Semantics and Thermodynamics

James P. Crutchfield

SFI WORKING PAPER: 1991-09-033

SANTA FE INSTITUTE

Semantics and Thermodynamics

James P. Crutchfield

Physics Department*
University of California
Berkeley, California 94720 USA

Abstract

Inferring models from given data leads through many different changes in representation. Most are subtle and profitably ignored. Nonetheless, any such change affects the semantic content of the resulting model and so, ultimately, its utility. A model's semantic structure determines what its elements mean to an observer that has built and uses it. In the search for an understanding of how large-scale thermodynamic systems might themselves take up the task of modeling and so evolve semantics from syntax, the present paper lays out a constructive approach to modeling nonlinear processes based on computation theory. It progresses from the microscopic level of the instrument and individual measurements, to a mesoscopic scale at which models are built, and concludes with a macroscopic view of their thermodynamic properties. Once the computational structure of the model is brought into the analysis it becomes clear how a thermodynamic system can support semantic information processing.

* Internet: chaos@gojira.berkeley.edu.

NONLINEAR MODELING: FACT OR FICTION?

> These ambiguities, redundances, and deficiencies recall those attributed by Dr. Franz Kuhn to a certain Chinese encyclopedia entitled *Celestial Emporium of Benevolent Knowledge*. On those remote pages it is written that animals are divided into (a) those that belong to the Emperor, (b) embalmed ones, (c) those that are trained, (d) suckling pigs, (e) mermaids, (f) fabulous ones, (g) stray dogs, (h) those that are included in this classification, (i) those that tremble as if they were mad, (j) innumerable ones, (k) those drawn with a very fine camel's brush hair, (l) others, (m) those that have just broken a flower vase, (n) those that resemble flies from a distance.
>
> J. L Borges, "The Analytical Language of John Wilkins", page 103.[5]

What one intends to do with a model colors the nature of the structure captured by it and determines the effort used to build it. Unfortunately, such intentions most often are not directly stated, but rather are implicit in the choice of representation. To model a given time series, should one use (Fourier) power spectra, Laplace transforms, hidden Markov models, or neural networks with radial basis functions?

Two problems arise. The first is that the choice made might lead to models that miss structure. One solution is to take a representation that is complete: a sufficiently large model captures the data's properties to within an error that vanishes with increased model size. The second, and perhaps more pernicious, problem is that the limitations imposed by such choices are not understood *vis á vis* the underlying mechanisms. This concerns the appropriateness of the representation.

The basis of Fourier functions is complete. But the Fourier model of a square wave contains an infinite number of parameters and so is of infinite size. This is not an appropriate representation, since the data is simply described by a two state automaton.† Although completeness is a necessary property, it simply does not address appropriateness and should not be conflated with it.

Nonlinear modeling, which I take to be that endeavor distinguished by a geometric analysis of processes represented in a state space, offers the hope of describing more concisely and appropriately a range of phenomena hitherto considered random. It can do this since it enlarges the range of representations and forces an appreciation, at the first stages of modeling, of nonlinearity's effect on behavior. Due to this nonlinear modeling necessarily will be effective.

From the viewpoint of appropriateness, however, nonlinear modeling is an ill-defined science: discovered nonlinearity being the product largely of assumptions made by and resources available to the implementor; and not necessarily a property of the process modeled. There is, then, a question of scientific principle that transcends its likely operational success: How does nonlinearity allow a process to perform different classes of computation and so exhibit more or less complex behavior? This is where I think nonlinear modeling can make a contribution

† It is an appropriate representation, though, of the response of free space to carrying weak electromagnetic pulses. Electromagnetic theory is different from the context of modeling *only* from given data. It defines a different semantics.

beyond engineering concerns. The contention is that incorporating computation theory will go some distance to basing modeling on first principles.

COMPUTATIONAL MECHANICS

The following discussion reviews an approach to these questions that seeks to discover and to quantify the intrinsic computation in a process. The rules of the inference game demand ignorance of the governing equations of motion. Each model is to be reconstructed from the given data. It follows in the spirit of the research program for chaotic dynamics introduced under the rubric of "geometry from a times series",[34] though it relies on many of the ideas and techniques of computation and learning theories.[2,21]

I first set up the problem of modeling nonlinear processes in the general context in which I continually find it convenient to consider this task.[10] This includes delineating the effect the measurement apparatus has on the quality and quantity of data. An appreciation of the manner in which data is used to build a model requires understanding the larger context of modeling; namely, given a fixed amount of data what is the best explanation? Once an acceptable model is in hand, there are a number of properties that one can derive. It becomes possible to estimate the entropy and complexity of the underlying process and, most importantly, to infer the nature of its intrinsic computation. Just as statistical mechanics explains macroscopic phenomena as the aggregation of microscopic states, the overall procedure of modeling can be viewed as going from a collection of microscopic measurements to the discovery of macroscopic observables; as noted by Jaynes.[23] The resulting model summarizes the relation between these observables. Not surprisingly its properties can be given a thermodynamic interpretation that captures the combinatorial constraints on the explosive diversity of microscopic reality. This, to my mind, is the power of thermodynamics as revealed by Gibbsian statistical mechanics.

The following sections are organized to address these issues in just this order. But before embarking on this, a few more words are necessary concerning the biases brought to the development.

The present framework is "discrete unto discrete." That is, I assume the modeler starts with a time series of quantized data and must stay within the limits of quantized representations. The benefit of adhering to this framework is that one can appeal to computation theory and to the Chomsky hierarchy, in particular, as giving a complete spectrum of model classes.[21] By complete here I refer to a procedure that, starting from the simplest, finite-memory models and moving toward the universal Turing machine, will stop with a finite representation at the least powerful computational model class. In a few words that states the overall inference methodology.[11] It addresses, in principle, the ambiguity alluded to above of selecting the wrong modeling class. There will be somewhere in the Chomsky hierarchy an optimal representation which is finitely expressed in the language of the least powerful class.

Finally, note that this framework does not preclude an observer from employing finite precision approximations of real-valued probabilities. I have in mind here using arithmetic codes to represent or transmit approximate real numbers.[4] That is, real numbers are algorithms. It is a mistake, however, to confuse these with the real numbers that are a consequence of the inference methodology, such as the need at some point in time to solve a Bayesian or a maximum

entropy estimation problem; as will be done in a later section. This is a fact since an observer constrained to build models and make predictions within a finite time, or with infinite time but access to finite resources, cannot make use of such infinitely precise information. The symbolic problems posed by an inference methodology serve rather to guide the learning process and, occasionally, give insight when finite manipulations of finite symbolic representations lead to finite symbolic answers.

Fuzzy β-Instruments

The universe of discourse for nonlinear modeling consists of a process P, the measuring apparatus $\mathcal{I}$, and the modeler itself. Their relationships and components are shown schematically in Figure 1. The goal is for the modeler, taking advantage of its available resources, to make the "best" representation of the nonlinear process. In this section we concentrate on the measuring apparatus. The modeler is the subject of a later section. The process, the object of the modeler's ultimate attention, is the unknown, but hopefully knowable, variable in this picture. And so there is little to say, except that it can be viewed as governed by stochastic evolution equations

$$\vec{X}_{t+\Delta t} = \vec{F}\left(\vec{X}_t, \vec{\xi}_t, t\right) \tag{1}$$

where $\vec{X}_t$ is the configuration at time t, $\vec{\xi}_t$ some noise process, and $\vec{F}$ the governing equations of motion.* The following discussion also will have occasion to refer to the process's measure $\mu\left(\vec{X}\right)$ on its configuration space and the entropy rate $h_\mu\left(\vec{X}\right)$ at which it produces information.

The measuring apparatus is a transducer that maps $\vec{X}_t$ to some accessible states of an instrument $\mathcal{I}$. This instrument has a number of characteristics, most of which should be under the modeler's control. The primary interaction between the instrument and the process is through the measurement space $\mathcal{R}^{\mathrm{D}}$ which is a projection $\mathcal{P}$ of $\vec{X}_t$ onto (say) a Euclidean† space whose dimension is given by the number D of experimental probes. The instrument's resolution ϵ in distinguishing the projected states partitions the measurement space into a set $\Pi_\epsilon(\mathrm{D}) = \left\{\pi_i : \pi_i \subset \mathcal{R}^{\mathrm{D}}, i = 0, \ldots, \epsilon^{-\mathrm{D}}\right\}$ of cells. Each cell π_i is the equivalence class of projected states that are indistinguishable using that instrument. The instrument represents the event of finding $\mathcal{P}\left(\vec{X}_t\right) \in \pi_i$ by the cell's label i. With neither loss of generality nor information, these indices are then encoded into a time-serial binary code. As each measurement is made its code is output into the data stream. In this way, a time series of measurements made by the instrument becomes a binary string, the data stream s available to the modeler. This is a discretized set of symbols $\mathrm{s} = \ldots s_{-4}s_{-3}s_{-2}s_{-1}s_0s_1s_2s_3s_4\ldots$ where in a single measurement made by the modeler the instrument returns a symbol $s_t \in \mathrm{A}$ in an alphabet A at time index $t \in \mathrm{Z}$. Here we take a binary alphabet $\mathrm{A} = \{0, 1\}$.

This gives the overall idea, but it is in fact a gross simplification. I will discuss two important elements that are left out: the instrument temperature and the cell dwell time.

* I explicitly leave out specifying the (embedding) dimension of the process. This is a secondary statistic that is estimated[12] as a topological property of the model, not something intrinsic to the present view of the process.

† If measuring p phases, for example, then the associated topology would be $\mathcal{R}^{\mathrm{D}-\mathrm{p}} \times \mathcal{T}^{\mathrm{p}}$.

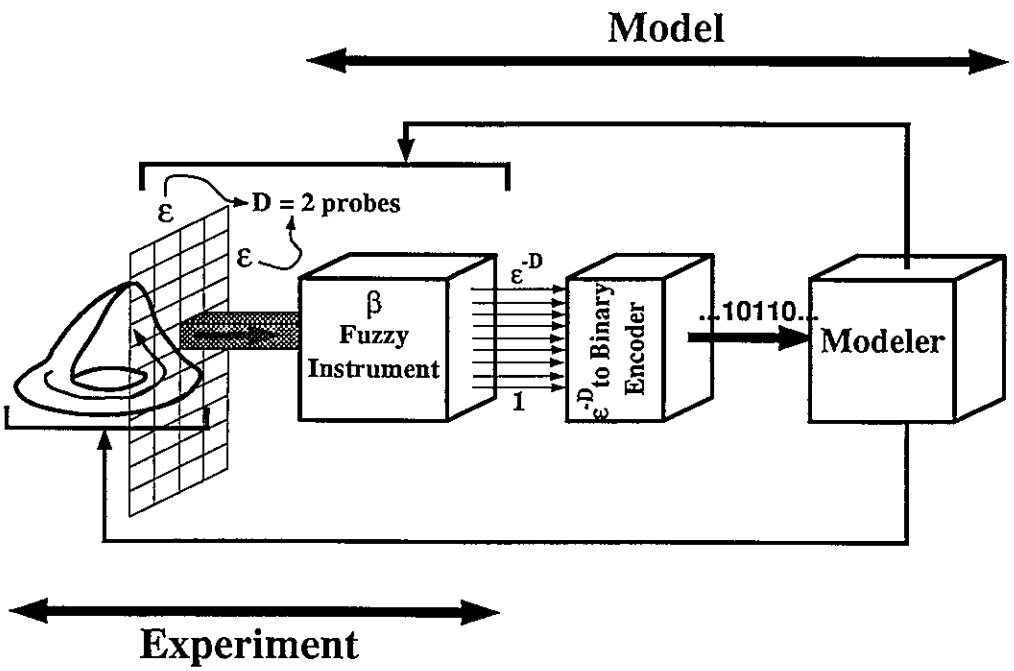

Figure 1 The Big Channel. The flow of information (measurements) on the shortest time scales is from the left, from the underlying process, to the right toward the modeler. The latter's task is to build the "best" representation given the available data set and computational resources. On longer time scales the modeler may modify the measuring apparatus and vary the experimental controls on the process. These actions are represented by the left-going arrows. Notice that from the modeler's perspective there is a region of ambiguity between the model and the experiment. The model includes the measuring apparatus since it instantiates many of the modeler's biases toward what is worth observing. But the experiment also includes the measuring apparatus since it couples to the process. Additionally, the apparatus is itself a physical device with its own internal dynamics of which the modeler may be unaware or incapable of controlling.

As described, the measurement partition $\Pi_\epsilon(\mathrm{D})$ is "crisp". Each partition cell is associated with an indicator function that maps the state $\vec{x} = \mathcal{P}\left(\vec{X}\right) \in \mathcal{R}^{\mathrm{D}}$ onto a symbolic label for that element depending on whether the state is or is not in the domain of that indicator function. But no real instrument implements a crisp measurement partition. There are errors in the assignment of a state $\vec{x}$ to a cell and so an error in the resulting symbol. There are two kinds of errors that one might consider.

The first is a classification error in which the cell is misidentified with the projected state $\vec{x} = \mathcal{P}\left(\vec{X}_t\right)$ independent of its location within the measurement cell. If the error rate probability is taken to be p, then the instrument's effective temperature $T_{\mathrm{inst}} = \left(k_{\mathrm{Boltzmann}}\beta_{\mathrm{inst}}\right)^{-1}$ is simply $\beta_{\mathrm{inst}} = \log_2(1-p)/p$. This is not a very realistic view of classification error. Physical devices, such as analog-to-digital converters, fail more in correct classification near the cell boundaries since they cannot implement exact decision thresholds. In this case error is not uniform over the partition cells.

One solution to this follows the spirit of fuzzy logic which suggests that the cell indicator function be generalized to a membership function that decays outside a cell.[43] An example of a fuzzy instrument that accounts for this somewhat realistically is to convolve the boundaries of the cells in the crisp partition $\Pi_\epsilon(\mathrm{D})$ with a Fermi-Dirac density. The membership function

then becomes

$$\pi_\epsilon^{\beta_{\text{inst}}}\left(\vec{X}\right) \propto \frac{1}{e^{\beta_{\text{inst}}(\|\mathcal{P}(\vec{X})-\vec{x}_\pi\|-\epsilon/2)}+1} \tag{2}$$

where β_{inst} is the fuzzy partition's inverse temperature and $\vec{x}_\pi \in \mathcal{R}^{\text{D}}$ is the cell's center in the measurement space. At zero temperature the crisp partition is recovered, $\Pi_\epsilon^\beta(\text{D}) \underset{\beta\to\infty}{\to} \Pi_\epsilon(\text{D})$ At sufficiently high temperatures, the instrument outputs random sequences uncorrelated with the process within the cell.

The algebra of fuzzy measurements will not be carried through the following. I will simply leave behind at this point knowledge of the fuzzy partition. The particular consequences for doing this correctly, though, will be reported elsewhere. The main result is that when done in this generality, the ensuing inference process is precluded from inferring too much and too precise a structure in the source.

The second element excluded from the Big Channel concerns the time $\vec{X}_t$ spends in each partition cell. To account for this there should be an additional time series that gives the cell dwell time for each state measurement. Only in special circumstances will the dwell time be constant, if the partition is a uniform coarse-graining. When ergodicity can be appealed to the average dwell time τ can be used. In any case, it is an important parameter and one that is readily available, but often unused.

The dwell time suggests another instrument parameter, the frequency response; or, more properly dropping Fourier modeling bias, the instrument's dynamic response. On short time scales the instrument's preceding internal states can affect its resolution in determining the present state and the dwell time. In the simplest case, there is a shortest time below which the instrument cannot respond. Then passages through a cell that are too brief will not be detected or will be misreported.

All of these detailed instrumental properties can be usefully summarized by the information acquisition rate $\dot{I}$. In its most general form it is given by the information gain of the fuzzy partition Π_ϵ^β with respect to the process's asymptotic distribution $\mu\left(\vec{X}\right)$ projected onto the measurement space. That is,

$$\dot{I}(\tau,\beta,\epsilon,\text{D}) = \tau^{-1} H\left(\Pi_\epsilon^\beta(\text{D})\,|\mathcal{P}\left(\mu\left(\vec{X}\right)\right)\right) \tag{3}$$

where $H(P|Q)$ is the information gain of distribution P with respect to Q. Assuming ignorance of the process's distribution allows some simplification and gives the measurement channel capacity

$$\begin{aligned}\dot{I}(\tau,\beta,\epsilon,\text{D}) &= \tau^{-1} H\left(\Pi_\epsilon^\beta(\text{D})\right) \\ &= \dot{I}(\tau,\infty,\epsilon,\text{D}) - \tau^{-1} H\left(\pi_\epsilon^\beta(\text{D})\right) \\ \text{where } \dot{I}(\tau,\infty,\epsilon,\text{D}) &= \tau^{-1}\log_2\|\Pi_\epsilon(\text{D})\| = -\tau^{-1}\text{D}\log_2\epsilon \end{aligned} \tag{4}$$

and where $H\left(\pi_\epsilon^\beta(\text{D})\right)$ is the entropy of a cell's membership function and $\|\Pi_\epsilon(\text{D})\|$ is the number of cells in the crisp partition. At high temperature $H\left(\pi_\epsilon^\beta(\text{D})\right) \underset{\beta\to 0}{\to} \log_2\left\|\Pi_\epsilon^\beta(\text{D})\right\|$ and the information acquisition rate vanishes, since each cell's membership function widens to cover the measurement space.

The Modeler

Beyond the instrument, one must consider what can and should be done with information in the data stream. Acquisition of, processing, and inferring from the measurement sequence are the functions of the modeler. The modeler is essentially defined in terms of its available inference resources. These are dominated by storage capacity and computational power, but certainly include the inference method's efficacy, for example. Delineating these resources constitutes the barest outline of an observer that builds models. Although the following discussion does not require further development at this abstract a level, it is useful to keep in mind since particular choices for these elements will be presented.

The modeler is presented with s, the bit string, some properties of which were just given. The modeler's concern is to go from it to a useful representation. To do this the modeler needs a notion of the process's effective state and its effective equations of motion. Having built a model representing these two components, any residual error or deviation from the behavior described by the model can be used to estimate the effective noise level of the process. It should be clear when said this way that the noise level and the sophistication of the model depend directly on the data and on the modeler's resources. Finally, the modeler may have access to experimental control parameters. And these can be used to aid in obtaining different data streams useful in improving the model by (say) concentrating on behavior where the effective noise level is highest.

The central problem of nonlinear modeling now can be stated. Given an instrument, some number of measurements, and fixed *finite* inference resources, how much computational structure in the underlying process can be extracted?

Limits to Modeling

Before pursuing this goal directly it will be helpful to point out several limitations imposed by the data or the moder's interpretation of it.

In describing the data stream's character it was emphasized that the individual measurements are only indirect representations of the process's state. If the modeler interprets the measurements as the process's state, then it is unwittingly forced into a class of computationally less powerful representation.[5] These consist of finite Markov chains with states in A or in some arbitrarily selected state alphabet.* This will become clearer through several examples used later on. It is important at this early stage to not over-interpret the measurements' content as this might limit the quality of the resulting models.

The instrument itself obviously constrains the observer's ability to extract regularity from the data stream and so it directly affects the model's utility. The most basic of these constraints are given by Shannon's coding theorems.[40] The instrument was described as a transducer, but it also can be considered to be a communication channel between the process and the modeler. The capacity of this channel is $\dot{I} = \tau^{-1} H\left(\Pi_\epsilon^\beta(\mathrm{D})\right)$. As $\beta \to \infty$ and if the process is deterministic and has entropy $h_\mu\left(\vec{X}\right) > 0$, a theorem of Kolmogorov's says that this rate is maximized for a given process if the crisp partition $\Pi_\epsilon(\mathrm{D})$ is generating.[26] This property requires infinite sequences of cell indices to be in a finite-to-one correspondence with the process's states. A

* As done with hidden Markov models.[18,37]

similar result was shown to hold for the classes of process of interest here: deterministic, but coupled to an extrinsic noise source.[13] Note that the generating partition requirement necessarily determines the number D of probes required by the instrument.

For an instrument with a crisp generating partition, Shannon's noiseless coding theorem says that the measurement channel must have a capacity higher than process's entropy

$$\dot{I} \geq h_\mu \left(\vec{X} \right) \tag{5}$$

If this is the case then the modeler can use the data stream to reconstruct a model of the process and, for example, estimate its entropy and complexity. These can be obtained to within error levels determined by the process's extrinsic noise level.

If $\dot{I} < h_\mu \left(\vec{X} \right)$, then Shannon's theorem for a channel with noise says that the modeler will not be able to reconstruct a model with an effective noise level less than the equivocation $h_\mu \left(\vec{X} \right) - \dot{I}$ induced by the instrument. That is, there will be an "unreconstructable" portion of the dynamics represented in the signal.

These results assume, as is also done implicitly in Shannon's existence proofs for codes, that the modeler has access to arbitrary inference resources. When these are limited there will be yet another corresponding loss in the quality of the model and an increase in the apparent noise level. It is interesting to note that if one were to adopt Laplace's philosophical stance that all (classical) reality is deterministic and update it with the modern view that it is chaotic, then the instrumental limitations discussed here are the general case. And apparent randomness is a consequence of them.

The Explanatory Channel

A clear statement of the observer's goal is needed, beyond just estimating the best model. Surely a simple model is to be desired from the viewpoint of understandability of the process's mechanism and as far as implementation of the model in (say) a control system is concerned. Too simple a model, though, might miss important structure, rendering the process apparently stochastic and highly unpredictable when it is deterministic, but nonlinear. The trade-off between model simplicity and large unpredictability can be explained in terms of a larger goal for the modeler: to explain to another observer the process's behavior in the most concise manner, but in detail as well. Discussion of this interplay will be couched in terms of the explanatory channel of Figure 2.

Before describing this view, it is best to start from some simple principles. To make contact with existing approaches and for the brevity's sake, the best model will be taken to be the most likely. If one had access to a complete probabilistic description of the modeling universe, then the goal would be to maximize the conditional probability $Pr\left(\mathrm{M}|\mathrm{s}\right)$ of the model M given the data stream s. This mythical complete probabilistic description $Pr\left(\mathrm{M},\mathrm{s}\right)$ is not available, but an approximation can be developed by factoring it using Bayes' rule[22]

$$Pr\left(\mathrm{M}|\mathrm{s}\right) = \frac{Pr\left(\mathrm{s}|\mathrm{M}\right) P\left(\mathrm{M}\right)}{\sum\limits_{\mathrm{M}\in\mathcal{M}} P\left(\mathrm{s}|\mathrm{M}\right)} \tag{6}$$

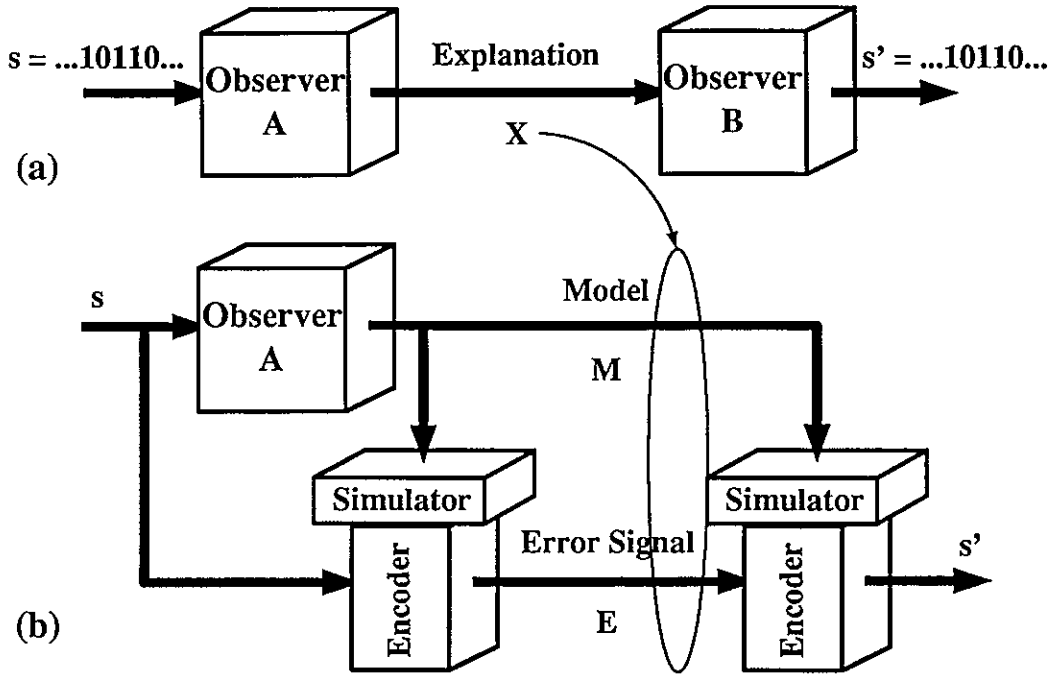

Figure 2 The Explanatory Channel. The upper portion (a) of the figure illustrates the bare channel: observer A communicates an explanation X to observer B. The lower portion (b) shows in more detail the two subchannels making up the explanatory channel: the model channel transmits model M and the error channel transmits an error signal E. M is first built by observer A and then loaded into A's simulator. It is also transmitted to B which loads into its simulator. A develops the error signal E as the deviation of the measurements in the data stream s from those predicted by simulating the model. Only these deviations are transmitted, and at less precision than the original individual measurements. B is to resynthesize a data stream s' by simulating the model and when that is not predictive, to use information from the error signal. X then *explains* s if s' = s.

There are several comments. First and foremost, all of these probabilities are conditioned on the choice of model class $\mathcal{M}$. Second, all of the terms on the right hand side refer to a single data stream s. Third, $Pr(\mathrm{s}|\mathrm{M})$ is the probability that a model $\mathrm{M} \in \mathcal{M}$ produces the given data. Thus, candidate models are considered to be generators of data. With sufficient effort, then, $Pr(\mathrm{s}|\mathrm{M})$ can be estimated. Finally, the normalization $Pr(\mathrm{s}) = \sum_{\mathrm{M}\in\mathcal{M}} Pr(\mathrm{s}|\mathrm{M})$ depends only on the given data and so can be dropped since it is a constant when maximizing $Pr(\mathrm{M}|\mathrm{s})$ over the model class.

Shannon's coding theorem established that an event of probability p can be optimally represented by a code with length $-\log_2 p$ bits.[40] The search for the most likely explanation is tantamount to constructing the shortest code $\hat{\mathrm{X}}$ for the data s. The length of the optimal code is then $\left\|\hat{\mathrm{X}}\right\| = -\log_2 Pr(\mathrm{M}|\mathrm{s})$. Using the above Bayesian decomposition of the likelihood, it follows that

$$\left\|\hat{\mathrm{X}}\right\| \propto -\log_2 Pr(\mathrm{s}|\mathrm{M}) - \log_2 Pr(\mathrm{M}) \tag{7}$$

The resulting optimization procedure can be described in terms of the explanatory channel of Figure 2. There are two observers A and B that communicate an explanation X via a channel.

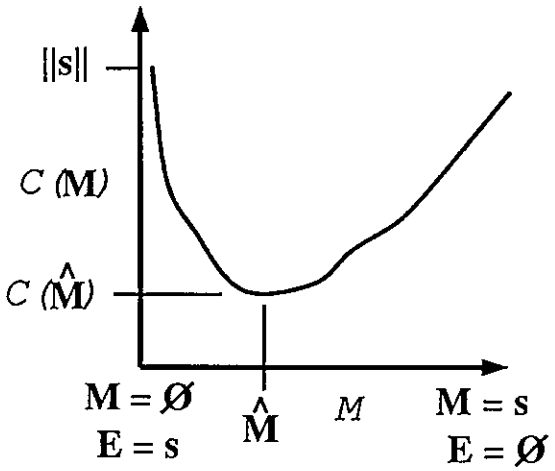

Figure 3 Model optimality. A schematic illustration of the features of the cost function $\mathcal{C}(\mathrm{M}, \mathrm{s})$ over the model class $\mathcal{M}$. The topology of the latter is extremely important and is by no means one dimensional. The right portion of the graph is the region of overfitting: so many parameters are used in the model that they begin to directly reflect individual measurements. The left portion of the graph is the region of high apparent randomness: the model captures so little of the data that there is large prediction error.

The input to this explanatory channel, what the modeler A sees, is the data stream s; the output, what B can resynthesize given the explanation X will be denoted s'.

As shown in Figure 2(b) X is transmitted over two subchannels. The first is the modeling channel along which a model M is communicated. The second is the error channel along which an error signal E is transmitted. E is that portion of s unexplained by the model M.

There are two criteria for a good explanation:

1. X must explain s. That is, B must be able to resynthesize the original data: $\mathrm{s}' = \mathrm{s}$.
2. The explanation must be as short as possible. That is, the length $\|\mathrm{X}\| = \|\mathrm{M}\| + \|\mathrm{E}\|$ in bits of X must be minimized.

The efficiency of an explanation or, equivalently, of the model is measured by the compression ratio

$$\mathcal{C}(\mathrm{M}, \mathrm{s}) = \frac{\|\mathrm{X}\|}{\|\mathrm{s}\|} = \frac{\|\mathrm{M}\| + \|\mathrm{E}\|}{\|\mathrm{s}\|} \tag{8}$$

This quantifies the efficacy of an explanation employing model M. $\mathcal{C}$ is then a cost function over the space $\mathcal{M}$ of possible models. The optimal model $\hat{\mathrm{M}}$ then minimizes this cost

$$\mathcal{C}\left(\hat{\mathrm{M}}, \mathrm{s}\right) = \inf_{\mathrm{M} \in \mathcal{M}} \mathcal{C}(\mathrm{M}, \mathrm{s}) \tag{9}$$

Figure 3 illustrates the basic behavior of the cost function $\mathcal{C}$ over the model space $\mathcal{M}$. There are two noteworthy extremes. When the model is trivial, $\mathrm{M} = \emptyset$, it predicts nothing about the data stream. In this case, the entire data stream must be sent along the error channel. $\|\mathrm{E}\| = \|\mathrm{s}\|$ is as large as it can be; but the model is small $\|\mathrm{M}\| = 0$. The second, complementary extreme occurs when the data stream is taken as its own model $\mathrm{M} = \mathrm{s}$. No information needs to be sent on the error channel since the model explains all of the data stream. In this case, $\|\mathrm{M}\| = \|\mathrm{s}\|$ and $\|\mathrm{E}\| = 0$. This is the overfitting regime: model parameters come to fit each measurement.

The view provided by the explanatory channel turns on the existence of an optimal code for a given information source. The semantics is decidedly different, though, in that it interprets the code as consisting of two parts: the model and the error. This is the main innovation beyond Shannon's theory. It apparently was given its first modern articulation by Kemeny[24] as an implementation of Ockham's dictum that "diversity should not be multiplied beyond necessity". It has been put on a more rigorous foundation by Rissanen[38] who adapted the Kolmogorov-Chaitin-Solomonoff algorithmic theory of inductive inference[29] to the needs of universal coding theory, and by Wallace[41] in the domain of statistical classification.

The notion of the explanatory channel might seem nonetheless to be a bit of an abstraction as far as modeling nonlinear processes is concerned. It was implemented, in effect, in a software system for reconstructing the equations of motion from dynamical system time series.[12] The system contained a symbolic dynamical system interpreter as the simulation portion of the channel. The error signal was determined as the deviation of the input trajectory from the deterministic dynamic. Initially, it was averaged by assuming the deviations to be IID Gaussian variables. Optimal models were then selected by minimizing the model entropy* which consisted of a model complexity term and a prediction error term. In this view, the precision of the error signal along different directions in the tangent space to the dynamic is modulated by the spectrum of the associated local Lyapunov characteristic exponents.

Optimal Instruments

The quantitative search for an optimal model extends to criteria for building optimal instruments. In one view at least, the instrument is part of the model. There are two basic principles that are easily summarized

1. Use all the data, and
2. Nothing but the data.

Formally, these translate into the following criteria for instruments.

1. Maximize the conditional entropy $h_\mu(\mathbf{s}|\mathcal{I}(\tau,\epsilon,\beta,\mathrm{D}))$ of the data over the space of instruments. As will be seen later h_μ is readily estimated using the reconstructed model $\hat{\mathrm{M}}(\mathbf{s}|\mathcal{I}(\tau,\epsilon,\beta,\mathrm{D}))$.
2. Minimize the complexity of the reconstructed machine

$$\inf_{\mathrm{M}\in\mathcal{M}} \|\mathrm{M}(\mathbf{s}|\mathcal{I}(\tau,\epsilon,\beta,\mathrm{D}))\| \tag{10}$$

With sufficiently large data sets prediction errors dominate over model size and only the first optimization need be performed. In this regime, early results demonstrated that in accordance with Kolmogorov's theorem the maxima are attained at generating partitions. Going somewhat beyond the theorem, they also showed that the dependence near the maxima was smooth. Later results showed that the order of the conditional entropy maximum is determined by, and so is an indication of, the smoothness of the equations of motion.[10] For finite and especially small data

* A variant of Akaike's Boltzmann Information Criterion for model order selection.[1]

sets, however, the model size plays a significant role. In that regime the criteria are optimized simultaneously over the space of instruments. Exactly how this is done to select the optimal instrument $\hat{I}$ will be left for discussion elsewhere.

The overall picture here is a formalism for implementing the Baconian scientific algorithm of experimentation and refinement. In the drive to understand and predict more of the process, the modeler updates the instrument. An improved model allows the instrument to be modified to remove discovered regularity from the measurements before the information is put into the data stream. In this way, over long times the instrument as transducer provides an increasingly more informative data stream that in principle narrows in on behavior that is less well modeled. One consequence of the coding theoretic view is that, as the instrument takes into account more and more regularity, the resulting data stream from it looks more and more like noise. Concomitantly, the residual regularity requires ever larger inference resources to extract.

Such a high level view of inductive inference is all very well and good; especially in light of the rather large number of parameters that appear. There is one problem, however, that goes to the heart of its coding theoretic premises. This is the almost complete lack of attention to the functional properties of the reconstructed models. It is exactly these properties that have scientific value. Furthermore, that value is independent of the amount of data used to find the model. This problem is reflected in the formalism's ignorance of the topological and metric properties of the model class and range of classes. The claim in the following is that these can be accounted for more directly with a measure of complexity and an investigation of computational properties of individual models. To address these the next section begins to focus on a particular class of models. Once their inference algorithm is outlined and some basic properties described, the discussion examines their utility and semantic content.

Computation from a Time Series

On what sort of structure in the data stream should the models be based? If the goal is prediction, as the preceding assumed, then a natural object to reconstruct from the data series is a representation of the instantaneous state of the process. Unfortunately, as already noted, individual measurements are only indirect representations of the process's state. Indeed, the instrument simply may not supply data of adequate quality in order to discover the true states independent of the amount of data. So how can the process's "effective" states be accessed?

The answer to this turns on a generalization of the "reconstructed states" introduced, under the assumption that the process is a continuous-state dynamical system, by Packard *et al.*[34] The contention there was that a single time series necessarily contained all of the information about the dynamics of that time series. The notion of reconstructed state was based on Poincaré's view of the intrinsic dimension of an object.[36] This was defined as the largest number of successive cuts through the object resulting in isolated points. A sphere in three dimensions by his method is two dimensional since the first cut typically results in a circle and then a second cut, of that circle, isolates two points. Packard *et al.* implemented this using probability distributions conditioned on values of the time series' derivatives. This was, in fact, an implementation of the differential geometric view of the derivatives as locally spanning the graph of the dynamic.

In this reconstruction procedure a state of the underlying process is identified once the conditional probability distribution is peaked. It was noted shortly thereafter that in the presence of extrinsic noise a number of conditions is reached beyond which the conditional distribution is no longer sharpened.[13] And, as a result the process's state cannot be further identified. The width of the resulting distribution then gives an estimate of the effective extrinsic noise level and the minimum number of conditions first leading to this situation, an estimate of the effective dimension.

The method of time derivative reconstruction gives the key to discovering states in discrete times series.* For discrete time series a state is defined to be the set of subsequences that render the future conditionally independent of the past.[14]† Thus, the observer identifies a state at different times in the data stream as its being in identical conditions of ignorance about the future. The set of future subsequences following from a state is called its **morph**.

For this definition of state several reconstruction procedures have been developed. In brief, the simplest method consists of three steps. In the first all length D subsequences in the data stream are represented as paths in a depth D binary "parse" tree. In the second, the morphs are discovered by associating them with the distinct depth $L = D/2$ subtrees found in the parse tree down to depth $D/2$. The number of morphs is then the number of effective states. In the final step, the state to state transitions are found by looking at how each state's associated subtrees map into one another on the parse tree.[11,14,15]

This procedure reconstructs from a data stream a "topological" machine: the skeleton of states and allowed transitions. There are a number of issues concerning statistical estimation, including error analysis and probabilistic structure, that need to be addressed.[16] But this outline suffices for the present purposes. The estimated models are referred to as ϵ-machines in order to indicate their dependence not only on measurement resolution, but also indirectly on all of the instrumental and inferential parameters discussed so far.

ϵ-Machines

The product of machine reconstruction is a set of states that will be associated with a set $\mathbf{V} = \{v\}$ of vertices and a set of transitions associated with a set $\mathbf{E} = \left\{e : e \sim v \underset{s}{\rightarrow} v', \quad v, v' \in \mathbf{V}, s \in \mathbf{A}\right\}$ of labeled edges. Formally, the reconstruction procedure puts no limit on the number of machine states inferred. Indeed, in some important cases the number is infinite, such as at phase transitions.[15] In the following $\mathbf{V}$ will be a finite set and the machines "finitary". One depiction of the reconstructed machine M is as a labeled directed graph $G = \{\mathbf{V}, \mathbf{E}\}$. Examples will be seen shortly. The full probabilistic structure is described by a set of transition matrices

$$\mathcal{T} = \left\{T^{(s)} : \left(T^{(s)}\right)_{vv'} = p_{v \underset{s}{\rightarrow} v'}, \quad v, v' \in \mathbf{V}, s \in \mathbf{A}\right\} \tag{11}$$

* The time delay method appears not to generalize.

† This notion of state is widespread; appearing in various guises in early symbolic dynamics, ergodic, and automata theories. It is the basic notion of state in Markov chain theory.

where $p_{v \underset{s}{\rightarrow} v'}$ denotes the conditional probability to make a transition to state v' from state v on observing symbol s.

A stochastic machine is a compact way of describing the probabilities of a possibly infinite number of measurement sequences. The probability of a given sequence $\mathbf{s}^L = s_0 s_1 s_2 \ldots s_{L-1}, \quad s_i \in \mathbf{A}$, is recovered from the machine by the telescoping product of conditional transition probabilities

$$p\left(\mathbf{s}^L\right) = p_{v_0} p_{v_0 \underset{s_0}{\rightarrow} v_1} p_{v_1 \underset{s_1}{\rightarrow} v_2} \cdots p_{v_{L-1} \underset{s_{L-1}}{\rightarrow} v_n} \tag{12}$$

Here v_0 is the unique start state. It is the state of total ignorance, so that at the first time step we take $p_{v_0} = 1$. The sequence $v_0, v_1, v_2, \ldots, v_{n-1}, v_n$ consists of those states through which the sequence drives the machine. To summarize, a machine is the set $M = \{\mathbf{V}, \mathbf{E}, \mathbf{A}, T, v_0\}$.

Several important statistical properties are captured by the stochastic connection matrix

$$T = \sum_{s \in \mathbf{A}} T^{(s)} \tag{13}$$

where $(T)_{vv'} = p_{v \rightarrow v'}$ is the state to state transition probability, unconditioned by the measurement symbols. By construction every state has an outgoing transition. This is reflected in the fact that T is a stochastic matrix: $\sum_{v' \in \mathbf{V}} p_{vv'} = 1$. It should be clear that by dropping the input alphabet transition labels from the machine the detailed, call it "computational", structure of the input data stream has been lost. All that is retained in T is the state transition structure and this is a Markov chain. The interesting fact is that Markov chains are a proper subset of stochastic finitary machines. Examples later on will support this contention. It is at exactly this step of unlabeling the machine that the "properness" appears.

The stationary state probabilities $\vec{p}_{\mathbf{V}} = \left\{ p_v : \sum_{v \in \mathbf{V}} p_v = 1, v \in \mathbf{V} \right\}$ are given by the left eigenvector of T

$$\vec{p}_{\mathbf{V}} T = \vec{p}_{\mathbf{V}} \tag{14}$$

The entropy rate of the Markov chain is then

$$h_\mu(T) = -\sum_{v \in \mathbf{V}} p_v \sum_{v' \in \mathbf{V}} p_{v \rightarrow v'} \log_2 p_{v \rightarrow v'} \tag{15}$$

This measures the information production rate in bits per time step of the Markov chain. Although the mapping from input strings to the chain's transition sequences is not in general one-to-one, it is finite-to-one. And so, the Markov chain entropy rate is also the entropy rate of the original data source

$$h_\mu(M) = -\sum_{v \in \mathbf{V}} p_v \sum_{v' \in \mathbf{V}} \sum_{s \in \mathbf{A}} p_{v \underset{s}{\rightarrow} v'} \log_2 p_{v \underset{s}{\rightarrow} v'} \tag{16}$$

The complexity* quantifies the information in the state-alphabet sequences

$$C_\mu(M) = H(\vec{p}_{\mathbf{V}}) = -\sum_{v \in \mathbf{V}} p_v \log_2 p_v \tag{17}$$

It measures the amount of memory in the process. For completeness, note that there is an edge-complexity that is the information contained in the asymptotic edge distribution $\vec{p}_{\mathbf{E}} = \left\{ p_e = p_v p_{v \underset{s}{\to} v'} : v, v' \in \mathbf{V}, e \in \mathbf{E}, s \in \mathbf{A} \right\}$

$$C_\mu^e(M) = -\sum_{e \in \mathbf{E}} p_e \log_2 p_e \tag{18}$$

These quantities are not independent. Conservation of information at each state leads to the relation

$$C_\mu^e = C_\mu + h_\mu \tag{19}$$

And so, there are only two independent quantities when modeling a process as a stochastic finitary machine. The entropy h_μ, as a measure of the diversity of patterns, and the complexity C_μ, as a measure of memory, have been taken as the two elementary coordinates with which to analyze a range of sources.[15]

There is another set of quantities that derive from the skeletal structure of the machine. Dropping all of probabilistic structure, the growth rate of the number of sequences it produces is the topological entropy

$$h = \log_2 \lambda(T_0) \tag{20}$$

where λ_0 is the principle eigenvalue of the connection matrix $T_0 = \sum_{s \in \mathbf{A}} T_0^{(s)}$. The latter is formed from the labeled matrices

$$\left\{ T_0^{(s)} : \left(T_0^{(s)}\right)_{vv'} = \begin{cases} 1 & p_{v \underset{s}{\to} v'} > 0 \\ 0 & \text{otherwise} \end{cases} s \in \mathbf{A} \right\} \tag{21}$$

The state and transition topological complexities are

$$\begin{aligned} C &= \log_2 \|\mathbf{V}\| \\ C^e &= \log_2 \|\mathbf{E}\| \end{aligned} \tag{22}$$

In computation theory, an object's complexity is generally taken to be the size in bits of its representation. The quantities just defined measure the complexity of the reconstructed machine. As will be seen in the penultimate section, when these entropies and complexities, both topological and metric, are integrated into a single parametrized framework, a thermodynamics of machines emerges.

* Within the reconstruction hierarchy this is actually the finitary complexity, since the context of the discussion implies that we are considering processes with a finite amount of memory. However, I have not introduced this restriction in unnecessary places in the discussion. The finitary complexity has been considered before in the context of generating partitions and known equations of motion.[19,31,42]

Complexity

It is useful at this stage to stop and reflect on some properties of the models that we have just described how to reconstruct. Consider two extreme data sources. The first, highly predictable, produces a streams of 1s; the second, highly unpredictable, is an ideal random source of a binary symbols. The parse tree of the predictable source is a single path of 1s. And there is a single subtree, at any depth. As a result the machine has a single state and a single transition on $s = 1$: a simple model of a simple source. For the ideal random source the parse tree, again to any depth, is the full binary tree. All paths appear in the parse tree since all binary subsequences are produced by the source. There is a single subtree, of any morph depth at all parse tree depths: the full binary subtree. And the machine has a single state with two transitions; one on $s = 1$ and one on $s = 0$. A simple machine, even though the source produces the widest diversity of binary sequences.

A simple gedanken experiment serves to illustrate how complexity is a measure of a machine's memory capacity. Consider two observers **A** and **B**, each with the same model **M** of some process. **A** is allowed to start machine **M** in any state and uses it to generate binary strings that are determined by the edge labels of the transitions taken. These strings are passed to observer **B** which traces there effect through its own copy of **M**. On average how much information about **M**'s state can **A** communicate to **B** via the binary strings? If the machine describes (say) a period three process, e.g. it outputs strings like 101101101..., 011011011..., and 110110110..., it has $\|\mathbf{V}\| = 3$ states. Since **A** starts **M** in different states, **B** can learn only the information of the process's phase in the period 3 cycle. This is $\log_2 \|\mathbf{V}\| = 1.584...$ bits of information about the process's state, if **A** chooses the initial states with equal probability. However, if the machine describes an idea random binary process, by definition **A** can communicate no information to **B**, since there is no structure in the sequences to use for this purpose. This is reflected in the fact, as already noted above, that the corresponding machine has a single state and its complexity is $\log_2 1 = 0$. In this way, a process's complexity is the amount of information that someone controlling its start state can communicate to another.

These examples serve to highlight one of the most basic properties of complexity, as I use the term. Both predictable and random sources are simple in the sense that their models are small. Complex processes in this view have large models. In computational terms, complex processes have, as a minimum requirement, a large amount of memory as revealed by many internal states in the reconstructed machine. Most importantly, that memory is structured in particular ways that support different types of computation. The sections below on knowledge and meaning show several consequences of computational structure.

In the most general setting, I use the word "complexity" to refer to the amount of information contained in observer-resolvable equivalence classes. For finitary machines, the complexity is measured by the quantities labeled above by C. This notion has been referred to as the "statistical complexity" in order to distinguish it from the Chaitin-Kolmogorov complexity,[9,27] the Lempel-Ziv complexity,[28] Rissanen's stochastic complexity,[38] and others[45,44] which are all equivalent in the limit of long data streams to the process's Kolmogorov-Sinai entropy $h_\mu\left(\vec{X}\right)$. If the instrument is generating and $\mu\left(\vec{X}\right)$ is absolutely continuous, these quantities are given by the

entropy rate of the reconstructed machine, Eq. (16).[7] Accordingly, I use the word "entropy" to refer to such quantities. They measure the diversity of sequences a process produces. Implicit in their definitions is the restriction that the modeler must pay computationally for each random bit. Simply stated, the overarching goal is exact description of the data stream. In the modeling approach advocated here the modeler is allowed to flip a coin or to sample the heat bath to which it may be coupled. "Complexity" is reserved in my vocabulary to refer to a process's structural properties, such as memory and other types of computational capacity.

This is not the place to review the wide range of alternative notions of "complexity" that have been discussed more recently in the physics and dynamics literature. The reader is referred to the comments and especially the citations elsewhere.[14,15] It is important to point out, however, that the notion defined here does not require knowledge of the equations of motion, the prior existence of exact conditional probabilities, Markov or even generating partitions of the state space, continuity and differentiability of the state variables, nor the existence of periodic orbits. Furthermore, the approach taken here differs from those based on the construction of universal codes in the emphasis on the model's structure. That emphasis brings it into direct contact with the disciplines of stochastic automata, formal language theory, and thermodynamics.

Finally, statistical complexity is a highly relative concept that depends directly on the assumed model class. In the larger setting of hierarchical reconstruction it becomes the finitary complexity since it measures the number of states in a finite state machine representation. But there are other versions appropriate, for example, when the finitary complexity diverges.[11]

Causality

There are a few points that must be brought out concerning what these reconstructed machines represent. First, by the definition of future-equivalent states, the machines give the minimal information dependency between the morphs. In this respect, they represent the causality of the morphs considered as events. The machines capture the information flow within the given data stream. If state B follows state A then A is a cause of B and B is one effect of A. Second, machine reconstruction produces minimal models up to the given prediction error level. This minimality guarantees that there are no other events (morphs) that intervene, at the given error level, to render A and B independent. In this case, we say that information flows from A to B. The amount of information that flows is the negative logarithm of the connecting edge probability. Finally, time is the natural ordering captured by machines. An ϵ-machine for a process is then the minimal causal representation reconstructed using the least powerful computational model class that yields a finite complexity.

KNOWLEDGE RELAXATION

The next two sections investigate how models can be used by an observer. An observer's knowledge $\mathcal{K}_P$ of a process P consists of the data stream, its current model, and how the information used to build the model was obtained.* Here the latter is given by the measuring

* In principle, the observer's knowledge also consists of the reconstruction method and its various assumptions. But it is best to not elaborate this here. These and other unmentioned variables are assumed to be fixed.

instrument $\mathcal{I} = \left\{\Pi_\epsilon^\beta(\mathrm{D}), \tau\right\}$. To facilitate interpretation and calculations, the following will assume a simple data acquisition discipline with uniform sampling interval τ and a time-independent zero temperature measurement partition Π_ϵ. Further simplification comes from ignoring external factors, such as what the observer intends or needs to do with the model, by assuming that the observer's goal is solely optimal prediction with respect to the model class of finitary machines.

The totality of knowledge available to an observer is given by the development of its $\mathcal{K}_P$ at each moment during its history. If we make the further assumption that by some agency the observer has at each moment in its history optimally encoded the available current and past measurements into its model, then the totality of knowledge consists of four parts: the time series of measurements, the instrument by which they were obtained, and the current model and its current state. Stating these points so explicitly helps to make clear the upper bound on what the observer can know about its environment. Even if the observer is allowed arbitrary computational resources, given either finite information from a process or finite time, only a finite amount of structure can be inferred.

An ϵ-machine is a representations of an observer's model of a process. To see its role in the change in $\mathcal{K}_P$ consider the situation in which the model structure is kept fixed. Starting from the state v_0 of total ignorance about the process's state, successive steps through the machine lead to a refinement of the observer's knowledge as determined by a sequence of measurements. The average increase in $\mathcal{K}_P$ is given by a diffusion of information throughout the model. The machine transition probabilities, especially those connected with transient states, govern how the observer gains more information about the process with longer measurement sequences.

A measure of information relaxation on finitary machines is given by the time-dependent finitary complexity

$$C_\mu(t) = H\left(\vec{p}_{\mathrm{V}}(t)\right) \tag{23}$$

where $H(P) = \sum_{p_i \in P} p_i \log_2 p_i$ is the Shannon entropy of the distribution $P = \{p_i\}$ and

$$\vec{p}_{\mathrm{V}}(t+1) = \vec{p}_{\mathrm{V}}(t)\, T \tag{24}$$

is the probability distribution at time t beginning with the initial distribution $p_{\mathrm{V}}(0) = (1, 0, 0, \ldots)$ concentrated on the start state. This distribution represents the observer's state of total ignorance of the process's state, i.e. before any measurements have been made, and correspondingly $C_\mu(0) = 0$. $C_\mu(t)$ is simply (the negative of) the Boltzmann H-function in the present setting. And we have the analogous result to the H-theorem for stochastic ϵ-machines: $C_\mu(t)$ converges monotonically when $\vec{p}_{\mathrm{V}}(t)$ is sufficiently close to $\vec{p}_{\mathrm{V}} = \vec{p}_{\mathrm{V}}(\infty)$: $C_\mu(t) \underset{t\to\infty}{\to} C_\mu$. That is, the time-dependent complexity limits on the finitary complexity. Furthermore, the observer has the maximal amount of information about the process, i.e. the observer's knowledge is in equilibrium with the process, when $C_\mu(t+1) - C_\mu(t)$ vanishes for all $t > t_{\mathrm{lock}}$, where t_{lock} is some fixed time characteristic of the process.

For finitary machines there are two convergence behaviors for $C_\mu(t)$. These are illustrated in figure 4 for three processes: one P_3 which is period 3 and generates $(101)^*$, one P_2 in which

only isolated zeros are allowed, and one P_1 that generates 1s in blocks of even length bounded by 0s. The first behavior type, illustrated by P_3 and P_2, is monotonic convergence from below. In fact, the asymptotic approach occurs in finite time. This is the case for periodic and recurrent Markov chains, where the latter refers to finite state stochastic processes whose support is a subshift of finite type (SSFT). The convergence here is over-damped.

The second convergence type, illustrated by P_1, is only asymptotic; convergence to the asymptotic state distribution is only at infinite time. There are two subcases. The first is monotonic increasing convergence; the conventional picture of stochastic process convergence. The second subcase (P_1) is nonmonotonic convergence. In this case, starting in the condition of total ignorance leads to a critically-damped convergence with a single overshoot of the finitary complexity. With other initial distributions oscillations, i.e. underdamped convergence, can be seen. Exact convergence is only at infinite time. This convergence type is associated with machines having cycles in the transient states or, in the classification of symbolic dynamics, with machines whose support is a strictly Sofic[32] system (SSS).* For these, at some point in time the initial distribution spreads out over more than just the recurrent states. $C_\mu(t)$ can then be larger than C_μ. Beyond this time, it converges from above. Much of the detailed convergence behavior is determined, of course, by T's full eigenvalue spectrum. The interpretation just given, though, can be directly deduced by examining the reconstructed machine's graph G. One aspect which is less immediate is that for SSSs the initial distribution relaxes through an infinite number of Cantor sets in sequence space. For SSFTs there is only a finite number of Cantor sets.

This structural analysis indicates that the ratio

$$\Delta C_\mu(t) = \frac{|C_\mu - C_\mu(t)|}{C_\mu} \tag{25}$$

is largely determined by the amount of information in the transient states. For SSSs this quantity only asymptotically vanishes since there are transient cycles in which information persists for all time, even though their probability decreases asymptotically. This leads to a general definition of (chaotic or periodic) phase and phase locking. The phase of a machine at some point in time is its current state. There are two types of phase of interest here. The first is the process's phase and the second is the observer's phase which refers to the state of the observer's model having read the data stream up to some time. The observer has β-locked onto the process when $\Delta C_\mu(t_{lock}) < \beta$. This occurs at the locking time t_{lock} which is the longest time t such that $\Delta C_\mu(t) = \beta$. When the process is periodic, this notion of locking is the standard one from engineering. But it also applies to chaotic processes and corresponds to the observer knowing what state the process is in, even if the next measurement cannot be predicted exactly.

These two classes of knowledge relaxation lead to quite different consequences for an observer even though the processes considered above all have a small number of states (2 or 3) and share the same single-symbol statistics: $Pr(s=1) = \frac{2}{3}$ and $Pr(s=0) = \frac{1}{3}$. In the over-damped case, the observer knows the state of the underlying process with certainty after a finite time. In the critically-damped situation, however, the observer has only approximate knowledge for all times. For example, setting $\beta = 1\%$ leads to locking times shown in table 1.

* SSS shall also refer, in context, to stochastic Sofic systems.[25]

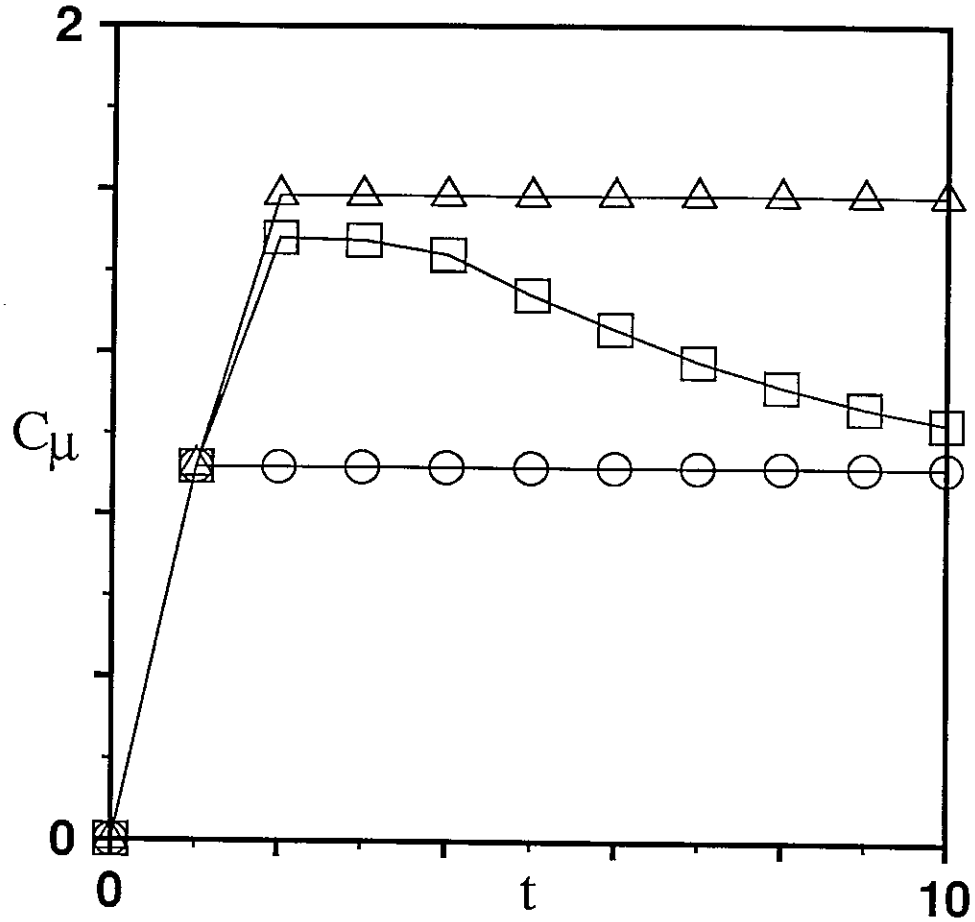

Figure 4 Temporal convergence of the complexity $C_\mu(t)$ for a period 3 process P_3 (triangles), a Markovian process P_2 whose support is a subshift of finite type (circles), and a process P_1 that generates blocks of even numbers of 1s surrounded by 0s (squares).

Thus, the ability of an observer to infer the state depends crucially on the process's computational structure, viz. whether its topological machine is a SSFT or a SSS. The presence of extrinsic noise and observational noise modify these conclusions systematically.

It is worthwhile to contrast the machine model of P_1 with a model based on histograms, or look-up tables, of the same process. Both models are given sufficient storage to exactly represent the length 3 sequence probability distribution. They are then used for predictions on length 4 sequences. The histogram model will store the probabilities for each length 3 sequence. This requires 8 bins each containing an 8 bit approximation of a rational number: 3 bits for the numerator and 5 for the denominator. The total is 67 bits which includes an indicator for the most recent length 3 sequence. The machine model, see Figure 5, must store the current state and five approximate rational numbers, the transition probabilities, using 3 bits each: one for the numerator and two for the denominator. This gives a model size of 17 bits.

Two observers, each given one or the other model, are presented with the sequence 101. What do they predict for the event that the fourth symbol is $s = 1$? The histogram model predicts

$$Pr(1|101) \approx Pr(1|01) = \frac{Pr(011)}{Pr(11)} = \frac{1/6}{4/9} = \frac{3}{8} \tag{26}$$

Locking Times at 1% Level	
Process	**Locked at time**
Period 3	2
Isolated 0s	1
Even 1 blocks	17

Table 1 β-locking times for the periodic P_3, isolated 0s P_2, and even 1s P_1, processes. Note that for the latter the locking time is substantially longer and depends on β. For the former two, the locking times indicate the times at which asymptotic convergence has been achieved. The observer knows the state of the underlying process with certainty at those locking times. For P_1, however, at $t = 17$ the observer is partially phase-locked with knowledge of 99% of the process's state information.

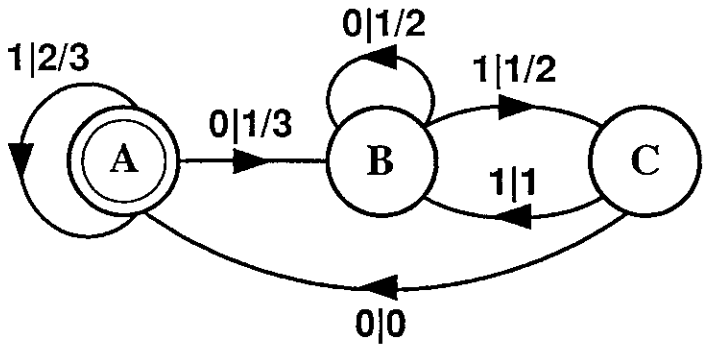

Figure 5 The even system generates sequences $\{\ldots 01^{2n}0\ldots : n = 0, 1, 2, \ldots\}$ of 1s of even length, i.e. even parity. There are three states $\mathrm{V} = \{\mathrm{A}, \mathrm{B}, \mathrm{C}\}$. The state A with the inscribed circle is the start state v_0. The edges are labeled $s|p$ where $s \in \mathrm{A}$ is a measurement symbol and $p \in [0, 1]$ is a conditional transition probability.

whereas the machine model predicts

$$Pr\,(1|101) = p_{\mathrm{C}\to\mathrm{B}} = 1 \tag{27}$$

The histogram model gives the wrong prediction. It says that the fourth symbol is uncertain when it is completely predictable. A similar analysis for the prediction of measuring $s = 1$ having observed 011 shows the opposite. The histogram model predicts $s = 1$ is more likely $p_{s=1} = 2/3$; when it is, in fact, not predictable at all $p_{s=1} = 1/2$. This example is illustrative of the superiority of stochastic machine models over histogram and similar look-up table models of time-dependent processes. In fact, there are processes with finite memory for which no finite-size sequence histogram will give correct predictions.

In order to make the physical relevance of SSSs and their slow convergence more plausible, the next example is taken from the Logistic map at a Misiurewicz parameter value. The Logistic map is an iterated mapping of the unit interval

$$x_{n+1} = f_r\,(x_n) = r x_n\,(1 - x_n)\,, \quad r \in [0, 4]\,, x_0 \in [0, 1] \tag{28}$$

The control parameter r governs the degree of nonlinearity. At a Misiurewicz parameter value the chaotic behavior is governed by an absolutely continuous invariant measure. The consequence is that the statistical properties are particularly well-behaved. These parameter values are

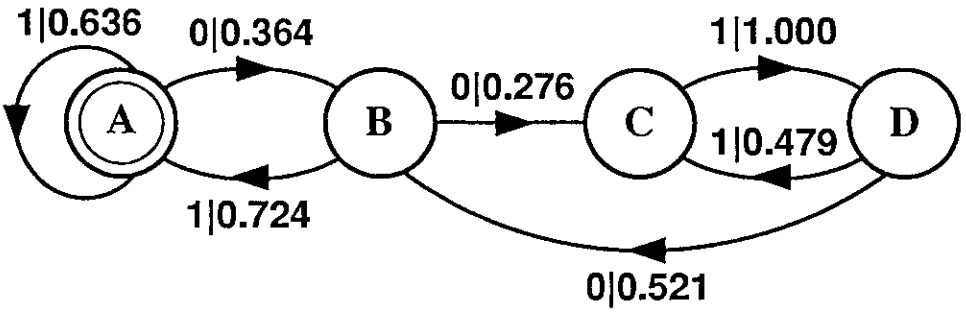

Figure 6 The machine $M_{r'}^{\rightarrow}$ reconstructed by parsing in forward presentation order a binary sequence produced using a generating partition of the Logistic map at a Misiurewicz parameter value.

determined by the condition that the iterates $f^n(x_c)$ of the map's maximum $x_c = 1/2$ are asymptotically periodic. The Misiurewicz parameter value r' of interest here is the first root of $f_{r'}^4(x_c) = f_{r'}^5(x_c)$ below that at $r = 4$. Solving numerically yields $r' \approx 3.9277370017867516$. The symbolic dynamics is produced from the measurement partition $\Pi_{1/2} = \{[0, x_c], (x_c, 1]\}$. Since this partition is generating the resulting binary sequences completely capture the statistical properties of the map. In other words, there is a one-to-one mapping between infinite binary sequences and almost all points on the attractor.

Reconstructing the machine from one very long binary sequence in the direction in which the symbols are produced gives the four state machine $M_{r'}^{\rightarrow}$ shown in figure 6. The stochastic connection matrix is

$$T = \begin{pmatrix} 0.636 & 0.364 & 0.000 & 0.000 \\ 0.724 & 0.000 & 0.276 & 0.000 \\ 0.000 & 0.000 & 0.000 & 1.000 \\ 0.000 & 0.521 & 0.479 & 0.000 \end{pmatrix} \tag{29}$$

Reconstructing the machine from the same binary sequence in the opposite direction gives the reverse-time machine $M_{r'}^{\leftarrow}$ shown in figure 7. It connection matrix is

$$T = \begin{pmatrix} 0.636 & 0.364 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.276 & 0.724 \\ 0.000 & 0.000 & 0.000 & 1.000 \\ 0.000 & 1.000 & 0.000 & 0.000 \end{pmatrix} \tag{30}$$

Notice that $M_{r'}^{\leftarrow}$ has a transient state and three recurrent states compared to the four recurrent states in $M_{r'}^{\rightarrow}$. This suggests the likelihood of some difference in complexity convergence. Figure 8 shows that this is the case by plotting $C_\mu(M_{r'}^{\rightarrow}, t)$ and $C_\mu(M_{r'}^{\leftarrow}, t)$ for positive and negative times, respectively. Not only do the convergence behaviors differ in type, but also in the asymptotic values of the complexities: $C_\mu(M_{r'}^{\rightarrow}) \approx 1.77$ bits and $C_\mu(M_{r'}^{\leftarrow}) \approx 1.41$ bits. This occurs despite the fact that the entropies must be and are the same for both machines: $h(M_{r'}^{\rightarrow}) = h(M_{r'}^{\leftarrow}) \approx 0.82$ bits per time unit and $h_\mu(M_{r'}^{\rightarrow}) = h_\mu(M_{r'}^{\leftarrow}) \approx 0.81$ bits per time unit. Although the data stream is equally unpredictable in both time directions, an observer learns about the process's state in two different ways and obtains different amounts of state information. The difference

$$\Delta C_{\leftrightarrows} = C_\mu(M_{r'}^{\rightarrow}) - C_\mu(M_{r'}^{\leftarrow}) \approx 0.36 \text{ bits} \tag{31}$$

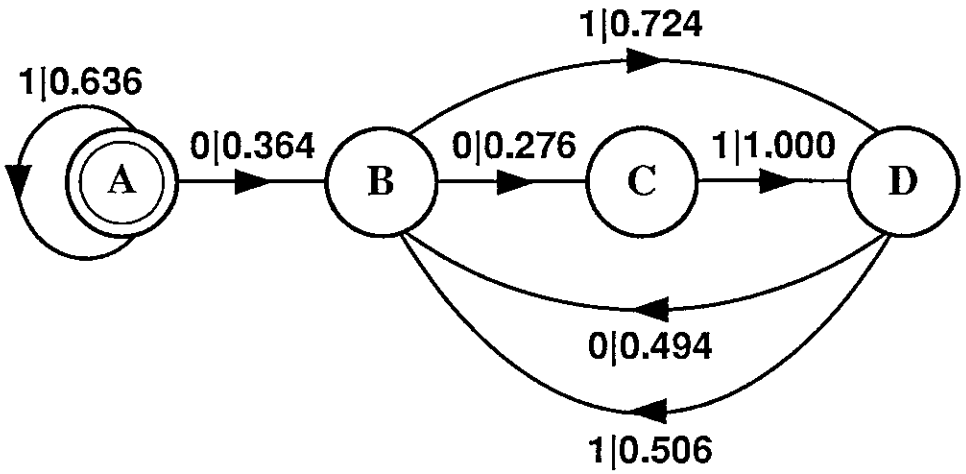

Figure 7 The machine $M_r^{\leftarrow}$ reconstructed by parsing in reverse presentation order a binary sequence produced using a generating partition of the Logistic map at a Misiurewicz parameter value.

is a measure of the computational irreversibility of the process. It indicates the process is not symmetric in time from the observer's viewpoint. This example serves to distinguish machine reconstruction and the derived quantifiers, such as complexity, from the subsequence-based measures, such as the two-point mutual information and the excess entropy.

MEASUREMENT SEMANTICS OF CHAOS

Shannon's communication theory tells one how much information a measurement gives. But what is the meaning of a particular measurement? Sufficient structure has been developed up to this point to introduce a quantitative definition of an observation's meaning. Meaning, as will be seen, is intimately connected with hierarchical representation. The following, though, concerns meaning as it arises when crossing a single change in representation and not in the entire hierarchy.[11]

A universe consisting of an observer and a thing observed has a natural semantics. The semantics describes the coupling that occurs during measurement. The attendant meaning derives from the dual interpretation of the information transferred at that time. As already emphasized, the measurement is, first, an indirect representation of the underlying process's state and, second, information that updates the observer's knowledge. The semantic information processing that occurs during a measurement thus turns on the relationship between two levels of representation of the same event.

The meaning of a message, of course, depends on the context in which its information is made available. If the context is inappropriate, the observation will have no basis with which to be understood. It will have no meaning. If appropriate, then the observation will be "understood". And if that which is understood — the content of the message — is largely unanticipated then the observation will be more significant than a highly likely, "obvious" message.

In the present framework context is set by the model held by the observer at the time of a measurement. To take an example, assume that the observer is capable of modeling using the class of stochastic finite automata. And, in particular, assume the observer has estimated a

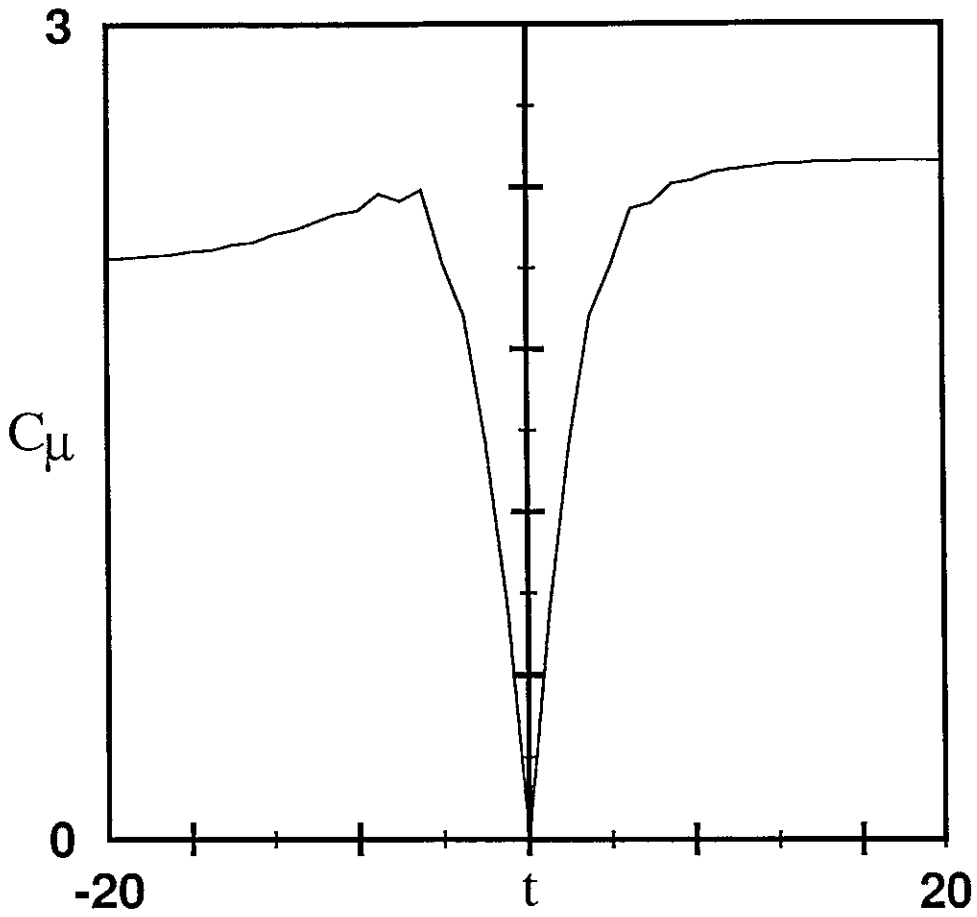

Figure 8 What the observer sees, on average, in forward and reverse lag time in terms of the complexity convergence $C_\mu(t)$ for $M_{\overrightarrow{r}}$ and $M_{\overleftarrow{r}}$. Data for the latter are plotted on the negative lag time axis. Note that not only do the convergence characteristics differ between the two time directions, but the asymptotic complexity values are not equal.

stochastic finite automaton* and has been following the process sufficiently long to know the current state with certainty. Then at a given time the observer measures symbol $s \in \mathrm{A}$. If that measurement forces a disallowed transition, then it has no meaning other than that it lies outside of the contexts (morphs) captured in the current model. The observer clearly does not know what the process is doing. Indeed, formally the response is for the observer to reset the machine to the initial state of total ignorance. If, however, the measurement is associated with an allowed transition, i.e. it is anticipated, then the amount $\Theta(s)$ of meaning is

$$\Theta(s) = -\log p_{\underset{s}{\rightarrow} v} \tag{32}$$

Here $\underset{s}{\rightarrow} v$ denotes the machine state $v \in \mathrm{V}$ to which the measurement brings the observer's knowledge of the process's state. $p_{\underset{s}{\rightarrow} v}$ is the corresponding morph's probability which is given

* Assume also that the estimated machine is deterministic in the sense of automata theory: the transitions from each state are uniquely labeled: $p_e = p(v, v', s) = p_v p_{v \underset{s}{\rightarrow} v'}$. This simplifies the discussion by avoiding the need to define the graph indeterminacy as a quantitative measure of ambiguity.[14] Ambiguity for an observer arises if its model is a stochastic nondeterministic automaton.

by the associated state's asymptotic probability. The meaning itself, i.e. the content of the observation, is the particular morph to which the model's updated state corresponds. In this view a measurement selects a particular pattern from a palette of morphs. The measurement's meaning is the selected morph* and the amount of meaning is determined by the latter's probability.

To clarify these notions, let's consider as an example a source that produces infinite binary sequences for the regular language[21] described by the expression $(0 + 11)^*$. We assume further that the choice implied by the "+" is made with uniform probability. An observer given an infinite sequence of this type reconstructs the stochastic finite machine shown in figure 5. The observer has discovered three morphs: the states $\mathbf{V} = \{\mathrm{A}, \mathrm{B}, \mathrm{C}\}$. But what is the meaning of each morph? First, consider the recurrent states B and C. State B is associated with having seen an even number of 1's following a 0; C with having seen an odd number. The meaning of B is "even" and C is "odd". Together the pair $\{\mathrm{B}, \mathrm{C}\}$ recognize the parity of the data stream. The machine as a whole accepts strings whose substrings of the form $01\ldots10$ have even parity of 1s. What is the meaning of state A? As long as the observer's knowledge of the process's state remains in state A, there has been some number of 1's whose parity is unknown, since a 0 must be seen to force the transition to the parity state B. This state, a transient, serves to synchronize the recurrent states with the data stream. This indicates the meaning content of an individual measurement in terms of the state to which it and its predecessors bring the machine.

Before giving a quantitative analysis the time dependence of the state probabilities must be calculated. Recall that the state probabilities are updated via the stochastic connection matrix

$$\vec{p}_{\mathbf{V}}(t+1) = \vec{p}_{\mathbf{V}}(t) \begin{pmatrix} \frac{2}{3} & \frac{1}{3} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 1 & 0 \end{pmatrix} \tag{33}$$

where $\vec{p}_{\mathbf{V}}(t) = (p_{\mathrm{A}}(t), p_{\mathrm{B}}(t), p_{\mathrm{C}}(t))$ and the initial distribution is $\vec{p}_{\mathbf{V}}(0) = (1, 0, 0)$. The time-dependent state probabilities are found using z-transforms to be

$$\begin{aligned} p_{\mathrm{A}}(t) &= \left(\frac{2}{3}\right)^t && t = 0, 1, 2, \ldots \\ p_{\mathrm{B}}(t) &= 2\left(\frac{2}{3}\right)^t - 2^{1-t} && t = 0, 1, 2, \ldots \\ p_{\mathrm{C}}(t) &= \begin{cases} \left(\frac{2}{3}\right)^t - 2^{1-t} & t = 1, 2, 3, \ldots \\ 0, & t = 0 \end{cases} \end{aligned} \tag{34}$$

Any time a disallowed transition is forced the current state is reset to the start state and $\vec{p}_{\mathbf{V}}(t)$ is reset to the distribution representing total ignorance which is given by $\vec{p}_{\mathbf{V}}(0)$.

What then is the quantitative degree of meaning of particular measurements? Let's consider all of the possibilities: all possible contexts, i.e. current states, and all possible measurements. t steps after a reset, the observer is

1. In the sync state and measures $s = 1$: $\Theta^t_{\mathrm{sync}}(1) = -\log_2 p_{\to \mathrm{A}} = t(\log_2 3 - 1)$;

* I simplify here. The best formal representation of meaning at present uses the set-theoretic structure that the machine induces over the set of observed subsequences. This in turn is formulated via the lattice theory[3] of machines.[20]

Observer's Semantic Analysis of Parity Source				
Observer in State	**Measures Symbol**	**Interprets Meaning as**	**Degree of Meaning (bits)**	**Amount of Information (bits)**
A	1	Unsynchronized	Infinity	0.585
A	0	Synchronize	0.585	1.585
B	1	Odd number of 1s	1.585	1
B	0	Even number of 1s	0.585	1
C	1	Even number of 1s	0.585	0
C	0	Confusion: lose sync, reset to start state	0	Infinity

Table 2 The observer's semantics for measuring the parity process of Figure 5.

2. In the sync state and measures $s = 0$: $\Theta^t_{\text{sync}}(0) = -\log_2 p_{\underset{0}{\rightarrow}\text{B}} = -\log_2 p_{\text{B}}(t)$; e.g. $\Theta^1_{\text{sync}}(0) = \log_2 3 \approx 1.584$ bits;
3. In the even state and measures $s = 1$: $\Theta^t_{\text{even}}(1) = -\log_2 p_{\underset{1}{\rightarrow}\text{C}} = -\log_2 p_{\text{C}}(t), t > 1$; e.g. $\Theta^2_{\text{even}}(1) = \log_2 6 \approx 2.584$ bits;
4. In the even state and measures $s = 0$: $\Theta^t_{\text{even}}(0) = -\log_2 p_{\underset{0}{\rightarrow}\text{B}} = -\log_2 p_{\text{B}}(t)$; e.g. $\Theta^2_{\text{even}}(0) = 1 + 2\log_2 3 - \log_2 7 \approx 1.372$ bits;
5. In the odd state and measures $s = 1$: $\Theta^t_{\text{odd}}(1) = -\log_2 p_{\underset{1}{\rightarrow}\text{B}} = -\log_2 p_{\text{B}}(t)$; e.g. $\Theta^3_{\text{odd}}(1) = 2 + 3\log_2 3 - \log_2 37 \approx 1.545$ bits;
6. In the odd state and measures $s = 0$, a disallowed transition. The observer resets the machine: $\Theta^t_{\text{odd}}(0) = -\log_2 p_{\underset{0}{\rightarrow}\text{A}} = -\log_2 p_{\text{A}}(0) = 0$.

In this scheme states B and C cannot be visited at time $t = 0$ nor state C at time $t = 1$.

Assuming no disallowed transitions have been observed, at infinite time $\vec{p}_{\text{V}} = \left(0, \frac{2}{3}, \frac{1}{3}\right)$ and the degrees of meaning are, if the observer is

1. In the sync state and measures $s = 1$: $\Theta_{\text{sync}}(1) = -\log_2 p_{\underset{1}{\rightarrow}\text{A}} = \infty$;
2. In the sync state and measures $s = 0$: $\Theta_{\text{sync}}(0) = -\log_2 p_{\underset{0}{\rightarrow}\text{B}} = \log_2 3 - 1 \approx 0.584$ bits;
3. In the even state and measures $s = 1$: $\Theta_{\text{even}}(1) = -\log_2 p_{\underset{1}{\rightarrow}\text{C}} = \log_2 3 \approx 1.584$ bits;
4. In the even state and measures $s = 0$: $\Theta_{\text{even}}(0) = -\log_2 p_{\underset{0}{\rightarrow}\text{B}} = \log_2 3 - 1 \approx 0.584$ bits;
5. In the odd state and measures $s = 1$: $\Theta_{\text{odd}}(1) = -\log_2 p_{\underset{1}{\rightarrow}\text{B}} = \log_2 3 - 1 \approx 0.584$ bits;
6. In the odd state and measures $s = 0$, a disallowed transition. The observer resets the machine: $\Theta_{\text{odd}}(0) = -\log_2 p_{\underset{0}{\rightarrow}\text{A}} = -\log_2 p_{\text{A}}(0) = 0$.

Table 2 summarizes this analysis for infinite time. It also includes the amount of information gained in making the specified measurement. This is given simply by the negative binary logarithm of the associated transition probability.

Similar definitions of meaning can be developed between any two levels in a reconstruction hierarchy. The example just given concerns the semantics between the measurement symbol level and the stochastic finite automaton level.[11] Meaning appears whenever there is a change in representation of events. And if there is no change, e.g. a measurement is considered only with respect to the population of other measurements, an important special case arises.

In this view Shannon information concerns degenerate meaning: that obtained within the same representation class. Consider the information of events in some set E of possibilities whose occurrence is governed by arbitrary probability distributions $\{P, Q, \ldots\}$. Assume that no further structural qualifications of this representation class are made. Then the Shannon self-information $-\log p_e$, $p_e \in P$, gives the degree of meaning $-\log_2 p_{\rightarrow e}$ in the observed event e with respect to total ignorance. Similarly, the information gain $I(P;Q) = \sum_{e \in E} p_e \log_2 \frac{p_e}{q_e}$ gives the average degree of "meaning" between two distributions. The two representation levels are degenerate: both are the events themselves. Thus, Shannon information gives the degree of meaning of an event with respect to the set E of events and not with respect to an observer's internal model; unless, of course, that model is taken to be the collection of events as in a histogram or look-up table. Although this might seem like vacuous re-interpretation, it is essential that general meaning have this as a degenerate case.

The main components of meaning, as defined above should be emphasized. First, like information it can be quantified. Second, conventional uses of Shannon information are a natural special case. And third, it derives fundamentally from the relationship *across* levels of abstraction. A given message has different connotations depending on an observer's model and the most general constraint is the model's level in a reconstruction hierarchy. When model reconstruction is considered to be a time-dependent process that moves up a hierarchy, then the present discussion suggests a concrete approach to investigating adaptive meaning in evolutionary systems: emergent semantics.

In the parity example above I explicitly said what a state and a measurement "meant". Parity, as such, is a human linguistic and mathematical convention, which has a compelling naturalness due largely to its simplicity. A low level organism, though, need not have such a literary interpretation of its stimuli. Meaning of (say) its model's states, when the state sequence is seen as the output of a preprocessor,* derives from the functionality given to the organism, as a whole and as a part of its environment and its evolutionary and developmental history. Said this way, absolute meaning in nature is quite a complicated and contingent concept. Absolute meaning derives from the global structure developed over space and through time. Nonetheless, the analysis given above captures the representation level-to-level origin of "local" meaning. The tension between global and local entities is not the least bit new to nonlinear dynamics. Indeed, much of the latter's subtlety is a consequence of their inequivalence. Analogous insights are sure to follow from the semantic analysis of large hierarchical processes.

* This preprocessor is a transducer version of the model that takes the input symbols and outputs strings in the state alphabet V.

Machine Thermodynamics

The atomistic view of nature, though professed since ancient times, was largely unsuccessful until the raw combinatorial complication it entailed was connected to macroscopic phenomena. Founding thermodynamics on the principles of statistical mechanics was one of, if not the major, influence on its eventual acceptance. The laws of thermodynamics give the coarsest constraints on the microscopic diversity of large many-particle systems. This same view, moving from microscopic dynamics to macroscopic laws, can be applied to the task of statistical inference of nonlinear models. And so it is appropriate after discussing the "microscopic" data of measurement sequences and the reconstruction of "mesoscopic" machines from them, to end with a discussion at the largest scale of description: machine thermodynamics. This gives a concise description of the structure of the infinite set of infinite sequences generated by a machine and also of their probabilities. It does this, in analogy with the conventional thermodynamic treatment of microstates, by focusing on different subsets of allowed sequences.

The first step is the most basic: identification of the microstates. Consistent with machine reconstruction's goal to approximate a process's internal states, microstates in modeling are the individual measurement subsequences.* Consider the set $\mathrm{sub}_L(\mathbf{s})$ of all length L subsequences occurring in a length N data stream $\mathbf{s}$. The probability of a subsequence $\omega \in \mathrm{sub}_L(\mathbf{s})$ is estimated by $p_\omega \approx N^{-1}N_\omega$, where N_ω is the number of occurrences of ω in the data stream. The connection with the physical interpretation of thermodynamics follows from identifying a microstate's energy with its self-information

$$U_\omega = -\log_2 p_\omega \tag{35}$$

That is, improbable microstates have high energy. Energy macrostates are then given by grouping subsequences of the same energy U into subsets $\{\omega : U_\omega = \mathrm{U}, \omega \in \mathrm{sub}_L(\mathbf{s})\}$. At this point there are two distributions: the microstate distribution and an induced distribution over energy macrostates. Their thermodynamic structure is captured by the parametrized microstate distribution

$$p_\omega(\beta) = \frac{2^{-\beta U_\omega}}{Z_\beta(L)} \tag{36}$$

where β accentuates or attenuates a microstate's weight solely according to its energy. This is the same role (inverse) temperature plays in classical thermodynamics. The partition function

$$Z_\beta(L) = \sum_{\omega \in \mathrm{sub}_L(\mathbf{s})} 2^{-\beta U_\omega} = \sum_{\omega \in \mathrm{sub}_L(\mathbf{s})} p_\omega^\beta \tag{37}$$

gives the total signal space volume of the distribution $P_\beta(L) = \{p_\omega(\beta) : \omega \in \mathrm{sub}_L(\mathbf{s})\}$. In this way statistical mechanics explains thermodynamic properties as constraints on how this volume changes under various conditions.

From these definitions an extensive, system-size-dependent thermodynamics follows directly. For example, given an infinitely long data stream $\mathbf{s}$ the average total energy in all length L

* Going from individual measurements in a data stream to subsequences is a change in representation from the raw data to the parse tree, a hierarchical data structure.

sequences is

$$\mathrm{U}(L) = \sum_{\omega \in \mathrm{sub}_L(\mathrm{s})} p_\omega(\beta) U_\omega = Z_\beta^{-1} \sum_{\omega \in \mathrm{sub}_L(\mathrm{s})} U_\omega 2^{-\beta U_\omega}$$

$$\mathrm{U}(L) = -Z_\beta^{-1} \sum_{\omega \in \mathrm{sub}_L(\mathrm{s})} p_\omega(\beta) \log_2 p_\omega \tag{38}$$

And the thermodynamic entropy is given by

$$\mathrm{S}(L) = H\left(P_\beta(L)\right) = -\sum_{\omega \in \mathrm{sub}_L(\mathrm{s})} p_\omega(\beta) \log_2 p_\omega(\beta) \tag{39}$$

where $H\left(P_\beta(L)\right)$ is the Shannon information of the microstate distribution.

These are definitions of the extensive, L-dependent thermodynamic parameters for a closed system thermally coupled to its environment. The total energy U exists in several forms. The most important of which is the thermal energy TS, where S is the thermodynamic entropy and T is the temperature. The remaining "free" energy is that which is stored via a reversible process and is retrievable by one. For a closed and nonisolated system it is the Helmholtz free energy F. The fundamental equation expressing energy conservation is then

$$\mathrm{U} = \mathrm{F} + \mathrm{TS} \tag{40}$$

In modeling, an observer is put into contact with the process and attempts, by collecting measurements and estimating models, to come to "inferential" equilibrium by finding the optimal model. The above thermodynamics describes the situation where the information in the data stream exists in two forms. The first is that which is randomized and the second is that responsible for the deviation from equilibrium. The thermodynamic analog of the Helmholtz free energy is

$$\mathrm{F}(L) = -\beta^{-1} \log_2 Z_\beta(L) \tag{41}$$

It measures the amount of nonrandom information in the ensemble described by $P_\beta(L)$ at the given temperature β^{-1}.

There are three temperature limits of interest in which the preceding thermodynamics can be simply described.

1. Equilibrium, $\beta = 1$: The original subsequence distribution is recovered: $p_\omega = p_\omega(1)$ and $Z_1 = 1$. All of the information is "thermalized" $\mathrm{U}_1 = \beta^{-1}\mathrm{S}_1$ and the Helmholtz free energy vanishes $\mathrm{F}_{\beta=1} = 0$.
2. Infinite temperature, $\beta = 0$: All microstates are "excited" and are equally probable: $p_\omega(0) = Z_0^{-1}$, where the partition function is equal to the total number of microstates: $Z_0 = \|\mathrm{sub}_L(\mathrm{s})\|$. The effective signal space volume is largest in this limit. The average energy is just the sum of the microstate energies: $\mathrm{U}_0 = Z_0^{-1} \sum U_\omega$. The entropy simply depends of the multiplicity of microstates $\mathrm{S}_0 = \log_2 \|\mathrm{sub}_L(\mathrm{s})\|$. The free energy diverges.
3. Zero temperature, $\beta = \infty$: The least energetic, or most probable, microstate ω_∞ dominates: $p_\omega = \delta_{\omega\omega_\infty}$, the signal space volume is the smallest $Z_\infty = 0$, $\mathrm{U}_\infty = U_{\omega_\infty}$, and the entropy vanishes $\mathrm{S}_\infty = 0$.

The goal for an observer is to build a model that reproduces the observed data stream, including the probability structure of the latter. In thermodynamic terms, the model should minimize the Helmholtz free energy. This is what machine reconstruction produces: a stochastic automaton that is in inferential equilibrium with the given data. How it does this is described elsewhere.[16] The following will cover the basic methods for this, using them to investigate the thermodynamic structure of a machine's invariant subsequences and distributions.

Dividing each of the extensive quantities by the volume L yields thermodynamic densities. And upon taking the thermodynamic limit of the densities, the asymptotic growth rates of the extensive parameters are obtained. These growth rates are intensive. They can be directly computed from the reconstructed machine. In a sense the machine itself is an intensive thermodynamic object: the effective computational equations of motion.

To obtain the intensive thermodynamics from a given stochastic machine M with $T = \left\{T^{(s)} : s \in \mathbf{A}\right\}$, a new set $\left\{T_\beta^{(s)} : s \in \mathbf{A}\right\}$ of parametrized transition matrices are defined by

$$\left(T_\beta^{(s)}\right)_{vv'} = e^{-\beta I_{v \xrightarrow{s} v'}} \tag{42}$$

where $I_{v \xrightarrow{s} v'} = -\log_2 p_{v \xrightarrow{s} v'}$ is the information obtained on making the transition from state v to state v' on symbol s. Note that as the parameter β is varied the transition probabilities in the original machine are given different weights while the overall "shape" of the transitions is maintained. This is the intensive analog of the effect β has on the extensive distribution P_β above.

Most of the thermodynamic properties are determined by the parametrized connection matrix

$$T_\beta = \sum_{s \in \mathbf{A}} T_\beta^{(s)} \tag{43}$$

There are two quantities required from this matrix. Its principal eigenvalue

$$\lambda_\beta = \sup_{i=0,\ldots,\|\mathbf{V}\|-1} \left\{\lambda_i : {}_i\vec{p}_{\mathbf{V}} T_\beta = {}_i\vec{p}_{\mathbf{V}} \lambda_i\right\} \tag{44}$$

and the associated right eigenvector $\vec{r}_{\mathbf{V}}$

$$T_\beta \vec{r}_{\mathbf{V}} = \lambda_\beta \vec{r}_{\mathbf{V}} \tag{45}$$

Note, however, that T_β is not a stochastic matrix. In fact,

$$\sum_{v' \in \mathbf{V}} \left(T_\beta\right)_{vv'} \begin{cases} \geq 1 & \beta < 1 \\ = 1 & \beta = 1 \\ \leq 1 & \beta > 1 \end{cases} \tag{46}$$

It does not directly describe, for example, the probabilities of the subset of sequences that are associated with the relative transition weightings at the given β; except, of course, at $\beta = 1$.

There is, however, an "equilibrium" machine, whose stochastic connection matrix is denoted S_β, that produces the same sequences but with the relative weights given by T_β. Recall that the state of macroscopic equilibrium is determined by one of the variational principles:

1. At given total entropy, the equilibrium state minimizes the energy;
2. At given total energy, it maximizes the thermodynamic entropy.

Using the latter entropy representation, the equilibrium machine is that with maximal thermodynamic entropy subject to the constraints imposed by T_β. That is, all of the nonzero edge probabilities are allowed to vary. $\mathcal{S}_\beta$ describes the process over the allowed subsequences which are in thermodynamic equilibrium at the given temperature. It is found using Shannon's entropy maximization formula[35,40]

$$\mathcal{S}_\beta = \frac{D^{-1}\left(\vec{r}_{\mathbf{V}}\right) T_\beta D\left(\vec{r}_{\mathbf{V}}\right)}{\lambda_\beta} \tag{47}$$

where $D\left(\vec{v}\right)$ is a diagonal matrix with the components of $\vec{v}$ on the diagonal. Since this is a stochastic matrix its principal eigenvalue is unity. However, the associated left eigenvector $\vec{p}_{\mathbf{V}}$

$$\vec{p}_{\mathbf{V}} \mathcal{S}_\beta = \vec{p}_{\mathbf{V}} \tag{48}$$

when normalized in probability gives the asymptotic state distribution.

The entropy rate, as seen in a previous section, is

$$h_\mu\left(\mathcal{S}_\beta\right) = \sum_{v \in \mathbf{V}} p_v \sum_{v' \in \mathbf{V}} p_{v \to v'} \log_2 p_{v \to v'} \tag{49}$$

where $p_{v \to v'} = \left(\mathcal{S}_\beta\right)_{vv'}$. The β-complexities are given by

$$\begin{aligned} C_\beta &= -\sum_{v \in \mathbf{V}} p_v \log_2 p_v \\ C^e_\beta &= -\sum_{e \in \mathbf{E}} p_e \log_2 p_e \end{aligned} \tag{50}$$

where $p_{e=(v,v')} = p_v p_{v \to v'}$. The metric ($\beta = 1$) and topological ($\beta = 0$) quantities are directly recovered. That is,

$$\begin{aligned} h = h_0 \text{ and } C = C_0 \\ h_\mu = h_1 \text{ and } C_\mu = C_1 \end{aligned} \tag{51}$$

The relation

$$C^e_\beta = C_\beta + h_\mu\left(\mathcal{S}_\beta\right) \tag{52}$$

again constrains the entropy rate and the complexities.

Physically speaking $I_{v \to v'} = -\log_2 p_{v \to v'}$ plays the role of an interaction energy between two states and β is related to the inverse temperature. Although the same support, i.e. set of sequences and topological machine, exists at all temperatures, varying β accentuates the measure of the process $\mathcal{S}_\beta$ over different paths in the machine or, equivalently, over different subsets of sequences. One subset's weight changes relative to others as dictated by T_β's elements.

In the limit of long sequences the partition function's growth rate is governed by the maximal eigenvalue λ_β of the machine matrix T_β. That is,

$$Z_\beta(L) \underset{L \to \infty}{\propto} \lambda_\beta^L \tag{53}$$

The machine Helmholtz free energy density becomes

$$F = -\lim_{L\to\infty} \frac{1}{L}\beta^{-1} \log_2 Z_\beta(L)$$
$$F = -\beta^{-1} \log_2 \lambda_\beta \tag{54}$$

and the thermodynamic entropy is

$$\mathrm{S} = k_B h_\mu \left(\mathcal{S}_{\beta^{-1}}\right) \tag{55}$$

where k_B is Boltzmann's constant. Using the basic thermodynamic relation between these, the total energy density is then readily computed by noting

$$\mathrm{U} = \mathrm{F} + \mathrm{TS}$$
$$\mathrm{U} = \beta^{-1}\left(h_\mu\left(\mathcal{S}_\beta\right) - \log_2 \lambda_\beta\right) \tag{56}$$

using the identification $\beta^{-1} = k_B T$.

In the entropy representation, the function $\mathrm{S}(\mathrm{U})$, computed from Eq. (55) and Eq. (56), determines the thermodynamic "potential" along an arc $\mathcal{S}_\beta$ in the model space $\mathcal{M}$ of consistent stochastic machines. Consistent machines are those having the same set of allowed sequences as those observed in the data stream. At each fixed β the equilibrium machine is estimated via Eq. (47). Here equilibrium refers to a closed and isolated system specified by a fixed temperature and so a fixed average energy U. In contrast, the graph of $\mathrm{S}(\mathrm{U})$ concerns a closed, but nonisolated system in contact with an energy reservoir at temperature $\beta^{-1} = \partial \mathrm{U}/\partial \mathrm{S}$. It gives the entropies and energies for the family of machines $\mathcal{S}_\beta$. The equilibrium machine occurs at $\beta = 1$ where the free energy vanishes and all of the unconstrained information is "thermal" or randomized.

This is the thermodynamic analog of a cost function like that over model space $\mathcal{M}$ as shown in Figure 3. It is not the same, however, since (i) $\mathrm{S}(\mathrm{U})$ is computed in the thermodynamic limits of a long data stream and long sequence length and (ii) it represents two different optimizations, one at each temperature and the other over all temperatures. This is the view of statistical estimation developed in large deviation theory.[8,17] It suggests a rather different appreciation of the Sinai-Ruelle-Bowen thermodynamic formalism[6,33,39] for invariant measures of dynamical systems as a foundation for nonlinear modeling.

Independent of this modeling interpretation Eq. (55) and Eq. (56) give a direct way to study macroscopic properties of the sequences produced by a stochastic machine. In particular, the shape of $\mathrm{S}(\mathrm{U})$ determines the variation in entropy and energy within subsets of sequences that are invariant under the process. It indicates rather directly the range of likely fluctuations in observed sequences. Two examples will serve to illustrate these points.

Figures 9 and 10 show the "fluctuation" spectra, the thermodynamic entropy density $\mathrm{S}(\mathrm{U})$ versus the energy density U, for the Misiurewicz machines $M_{r'}^{\rightarrow}$ and $M_{r'}^{\leftarrow}$. Notice the rather large difference in character of the two spectra. This is another indication of the computational irreversibility of the underlying process. The topological entropies, found at the spectra's maxima $h = S(U : \beta = 0)$, and the metric entropies, found at the unity slope point on the curves $h_\mu = S(U : \beta = 1)$, are the same, despite this. The energy extremes $\mathrm{U}_{\min}$ and $\mathrm{U}_{\max}$, as well as the thermodynamic entropies $\mathrm{S}(\mathrm{U}_{\min})$ and $\mathrm{S}(\mathrm{U}_{\max})$ differ significantly due to the irreversibility.

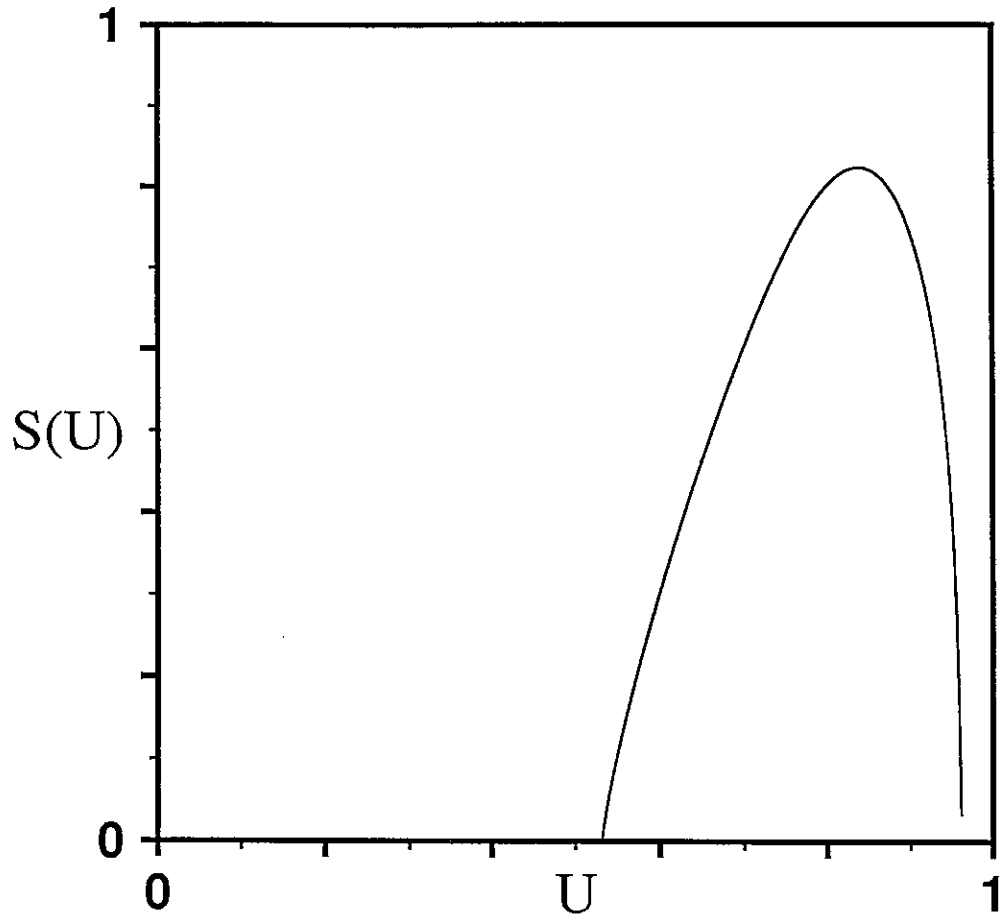

Figure 9 The fluctuation spectrum, thermodynamic entropy density S(U) versus internal energy density U, for the machine $M_{\vec{r},\vec{l}}$.

By way of ending this section, a final thermodynamic analogy will be mentioned. One of the first experimentally accessible measures of complexity was the excess entropy.[13] The total excess entropy $F_\beta(L)$ is a coarse measure of the average amount of memory in a measurement sequence above and beyond its randomized information. It is defined as follows

$$F_\beta(L) = H_\beta(L) - h_\beta L \tag{57}$$

where

$$H_\beta(L) = (1-\beta)^{-1} \log_2 Z_\beta(L) \tag{58}$$

is the total Renyi entropy and $h_\beta = (1-\beta)^{-1} \log_2 \lambda_\beta$ is the Renyi entropy rate. This was referred to as the free information[14] since it is easily seen to be analogous to a free energy. The free information is the Legendre transform of the Renyi entropy H_β that replaces its length dependence with an intensive parameter h_β. If subsequence length L is again associated with the volume V, a thermodynamic pressure can be associated with h_β. Finally, since the free information is an approximation of the finitary complexity,[14] the latter is also seen to be a type of free energy.

A more detailed development of machine thermodynamics can be found elsewhere.[16] The preceding outline hopefully serves to indicate a bit of its utility and interest.

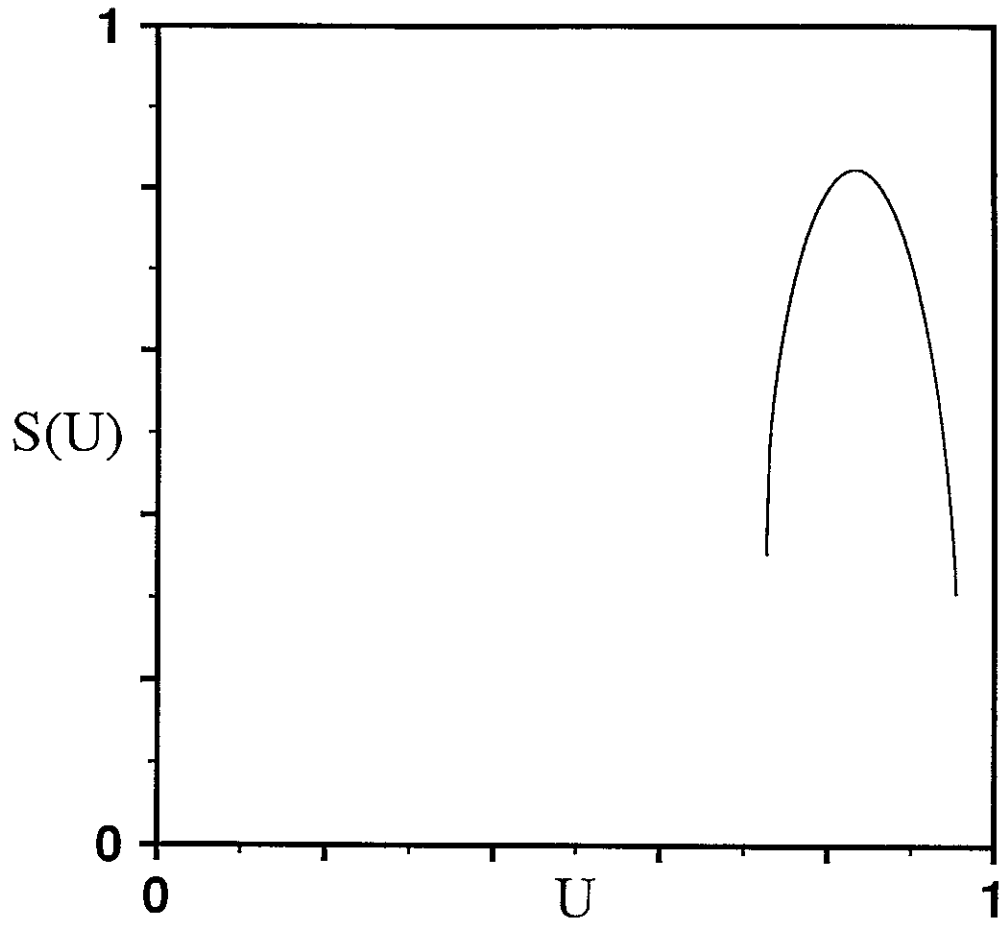

Figure 10 The fluctuation spectrum, thermodynamic entropy density S(U) versus internal energy density U, for the machine $M_{\epsilon}^{\leftarrow}$.

To summarize, the thermodynamic analysis suggests that the total information in a data stream, as extracted by machine reconstruction and as interpreted by an observer with a model, exists in two forms: one thermal and the other associated with information processing. The first, randomness in the data, is measured by the thermodynamic entropy. The second, structure in the data that causes it to deviate from simple thermal equilibrium, is that available to do mathematical work. This work might consist of communicating from a process to an observer; this is information transmission in space. It also can be available as static memory, which is information transmission across time. Most usefully, it can be available to do genuine computation and to support thereby semantic information processing.

SCIENCE AS DATA COMPRESSION?

Thinking back to the explanatory channel, these considerations lead me to disagree with the philosophical premise implicit in the universal coding theory approach to nonlinear modeling. While I accept the mathematics and use the optimization criteria, its own semantics appears wanting. Science is not data compression. The structure of models is ultimately more important than their use in encoders and decoders for the efficient encapsulation of experience. In the limit of large data streams and positive entropy processes, i.e. the realm of universal coding theory,

the model is essentially ignored and prediction error dominates. At the end of the day, though, good models are independent of the amount of data used to originally infer to them.* This point was emphasized in the preceding by the analysis of the effects their computational structure had on knowledge relaxation and on their semantic structure. Even these naked, mathematical objects, with which one typically does not associate meaning, do imply a semantic structure for the act of measurement. And it is this semantics that gives models their scientific value.

The preceding discussion, though only an outline, attempted to put these issues in a sufficiently large arena so that they can stand on their own. At the beginning there are dynamical systems whose diverse and complicated phenomenology has rapidly become better understood. They enrich our view of natural phenomena; though they do not necessarily deepen it. The contrast between their often simple specification and their creation of apparent complexity leads to computational mechanics. Computation theory in this development appears as the theory *par excellence* of structure. During the 1960's it gave the foundation for a theory of randomness. But that success should not blind us to the pressing need for constructive measures of complexity for physical, chemical, biological, and economic systems that go beyond randomness. Descriptions of complexity need not always pay for randomness. This is as true of statistical inference applied to nonlinear modeling as it is of thermodynamic and evolutionary systems. Indeed, it is one of the primary lessons of nonlinear dynamics that effective randomness is cheap and easily regenerated. Concomitantly, it also shows that ideal randomness is just that: an ideal that is expensive and, in principle, impossible to objectively obtain. Fortunately, nature does not seem to need it. Often only randomness effective for the task at hand is required.

This tension between randomness and order, the result of which is complexity, has always been a part of the problem domain of thermodynamics. Indeed, phase transitions and, especially, critical phenomena are the primary evidence of nature's delicate balance between them. Given this observation, the question now presents itself to nonlinear modeling, What types of computation are supported by physical systems at phase transitions, at the interface between order and chaos?† Away from "critical" processes, classical thermodynamics forms a solid basis on which to build nonlinear modeling. To the extent Gibbsian statistical mechanics is successful, so too will optimal modeling be. Though, as I just mentioned, there is much to question within this framework. Having described the analogy between thermodynamics and optimal modeling, another deeper problem suggests itself.

Classical thermodynamics foundered in its description of critical phenomena due to its confusion of the (observable) average value of the order parameter with its most likely value. So too the universal coding theoretic association of the optimal model with the most likely, Eq. (7), can fail for processes with either low entropy or near phase transitions. This will be especially exaggerated for "critical" processes that exhibit fluctuations on all scales. In these cases, fluctuations dominate behavior and averages need not be centered around the most likely value of an observable. This occurs for high complexity processes, such as those described by stochastic context-free and context-sensitive grammars,[15,30] since they have the requisite internal

* To describe the behavior of a thermodynamic system it suffices to communicate the equations of state, approximate macroscopic parameters, and possibly the force laws governing the microscopic constituents. Exact description is not only undesirable, but well nigh impossible.

† A first, constructive answer can be found elsewhere.[15]

computational capacity to cause the convergence of observable statistics to deviate from the Law of Large Numbers.

Having told this modeling story somewhat briefly, I hope it becomes at least a little clearer how the view of microscopic processes offered by statistical mechanics needs to be augmented. The examples analyzed demonstrate that the computational structure and semantic content of processes are almost entirely masked and cannot be articulated within the conventional framework. But it is exactly these properties that form the functional substrate of learning and evolutionary systems. The claim here is that an investigation of the intrinsic computation performed by dynamical systems is a prerequisite for understanding how physical systems might spontaneously take up the task of modeling their nonlinear environments. I believe the engineering by-products of this program for forecasting, design, and control, will follow naturally.

> The greatest sorcerer [writes Novalis memorably] would be the one who bewitched himself to the point of taking his own phantasmagorias for autonomous apparitions. Would not this be true of us?
>
> J. L Borges, "Avatars of the Tortoise", page 115.[5]

ACKNOWLEDGEMENTS

I would like to express my appreciation for discussions with Karl Young and Jim Hanson. Many thanks are due to the Santa Fe Institute, where the author was supported by a Robert Maxwell Foundation Visiting Professorship, for the warm hospitality during the writing of the present review. Funds from ONR contract N00014-90-J-1774, NASA-Ames University Interchange NCA2-488, and the AFOSR, also contributed to this work.

REFERENCES

1. H. Akaike. An objective use of bayesian models. *Ann. Inst. Statist. Math.*, 29A:9, 1977.
2. D. Angluin and C. H. Smith. Inductive inference: Theory and methods. *Comp. Surveys*, 15:237, 1983.
3. G. Birkhoff. *Lattice Theory*. American Mathematical Society, Providence, third edition, 1967.
4. R. E. Blahut. *Principles and Practice of Information Theory*. Addision-Wesley, 1987.
5. J. L. Borges. *Other Inquisitions 1937 - 1952*. Simon and Schuster, New York, 1964.
6. R. Bowen. *Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms*, volume 470 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1975.
7. A. A. Brudno. Entropy and the complexity of the trajectories of a dynamical system. *Trans. Moscow Math. Soc.*, 44:127, 1983.
8. J. A. Bucklew. *Large Deviation Techniques in Decision, Simulation, and Estimation*. Wiley-Interscience, New York, 1990.
9. G. Chaitin. On the length of programs for computing finite binary sequences. *J. ACM*, 13:145, 1966.
10. J. P. Crutchfield. *Noisy Chaos*. PhD thesis, University of California, Santa Cruz, 1983. published by University Microfilms Intl, Minnesota.
11. J. P. Crutchfield. Reconstructing language hierarchies. In H. A. Atmanspracher, editor, *Information Dynamics*, New York, 1990. Plenum.
12. J. P. Crutchfield and B. S. McNamara. Equations of motion from a data series. *Complex Systems*, 1:417, 1987.
13. J. P. Crutchfield and N. H. Packard. Symbolic dynamics of noisy chaos. *Physica*, 7D:201, 1983.
14. J. P. Crutchfield and K. Young. Inferring statistical complexity. *Phys. Rev. Let.*, 63:105, 1989.
15. J. P. Crutchfield and K. Young. Computation at the onset of chaos. In W. Zurek, editor, *Entropy, Complexity, and the Physics of Information*, volume VIII of *SFI Studies in the Sciences of Complexity*, page 223. Addison-Wesley, 1990.
16. J. P. Crutchfield and K. Young. ϵ-machine spectroscopy. preprint, 1991.
17. R. S. Ellis. *Entropy, Large Deviations, and Statistical Mechanics*, volume 271 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, New York, 1985.
18. A. Fraser. Chaotic data and model building. preprint, 1990.
19. P. Grassberger. Toward a quantitative theory of self-generated complexity. *Intl. J. Theo. Phys.*, 25:907, 1986.
20. J. Hartmanis and R. E. Stearns. *Algebraic Structure Theory of Sequential Machines*. Prentice-Hall, Englewood Cliffs, New Jersey, 1966.
21. J. E. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley, Reading, MA, 1979.
22. C. Howson and P. Urbach. *Scientific Reasoning: The Bayesian Approach*. Open Court, La Salle, Illinois, 1989.

23. E. T. Jaynes. Where do we stand on maximum entropy? In *Delaware Symposium in the Foundations of Physics*, volume 1, Berlin, 1967. Springer-Verlag.
24. J. G. Kemeny. The use of simplicity in induction. *Phil. Rev.*, 62:391, 1953.
25. B. Kitchens and S. Tuncel. Finitary measures for subshifts of finite type and sofic systems. *MSRI Journal*, page 1, 1984.
26. A. N. Kolmogorov. A new metric invariant of transient dynamical systems and automorphisms in lebesgue spaces. *Dokl. Akad. Nauk. SSSR*, 119:861, 1958. (Russian) Math. Rev. vol. 21, no. 2035a.
27. A. N. Kolmogorov. Three approaches to the concept of the amount of information. *Prob. Info. Trans.*, 1:1, 1965.
28. A. Lempel and J. Ziv. On the complexity of individual sequences. *IEEE Trans. Info. Th.*, IT-22:75, 1976.
29. M. Li and P. M. B. Vitanyi. Kolmogorov complexity and its applications. Technical Report CS-R8901, Centruum voor Wiskunde en Informatica, Universiteit van Amsterdam, 1989.
30. W. Li. Generating non-trivial long-range correlations and 1/f spectra by replication and mutation. preprint, 1990.
31. K. Lindgren and M. G. Nordahl. Complexity measures and cellular automata. *Complex Systems*, 2:409, 1988.
32. B. Marcus. Sofic systems and encoding data. *IEEE Transactions on Information Theory*, 31:366, 1985.
33. Y. Oono. Large deviation and statistical physics. *Prog. Theo. Phys.*, 99:165, 1989.
34. N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw. Geometry from a time series. *Phys. Rev. Let.*, 45:712, 1980.
35. W. Parry and S. Tuncel. *Classification Problems in Ergodic Theory*, volume 67 of *London Mathematical Society Lecture Notes Series*. Cambridge University Press, London, 1982.
36. H. Poincare. *Science and Hypothesis*. Dover Publications, New York, 1952.
37. L. R. Rabiner. A tutorial on hidden markov models and selected applications. *IEEE Proc.*, 77:257, 1989.
38. J. Rissanen. Stochastic complexity and modeling. *Ann. Statistics*, 14:1080, 1986.
39. D. Ruelle. *Thermodynamic Formalism*. Addision Wesley, Reading, MA, 1978.
40. C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, Champaign-Urbana, 1962.
41. C. S. Wallace and F. P. Freeman. Estimation and inference by compact coding. *J. R. Statist. Soc. B*, 49:240, 1987.
42. S. Wolfram. Computation theory of cellular automata. *Comm. Math. Phys.*, 96:15, 1984.
43. L. A. Zadeh. *Fuzzy sets and applications: selected papers*. Wiley, New York, 1987.
44. J. Ziv. Complexity and coherence of sequences. In J. K. Skwirzynski, editor, *The Impact of Processing Techniques on Communications*, page 35, Dordrecht, 1985. Nijhoff.
45. W. H. Zurek. Thermodynamic cost of computation, algorithmic complexity, and the information metric. preprint, 1989.