measure on the orbits which singles out atypical (measure zero) behavior. It is particularly important to the application of learning theory that the practical problems of approximation be taken in account from the start. In contrast, complexity theoretic approaches tend to emphasize worst case behavior in order to obtain hardness results. If these apply only to measure zero orbits or to nongeneric dynamical systems, then they might be of little practical value.

If it has not become sufficiently clear at this point let me mention before closing the underlying motivation for a computational learning theory of dynamical systems. The goal is *artificial science*. Within the limited domain of nonlinear dynamical systems subject to extrinsic fluctuations, this is the complete automation of inferring deterministic dynamics from noisy data. To the extent that this is possible, implementations will greatly accelerate experimental science and presumably scientific discovery generally. To the extent that it is not feasible, we will have an understanding of yet another limitation of scientific method.

The foregoing has briefly reviewed several approaches to consistently modeling noisy dynamical systems. It has turned on the intimate connection between complex dynamics, modeling theory, statistics, and computation. The role of modeling was emphasized, although there is an equally important scientific, rather than engineering, motivation. That is the definition of a measure of complexity that is appropriate to and *implementable* for physical processes. The framework outlined above for reconstructing ϵ -machines and for their statistical mechanical description unifies a number of proposed definitions of physical complexity, that range from the use of entropy convergence critical exponents to the relaxation rates for diffusion on hierarchical barriers.^{20,7,21,22,23,24,16,25} Not only does the current framework indicate how these are related, but also how they fail to capture important structural properties of complex systems. This structure is given explicitly by the ϵ -machines.

The exclusive use of information theoretic analysis[§] restricts one to the lowest level in Chomsky's hierarchy. Indeed, conventional statistical mechanics is similarly limited due to its reliance on correlation functions, that is, on $\alpha = 2$ information statistics. Such a restriction misses higher level computation that dominates at phase transitions. The general conclusion is that only systems at phase transitions, to be understood broadly as being in a "critical" state, perform computation at levels beyond information transmission and storage. Being in a critical state is a first and a minimal requirement for nontrivial computation. This is as true of ferromagnets at the Curie temperature as it is of digital or analog computers considered as physical systems and of evolving biological organisms. These systems trade their ability to store information in certain quasi-static degrees of freedom against the underlying nonlinear dynamics necessary for innovation and reliable information transmission. These needs are balanced at the borders of chaos. Computationally critical states form the substrate supporting nontrivial information processing and, presumably, are prerequisite for evolutionary development. The question remains: Why would a bowl of primordial soup spontaneously take up a computationally critical state?

[§] I have in mind especially the recent resurgence of mutual information.

References

1. J. P. Crutchfield, N. H. Packard, J. D. Farmer, and R. S. Shaw. Chaos. *Sci. Am.*, 255:46, 1986.
2. J. P. Crutchfield and B. S. McNamara. Equations of motion from a data series. *Complex Systems*, 1:417, 1987.
3. J. P. Crutchfield and K. Young. Inferring statistical complexity. *Phys. Rev. Let.*, 63:105, 1989.
4. J. P. Crutchfield and K. Young. Computation at the onset of chaos. In W. Zurek, editor, *Entropy, Complexity, and the Physics of Information*. Addison-Wesley, 1989. in press.
5. J. P. Crutchfield and K. Young. Thermodynamics of minimal reconstructed machines. in preparation, 1989.
6. E. M. Gold. Language identification in the limit. *Info. Control*, 10:447, 1967.
7. J. P. Crutchfield and N. H. Packard. Symbolic dynamics of noisy chaos. *Physica*, 7D:201, 1983.
8. R. M. Wharton. Approximate language identification. *Info. Control*, 26:236, 1974.
9. E. M. Gold. Complexity of automaton identification from given data. *Info. Control*, 37:302, 1978.
10. D. Angluin. On the complexity of minimum inference of regular sets. *Info. Control*, 39:337, 1978.
11. J. P. Crutchfield. *Noisy Chaos*. PhD thesis, University of California, Santa Cruz, 1983. published by University Microfilms Intl, Minnesota.
12. N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw. Geometry from a time series. *Phys. Rev. Let.*, 45:712, 1980.
13. J. Rissanen. Modeling by shortest data description. *Automatica*, 14:462, 1978.
14. A. W. Biermann and J. A. Feldman. On the synthesis of finite-state machines from samples of their behavior. *IEEE Trans. Comp.*, C-21:592, 1972.
15. N. H. Packard. *Measurements of Chaos in the Presence of Noise*. PhD thesis, University of California, Santa Cruz, 1982.
16. P. Grassberger. Toward a quantitative theory of self-generated complexity. *Intl. J. Theo. Phys.*, 25:907, 1986.
17. J. P. Crutchfield. Time is the ultrametric of causality. in preparation, 1989.
18. J. P. Crutchfield. Compressing chaos. in preparation, 1989.
19. J. Rissanen. Universal coding, information, prediction, and estimation. *IEEE Trans. Info. Th.*, IT-30:629, 1984.
20. J. P. Crutchfield and N. H. Packard. Symbolic dynamics of one-dimensional maps: Entropies, finite precision, and noise. *Intl. J. Theo. Phys.*, 21:433, 1982.
21. S. Wolfram. Computation theory of cellular automata. *Comm. Math. Phys.*, 96:15, 1984.
22. R. Shaw. *The Dripping Faucet as a Model Chaotic System*. Aerial Press, Santa Cruz, California, 1984.
23. C. H. Bennett. On the nature and origin of complexity in discrete, homogeneous locally-interacting systems. *Found. Phys.*, 16:585, 1986.
24. C. P. Bachas and B.A. Huberman. Complexity and relaxation of hierarchical structures. *Phys. Rev. Let.*, 57:1965, 1986.
25. S. Lloyd and H. Pagels. Complexity as thermodynamic depth. *Ann. Phys.*, 188:186, 1988.