Order Number 9213359

The grammar and statistical mechanics of complex physical systems

Young, Karl Allen, Ph.D.

University of California, Santa Cruz, 1991

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

UNIVERSITY OF CALIFORNIA

SANTA CRUZ

THE GRAMMAR AND STATISTICAL MECHANICS OF COMPLEX PHYSICAL SYSTEMS

A dissertation submitted in partial satisfaction
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

PHYSICS

by

Karl Young

December 1991

The dissertation of Karl Young
is approved:

Dean of Graduate Studies and Research

Contents

List of Figures

Abstract

The Grammar and Statistical Mechanics of Complex Physical Systems

Karl Young

Constructive and computationally tractable techniques are developed for the classification of complex dynamical systems. These techniques subsume and extend standard dynamical characterizations based on Renyi dimensions and entropy spectra, invariant measures, algorithmic complexities and other dynamical and statistical measures. A natural way of quantifying the complexity of spatio-temporal data sets, and hence the combination of underlying physical system and measuring apparatus, is presented. These methods uniquely associate a data set with a formal language, given a specified model basis. The position of this language in a hierarchy of formal languages provides a measure of the complexity of the data set. This in turn provides a quantitative measure of the experimental or computational resources necessary for the complete characterization of a given system. Examples of data sets produced by various systems and models are analyzed using these techniques. It is shown that these techniques are particularly effective in detecting structure and elucidating underlying mechanisms for complex physical phenomena.

Acknowledgments

The text of this dissertation includes reprints of the following previously published material:

J. P. Crutchfield and K. Young. Inferring Statistical Complexity. Phys. Rev. Let., 63:105, 1989.

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.

The co-author listed in this publication directed and supervised the research which forms the basis for this dissertation.

Preface

This work would never have been possible without the support of a large number of people and I would like to thank them all. In particular the patience and support of my mother Nyda, my wife Sue, my son Miles, and my in-laws David and Randi Friedland was indispensable. For vital support and intellectual stimulation I thank Ralph Abraham, Bill Baird, James Crutchfield, David Doshay, Eric Friedman, James Hanson, Ken Hunter, Oliver Johns, Mary King, Jeff Royer, Jeff Scargle, Peter Scott, John Sidorowich, Ken Simpson, and Betty Young. Thanks to my friends at SF State, in particular Roger Bland, John Burke, Jerry Fisher, and Chris Hodges. I would like to thank Ralph Abraham, John Guckenheimer, Mike Nauenberg, and Steve Smale in whose classes I was able to first glimpse the beauty of dynamics. To James Crutchfield, Jeff Scargle, and Peter Scott I give heartfelt thanks for suffering through early drafts of this dissertation and offering invaluable editorial and technical advice. For financial support I thank James Crutchfield, Dave Donoho, Carson Jeffries, Jeff Scargle, NASA Ames, the Aspen Center for Physics, ONR (contracts N00014-86-K-0154 and N00014-90-J-1774), and AFOSR (contract AFOSR-91-0293). I would also like to thank the Santa Cruz Dynamical Systems Collective for providing much of the inspiration for this work.

PART I
Introductory Material

Chapter 1 Why Computational Mechanics ?

Section 1 Classification in Physics

Introduction

The subject of this thesis is the classification of the behavior of complex physical systems. Why bother trying to classify rather than just directly learning the basic laws that govern the behavior of a system's constituents? Perhaps the simplest answer is suggested in the context of thermodynamics by an example given by Callen.[12] When you go to the drugstore to buy a litre of alcohol it's probably a good thing that you don't have to specify the positions and momenta of 10^{23} or so atoms. To describe a litre of alcohol in a way that is "useful" certainly entails the specification of a vastly smaller set of numerical quantities.

Classification in thermodynamic terms requires the specification of just a few important quantities such as temperature, volume, and pressure to predict the way a system in equilibrium will behave on the average. A thermodynamic description thus classifies a tremendous number of microscopic systems according to their average behavior.

On the other hand consider a simple, classical deterministic system like a mass on a spring. In this case we would generally consider the problem of classification to be rather trivial and unimportant. To the degree that the system can in fact be regarded as classical and deterministic the specification of a few parameters like the mass, the spring constant and the initial position and velocity completely and uniquely specify a particular system and its behavior. Unlike the thermodynamic description no notion of averaging is required.

The necessity of developing a classification scheme seems to be linked to the complexity of the process under examination. In this section we shall attempt to outline what we mean by complexity as well as the relationship between complexity and classification. We shall discuss

the use of classification schemes in physics and in particular argue that a careful application of such schemes may clarify the concept of irreversibility in physics.

Recent results in nonlinear dynamics suggest that, in general, when trying to describe the behavior of an observed physical system it is hard to determine to what degree averaging is necessary to "explain" and predict the behavior of a system. In this thesis we propose quantitative methods for making just such a determination.

A direct outgrowth of the classification techniques proposed in this thesis is a method for determining the "appropriate coordinates" for the description of complex systems. In the case of a coupled linear system with a large number of degrees of freedom like a perfect crystal, a normal mode decomposition is straightforward. For strongly coupled nonlinear systems finding the equivalents of normal coordinates is generally impossible. We propose to "let the data" generate the appropriate representation in the spirit of [63]

Complexity and Classification

We will argue that "generic" physical systems exhibit complicated behavior which is not easily characterized by either simple stochastic models or by simple deterministic models. The attempt to classify a given system in terms of how complicated it's behavior is will lead to us to a characterization of the system's structure and a measure of it's intrinsic complexity.

In the case of equilibrium thermodynamics, as Boltzmann first showed, an underlying microscopic system of particles is required to exhibit "sufficiently random" behavior, i.e. completely uncorrelated motion of the constituents, for the concomitant thermodynamic behavior to be simple. This was the assumption of molecular chaos or "Stosszahl-Ansatz" that Boltzmann used to prove his H-Theorem.[73] In many cases a system is composed of too many constituents to specify its microscopic dynamics completely but is not random enough to use the assumption of molecular chaos. In other cases nonlinear coupling between the coordinates of a system with only a few constituents leads to complicated behavior. In both of these situations some sort of averaging is required for a description to be "useful". The averaging cannot be done as simply

as in the case of a system whose constituents undergo "completely random" interactions as is assumed for a system in thermodynamic equilibrium. This is the situation that occurs for most of the interesting systems studied in nonlinear dynamics.

In these more complicated cases a fully detailed description of the microscopic dynamics is usually unavailable. We replace this with an abstract description in terms of some set of quantities resembling the average quantities used in thermodynamics. The quantities we refer to as abstract are abstract with respect to the coordinates of the constituents of the system.[1]

In the thermodynamic limit for systems in equilibrium the law of large numbers and central limit theorem guarantee that fluctuations in average quantities are insignificant. In more general systems however, fluctuations about the average are often important. In fact the nature and importance of fluctuations is a heuristic measure of the complexity of a system as we shall see.

As a simple physical example of a complex system whose behavior we might wish to classify, consider Rayleigh-Bénard convection, one of the classic examples of the onset of turbulence. For this system a fluid is contained between two thermally conducting plates and the whole system is heated from below. This can also be considered as an extremely simplified model of the earth's atmosphere. If the temperature gradient between the upper and lower plates is small the steady state behavior of the system is such that the fluid conducts heat from the lower plate to the upper plate. This would certainly seem to qualify as simple behavior. As the temperature gradient is increased the fluid develops convection rolls, still simple but more complicated than the conducting state. As the temperature gradient is further increased the convection rolls develop wiggles and the spectrum of frequencies picks up new components. Finally, at a high enough temperature gradient the frequency spectrum is broadband and the fluid appears to exhibit completely random motion. We argue that the completely random state is simpler to describe than the highly complicated pattern of wiggling convection rolls that occurs before the onset of full turbulence. Taking a snapshot of the behavior of the system in these

[1] Ironically, in this context, thermodynamic quantities like pressure are considered abstract as compared to the position and velocity of molecular constituents.

different cases we note that it would be much easier to reproduce the overall characteristics of the random pattern than the pattern of wiggly convection rolls, although it would be virtually impossible to reproduce the exact pattern of the more random looking system. This is easy to understand in a statistical sense; there is a vastly larger number of systems with the same statistical characteristics as the random appearing states than there are for the more ordered states, i.e., the random states have a higher entropy considered as a set of states.

We will take complexity to be a measure of the difficulty involved in describing the behavior that a particular system might exhibit. The attempt to classify systems based on this measure will lead us to a characterization of the intrinsic structure of the system. By focusing on classification we emphasize the role that models of complex processes play in our scheme. We identify a process modulo the set of processes which are characterized by the same model with it's associated measure of complexity. This in turn allows for the hierarchical organization of processes based on the complexity of the associated models.

Complexity in Physics

Among other things, a classification based approach to the analysis of complex physical phenomena requires an unambiguous definition of the term *complexity*. The notion of complexity that will be used throughout the thesis will be defined carefully and precisely in the chapters on computational mechanics, i.e., parts II, III, IV, and V, but at this point we give a heuristic definition.

The intuitive basis for our measure of complexity is that periodic systems have a complexity that increases with the period of the process, spatial or temporal, and for stochastic systems, i.e. statistically described systems, the complexity decreases as a function of the randomness associated with the process. This definition is certainly consistent with the psychological impression of complexity we derive from sensory input. Rigid, low periodic visual or aural patterns strike us as simple, e.g., a checkerboard, as do completely random, white noise type patterns, e.g., the signal of your favorite television station after it "leaves the air". However

at the lowest level of description, i.e., independent of any model of a process, random patterns are harder to exactly describe. We usually think of complex patterns as those that are richly textured and consist of intricate sets of symmetries, and as a result are hard to generate. Many useful definitions, which roughly satisfy these intuitive criteria, have been suggested in the nonlinear dynamics, computation theory, and statistics literature.[4],[7],[18],[35],[64],[76],[84] For this reason we shall try to demonstrate the relationship of our measure of complexity to other proposed measures of complexity and randomness.[24]

An historical notion of a physical system's complexity is that it should roughly be a function of the number of fundamental entities, e.g. particles or excitations, of which the system is composed. This definition is insufficient in the present context. While most of the systems comprising the domain of condensed matter physics could certainly be considered complex under this definition, there are examples of systems composed of many bodies that do not exhibit complex behavior, e.g. a gas of billiard balls or a perfect crystal at low temperature. There are also examples of systems with few bodies that do exhibit complex behavior, e.g. three bodies under a central force and the zero bodies of the vacuum of quantum field theory. In physics the most useful analytical approach to modeling complex behavior approximately, has been to treat a system as fundamentally linear with couplings weak enough to be treated by the use of perturbation theory. Under such a scheme the above definition of complexity is basically independent of the nature of the coupling. For example in a system of coupled harmonic oscillators with weak nonlinearities the complexity of the behavior is certainly a monotonic function of the number of oscillators. But if the coupling cannot be treated in this manner the strength of the interactions at a particular scale place constraints on the behavior of the composite system. We shall argue that it is in part the nature of these constraints that leads to techniques for the classification of complex physical systems.[2]

---

[2] There are mathematical theories of generalized constraints, in particular the theory of Pfaffian forms.[13] Here we wish to develop more empirically based methods.

Irreversibility in Physics

A traditional problem in physics has been to understand the origin of macroscopic irreversibility when the microscopic laws of physics, e.g. Hamilton's equations of motion with the standard potentials, are invariant under time reversal.[3] One way of thinking about this problem is to consider a film crew of Maxwell demons who function on the scale of gas molecules (don't ask what their cameras are made of!). Any film these demons shoot, of the behavior of a gas at the molecular level, can be run in either direction without molecular scale viewers being able to detect whether the film is being run forwards or backwards. On the other hand consider a macroscopic film crew filming the same gas at the same time. Macroscopic viewers would be able to determine the asymmetry in the film directions easily. The problem is to explain this discrepancy.

A byproduct of the program of computational mechanics is the proposal for an outline of the solution to this problem. We shall argue that the need for an abstract definition of state, as is commonly used in thermodynamics when attempting to describe the behavior of complex systems, will lead to a simple context in which to pose the problem of irreversibility.

That an abstract description of state is necessary for generating irreversibility in macroscopic systems has long been a part of the lore of statistical mechanics. The abstract description that is commonly described in this context is that of a process that evolves on a coarse grained representation of a phase space.[29],[73],[50] The process is then automatically irreversible due to the many to one mapping of the "physical" states, i.e. points in phase space, onto the coarse grained or "observable" states , i.e., coarse grained bins in phase space. While this notion is intuitively appealing, the observable states seem to be defined in an arbitrary way, i.e., what

---

[3] It is, however, assumed by elementary particle physicists that some component of a complete theory of fundamental interactions must violate time reversal invariance by the CPT theorem and the fact that CP invariance is violated by the electro-weak interactions.

level of coarse graining yields observable states?[4] Furthermore no way of easily determining statistical criteria such as stationarity for processes defined over such sets of observable states is available. We shall define criteria, namely optimal predictability, for the determination of observable states. We argue that this is the minimal addition to the intrinsic dynamics of a system that is necessary for generating irreversibility via coarse graining in the above sense.

Irreversibility in computational mechanics is discussed in more detail in [18].

Conclusion

If, as has been argued above, we acknowledge the need of abstract descriptions for obtaining "useful" descriptions of complex systems, we are implicitly acknowledging the importance of classification. Any physical system that is describable by an abstract model is by definition a member of some class, that is, the class of systems describable by this model.

In the following sections we shall argue that there are techniques available for generating just such "useful" abstract descriptions of data sets produced by nonlinear systems. These techniques are adopted from ergodic theory, information theory, and computation theory. None of these disciplines, however, provide "off the shelf" techniques that are completely satisfactory in our context. It is the adaptation and extension of techniques developed in these areas that, it is felt, form an original contribution and serve as the basis of this thesis.

The basic techniques from ergodic theory, information theory, and computation theory that we begin with are rather standard in those disciplines but may not be familiar to physicists. Therefore we shall spend some time discussing them, including some brief historical comments in this introductory chapter.

The present work is to some extent conceptually rooted in a basic construct from computation theory known as the Chomsky hierarchy.[14] This hierarchy was an early attempt at complexity

---

[4] Quantum mechanics provides a natural coarse graining scale for phase space with partition elements having volume $\hbar^N$ where N represents the number of degrees of freedom, but here we are simply questioning the validity of coarse graining arguments made in the contest of classical statistical mechanics.

based classification of natural languages. Part of the inspiration for the use of these ideas can be found in Chomsky's statement "General linguistic theory can be viewed as a metatheory which is concerned with the problem of how to choose such a grammar in the case of each particular language on the basis of a finite corpus of sentences." We think in terms of having a finite corpus of sentences, i.e., a finite data set. This set consists of a finite number of sequences of symbols, produced by a nonlinear system whose grammar, or set of sentence generating rules, we wish to learn. The grammar we infer constitutes a "useful" description of the underlying structure of the nonlinear system and, in part, quantifies its complexity.

Computational mechanics generally focuses on an alternate representation of languages, usually referred to in the computation theory literature as automata or machines. Corresponding to any language and it's representation as a grammar there is also a machine representation. These machines can be thought of as functions which recognize the language. Given a particular "word" or sequence of symbols the function tells one whether or not this sequence is an element of the language. Alternatively one can think of a machine as something that produces legal sequences of the language.

Although useful, one of the limitations of this framework, as used in formal language theory, is that it does not provide for a statistical description of the "finite corpus of sentences". On the other hand ergodic theory, which does provide a statistical framework for describing sentences of symbols via symbolic dynamics, only does so for sets of "infinite sentences", i.e., infinite sequences. As we have argued, in the experimental analysis of nonlinear systems what is desired is a statistical description of finite sets of observed sequences.[5] Therefore an interpolation and extension of various techniques from computation and ergodic theories is found to be necessary in our context.

Some of the philosophical ramifications of taking a classification oriented approach to the analysis of data produced by nonlinear physical systems will be discussed in the final section of

[5] Hopefully one can also obtain a description of the asymptotic behavior of certain functions defined on these sequences as their length becomes large.

this chapter. Those of a pragmatic bent, only wishing to obtain an outline and some examples of computational mechanics may wish to simply read sections 4, 5, and 7 of this chapter, skim chapter 6 which gives a brief summary of computational mechanics, and read whichever of parts II, III, IV, and V are of interest. Those wishing a bit more detail and/or background in ergodic theory, information theory, computation theory, or thermodynamics might wish to read the above as well as the appropriate sections of chapters 2, 3, 4, and 5.

Section 2 Computational Mechanics as a Classification Scheme

Despite Rutherford's famous pronouncement that science falls into two categories, *physics* and *stamp collecting*, these two "categories" are not, nor should they always be considered to be, mutually exclusive, as we have argued above. In fact, to the degree that we acknowledge that physics is an empirical science, the lions share of activity in physics is stamp collecting, i.e., problems relating to the classification of observed phenomena. The implication of Rutherford's distinction is that there is a clear demarcation between understanding and classifying. We argue here that the clarity of this distinction is artificial and a result of the particular set of problems with which physicists have dealt. This would be of little concern if the set of problems physicists attempt to solve in the future remained in this set. Recent work, however, indicates that this is unlikely. In particular the attempt to analyze the behavior of complex systems with many degrees of freedom, and with more than weak coupling between the constituents, seems to indicate the existence of a different class of problem. Examples of such complex phenomena that have been of recent interest in physics and illustrate the point are spin glasses, plasmas, turbulent fluids and the QCD vacuum.

Despite roughly three hundred years of unprecedented success, it has become increasingly apparent in the last twenty years or so, that a naive reliance on the combined efficacy of reductionism and a vaguely defined sense of physical intuition, in the guise of appropriately perturbed linear theories, provides an inadequate foundation for the study of complex physical

systems.[85] That this is true in practice, has long been acknowledged by physicists working in condensed matter physics, and has resulted in the seminal work on the classification of condensed matter systems into universality classes via the use of renormalization group techniques.[69] That this is true in principle has been demonstrated most dramatically by recent work in dynamical systems theory. It is this basic fact that has led to a search for new techniques for the analysis of complex data sets. For example the powerful tools of spectral analysis, which in conjunction with perturbation theory have served as the basis of theoretical physics for much of the nineteenth and twentieth centuries simply fail to distinguish between physically distinct situations. This has been dramatically demonstrated in the difference between Landau's theory of the transition to turbulence in fluids [52] and that of Ruelle and Takens.[78] It is in this spirit that this thesis appeals to the techniques of ergodic theory and formal language theory in an attempt to develop new tools, which we feel are necessary for the optimal classification, prediction, and understanding of complex physical systems.

We shall specifically construct a parametrized spectrum of complexities which will serve as classification invariants for complex processes. Previous authors [60] have argued that the spectrum of dynamical entropies and dimensions serve as a complexity spectrum, but we shall argue that this is not the whole story and that the complexity spectrum we define contains information not contained in the dynamical entropies or dimensions.

Section 3 New Techniques for the Investigation of Complex Natural Phenomena

The phenomenon of universality, recently discovered by condensed matter and field theorists via the use of the renormalization group techniques, showed that a set of relatively simple rules of organization for composite systems at phase transitions can be generic. In fact the behavior can be characterized by a small set of parameters, i.e., critical exponents and scaling laws. This has led to a renewed interest in classification schemes for complex systems in physics. In parallel with these developments dynamical systems theorists have developed classification techniques

for the long term topological and statistical behavior of dynamical systems. These techniques are generally based on the determination of sets of local and global invariants. These allow for a description of more general asymptotic behavior than simple equilibria, i.e. fixed points and limit cycles, to include the type of bounded stochastic, or chaotic, motion so commonly observed in natural systems like coupled or damped driven oscillators. Similarly ergodic theorists have developed classification invariants that are useful in describing the asymptotic statistical behavior of complex systems. This has made the techniques of ergodic theory useful to physicists in ways that they haven't been since the work of Boltzmann. In fact a reevaluation of the goals of the founders of ergodic theory leads directly to a new theory of fluctuations. This theory is directly related to the statistical theory of large deviations [28] which provides a framework for the examination of "generalized" thermodynamic quantities, e.g. moments of distributions, at various scales of observation. These quantities can have highly nontrivial scaling behavior in multiplicative process such as energy transfer from large scales to smaller ones in fluids [47], the formation of clouds [81] and multiple high-energy scattering events [8]. This behavior will be examined for some archetypal multiplicative processes in part V.

Independent of developments in ergodic theory, workers in the foundations of mathematics and the theory of computation developed the theory of formal languages in an attempt to classify the complexity and computational capabilities of various formal systems. This work dates back to Hilbert's formalist program to put mathematics on a firm axiomatic basis. Though the work of Godel, Turing, Church and others [43] demonstrated the impossibility of completing Hilbert's program, and hence demonstrated the limits of axiomatization, their work laid the foundations of formal language theory. This allowed for the quantification of the notion of the difficulty of a particular task or equivalently the complexity of a model of computation. Chomsky [14] and others subsequently constructed hierarchical classification schemes for the complexity of computational models. Although Chomsky hoped, but was never able, to place a model of natural language in his hierarchy, we find his classification scheme particularly appropriate for

the comparison of models of physical systems.

Many have expressed the idea that Godel's undecidability results should have something important to say about the way scientists construct their theories. The most pronounced example is the statement, attributed to Wheeler, perhaps apocryphally, that Godel's theorem is too important to be left to the mathematicians. Here we take the opposite point of view, that while a beautiful result, the theorem itself is essentially irrelevant to the process of scientific theory construction.[6] The main reason that the theorem is irrelevant is that many of the theories that scientists have constructed, when placed in Chomsky's hierarchy, can be thought of as belonging to the simplest class of languages, called regular languages. A known result [43] states that learning an arbitrary regular language from particular instances of allowable sequences is undecidable, i.e., not capable of an algorithmic characterization. But almost by definition for scientists, learning the class of theories obtainable from data is a decidable proposition. Hence there is structure in the process of observing and categorizing observations, i.e., building theories, that significantly constrains the set of models or formal systems available to the empirical scientist, as opposed to the mathematician. It is precisely this structure that we attempt to utilize and quantify here.

As mentioned above, some of the subtleties involved in the joint use of techniques from ergodic theory and formal language theory derive from considering a way to relate the asymptotic quantities dealt with in ergodic theory to the finite quantities dealt with in formal language theory, i.e. those dealt with by computers and experimentalists. This type of problem has been discussed by various authors [77] under the rubric of general thermodynamic limits and many of the examples dealt with in this thesis can be thought of as special cases of attempts to determine how to obtain general thermodynamic limits from finite data sets.

We shall also introduce an abstract definition of the state of a physical system based on optimal prediction of future or adjacent spatio-temporal states. The use of abstract states in

---

[6] The techniques used in the proof, however, are just those that we find most useful.

the description of a system consists of acknowledging, as was discussed above, that while the behavior of a system may exhibit structure, the optimal way to represent this structure will not necessarily be obvious. This should come as no surprise to anyone familiar with the dilemma of the long and short lived kaon states. There the correct representation for understanding the structure was in terms of CP eigenstates and not the more obvious mass eigenstates. Similarly it may not have always been obvious that the simplest representation of the vibrational modes of a drumhead were in terms of functions that solved Bessel's differential equation. Consistent with these observations and the obvious fact that our available experimental resources, and not the dynamical states obtained directly from theory, determine the set of relevant observational states, we argue that a careful treatment of the general definition of such states is necessary, and shall be offered here. For a wonderfully careful and precise treatment of the notion of observation state very similar to that offered here see Penrose [66]. Part of the formal structure of computational mechanics is to automate the procedure of determining an optimal representation of observational states. We note that the conceptual basis we offer for the notion of observation state seems far less *ad hoc* than many current attempts to abstractly characterize this notion, e.g., via neural network models [44],[49],[54],[70].

Section 4 Putting it All Together

The unifying element in this broad spectrum of ideas is the simple notion that the spatio-temporal resolution of any measuring instrument is finite[7].[17] The discovery by dynamical systems theorists that information producing systems, i.e., systems with positive information theoretic entropy, are generic [38] leads us to a consistent, Bayesian[8] theory of stochastic modeling.[19]

---

[7] For theoretical purposes we could extend the resolution of a measuring instrument to being countably infinite with little modification of the theory.

[8] The finite nature of the measuring instrument allows us to develop a statistical model without assuming that there is anything fundamental about the probabilities derived in the model. Having given ourselves this justification we shall use frequentist techniques with abandon.

The basic technique we will introduce, and subsequently refer to as machine reconstruction, is based on the consideration of a series of spatio-temporal measurements, i.e., the output of some instrument. This output can be thought of as the result of imposing a partition on the state space of the system being studied, with bin sizes given by the finite resolution of the instrument. As an example consider again the case of Rayleigh-Bénard convection. Gollub, Benson [34] and others have experimentally identified various fluid instabilities by measuring the velocity of a fluid at various points in the fluid. As these measurements are of finite resolution, one obtains from a particular experiment a time series of finite symbols, corresponding to measurement bins, representing the fluid velocity at that point. In practice we shall see that for chaotic or information producing systems much can be learned from the use of a much smaller set of symbols than those given by the full resolution of the available experimental apparatus which in this case would lead to an enormous number of symbols.

In the simple case of a single time series the successive measurements are represented by a string of alphabet symbols and yield a symbolic dynamical system in the sense of [21]. The set of substrings of this symbolic dynamical system can be analyzed as a formal language and the frequency of occurrence of the various substrings allows for a statistical treatment of the formal language. This is a natural extension of the work of Shannon, Renyi and others in information and coding theory.[9] We shall utilize the statistical theory of large deviations to give bounds on the relative frequency of occurrence, and hence the relative importance, or "typicality", of various sets of substrings of a given length from the formal language representing a particular physical system. This is simply an extension of the notion of entropy in statistical mechanics. We emphasize again that a careful statement as to the assumed class of models we are using is crucial and somewhat novel, although not unprecedented in the analysis of physical systems.[19],[66],[76] As mentioned previously the definition of observational states, central to computational mechanics, depends critically on a careful statement of the model class being used.

That the program of computational mechanics is useful is the subject of this thesis. To

summarize, a set of tools with the properties necessary for classifying and predicting the spatio-temporal behavior of complex physical systems will be developed using elements from the theory of formal languages, ergodic theory, and information theory. In the next section we shall work through some simple examples to give a flavor of the techniques we will describe in detail later.

Section 5 Three Simple Examples

Introduction

In this section we provide examples of the implementation of computational mechanics on two particularly simple systems and a third, chaotic, example which is more interesting from our point of view. The first two examples, the biased coin and the harmonic oscillator, represent opposite ends of a spectrum regarding the utility of a statistical approach for the description of the systems behavior. For coin flipping the necessity of a statistical approach is obvious. Just as obvious is the lack of need of a statistical approach for the harmonic oscillator. There "complete deterministic information" about the system's behavior is given by a knowledge of the complete orthonormal set of solutions of the differential equation characterizing the oscillator, the initial conditions, and the values of a small number of parameters.

We argue that the most interesting systems in nonlinear physics are those for which a combination of deterministic and statistical description is the most efficient. We try to illustrate this with our third example, the logistic map at a particular chaotic parameter value. Computational mechanics is an attempt to develop a framework for just such descriptions.[24],[25],[23],[26]

The Biased Coin

We shall represent the occurrence of a head when we flip a coin by the symbol 1 and the occurrence of a tail by the symbol 0. Using a generic Unix™ random number generator, a simulation of two hundred tosses of a biased coin with $Pr(Heads) = 0.6$ and $Pr(Tails) = 0.4$,

yielded the sequence

S =0101001111011010110011000110010

1001011000111111010100100101111

0100011011111001111011000101110

1101111011100001111011011010111

0111011111010110011101110111010

1010111001111110000111111100010

10011111010110...

An accepted way of looking for complexity or order in some sequence is to look for correlations in the sequence. At least finding no correlations would indicate that the sequence is in some sense random.[9]. There are numerous ways that have been used for detecting correlations in sequences but we shall focus on two particularly simple ways. First we will look for forbidden subsequences and second we will calculate the frequencies of subsequences of a particular length. Provided that we have adequate statistics, a major assumption that is very hard to justify, forbidden sequences represent restrictions imposed by the dynamics of the system. Again, provided that we have adequate statistics, nonuniformities in the subsequence distribution are the result of dynamical effects.

To exhibit the characteristics of the set of subsequences associated with a given sequence we slide along the sequence with a window of given width D. We then organize the set of sequences observed in the sliding window on a binary tree as shown in figure 1. To obtain adequate statistics the length of this window obviously depends not only on the given process but on the length of the initial data sequence $N = ||\mathrm{S}||$. In our example, a window of length $D = 200$ would yield 1 subsequence out of a possible $2^{200} \simeq 1.6 \times 10^{60}$ subsequences. This is not very good statistics. In general there are $N - D + 1$ subsequences of a given sequence S.

[9] Technically a sequence generated by independent trials is considered to be the most random type of sequence. In this sense independence is a stronger property than being uncorrelated.[80]

For binary sequences, random ones like sequences of coin flips anyway, we expect to restrict D so that $N - D + 1 \gtrsim 2^D$.

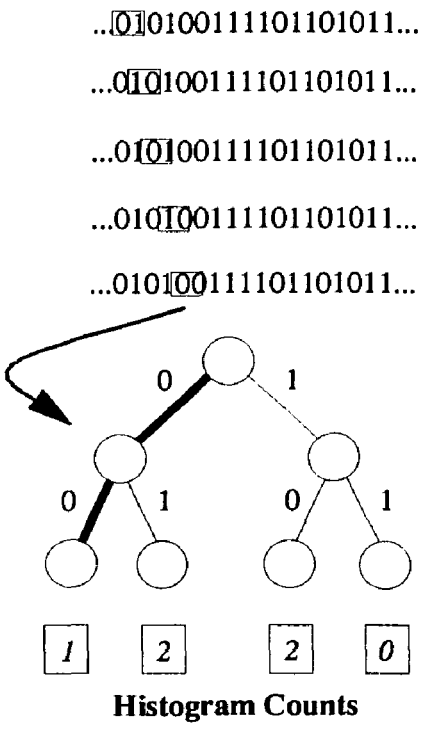


Figure 1 The first 18 symbols of the 200 symbol sequence S shown in the text are shown here. This sequence is "parsed" with a length $D = 2$ window. As the window slides along the sequence a binary tree is built by recording the occurrence of the various subsequences. Statistics are obtained by recording the number of occurrences of the particular subsequences as shown by the histogram count bins under each leaf of the tree in the figure.

In our case $N = 200$ so the maximum D satisfying the condition $N - D + 1 \gtrsim 2^D$ is $D = 7$ but that wouldn't give us very many counts per subsequence. Since we already know that the system we're trying to model is biased and hence the counts of sequences of straight tails will be relatively small we'll play it safe and choose $D = 4$.

In particular, using the "prior" knowledge that coin flipping represents a set of independent or *Bernoulli* trials we can calculate how many occurrences of the least likely sequence we expect, where that sequence is a string of four 0's,

$$\begin{aligned}
\#(0000) &= (N - D + 1) \times Pr(0000) \\
&= (197) \times \binom{4}{4} \times (Pr(0))^4 \times (Pr(1))^0 \\
&= 5.0432 \\
&\simeq 5
\end{aligned}$$

A similar calculation for $D = 7$ yields

$$\begin{aligned}
\#(0000000) &= (N - D + 1) \times Pr(0000000) \\
&= (194) \times \binom{7}{7} \times (Pr(0))^7 \times (Pr(1))^0 \\
&= 0.31875 \\
&\simeq 0
\end{aligned}$$

The binary tree for $D = 4$ is shown in figure 2 along with the frequency counts for the various sequences.

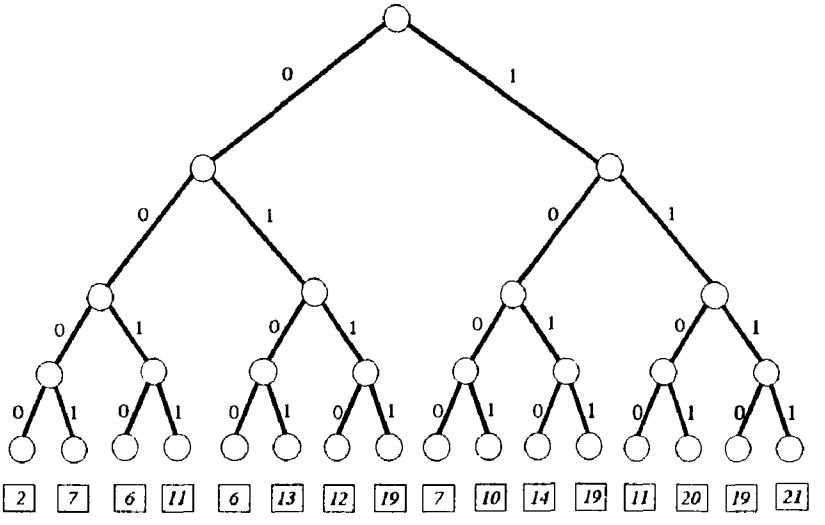


Figure 2 The binary tree with branches of length $D = 4$ for the simulated data stream given in the text, for a biased coin with $Pr(Heads) = 0.6$ and $Pr(Tails) = 0.4$. The histogram bin counts at the bottom give the number of occurrences of the binary string associated with that leaf of the tree.

One way to study correlations in the data stream S is to ask is how easy it is to predict the occurrence of future subsequences from a given node in the tree. More specifically what is the branching structure beneath any node in the tree? For example if there was a single path, with no branching beneath a given node in the tree, the set of possible future subsequences would consist of the single sequence labeling that path in the tree. Then our knowledge of the future, and hence ability to predict, would be completely correlated by the knowledge that we were at that node in the tree.

If in fact we had an infinite tree and the branching structure was identical beneath any two nodes for arbitrarily large subtrees, the future would look identical from these two nodes. We reiterate that by future what we mean here is the set of possible future subsequences. A more complete definition of identity is that the branching structure **and** the branching probabilities have to be identical. We shall say that any two tree nodes having this property belong to the

same class in terms of prediction. Formally these classes turn out to be equivalence classes as will be shown below.

To obtain these equivalence classes we group the nodes in the tree according to the branching structure beneath them. Since in practice we have only finite data sets and hence trees of finite depth D, we need a method for constructing approximate classes from finite data sets. First we choose the tree depth D based on the size of the data stream N as described above for the biased coin. Then we choose a *subtree* depth L that determines how far into the future we look from any tree node, i.e., L steps. This is clearly a parameter in our model as are N and D. If we have insufficient data, nodes that would be identical for a given process may have different branching structure due to poor statistics. Hence L, as well as D, must generally be chosen carefully.

If the branching structure beneath two tree nodes is identical to depth L then we say that these tree nodes belong to the same class at this level of approximation.[10]

For the biased coin example if we take $L = 2$ we can examine the branching structure beneath nodes of up to depth 2 in the tree of depth $D = 4$. For all 7 such nodes we observe identical branching structure, i.e., the fully branched depth $L = 2$ subtree as shown in figure 3. Therefore to this level of approximation, i.e., subtree depth $L = 2$ subtrees, all 7 nodes belong to a single class of future predictability.

---

[10] Note that, for a binary tree, there are 2^{2^L} possible subtrees of depth L.

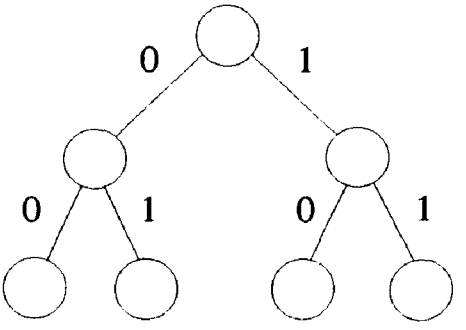


Figure 3 The single $L = 2$ subtree for the biased coin example.

The way we represent these classes is as *vertices* in a graph. We shall call the vertices *states* as they represent the various states (classes) of predictability given by the model obtained with our reconstruction technique.

For the biased coin all tree nodes belong to a single class and hence condense into a single vertex. To see how the vertices (classes) are connected we just look at the way the corresponding nodes are connected in the tree and draw an edge between any two vertices containing member nodes that are connected in the tree. Thus for every edge connecting vertices in the graph there are a set of representative node to node transitions in the tree. This is shown in figure 4 for the biased coin.

We label the edges of the graph with probabilities. The way we approximate these probabilities is to average over representative node to node transitions in the tree. We then use these relative frequencies as approximations to the branching probabilities of the corresponding edges in the graph.

For the biased coin there is only one class. Therefore we examine the branching probabilities beneath the root node to the nodes at level 1 in the tree. Since all nodes are in the same class these branchings represent edges that leave and return to the only vertex in the graph. The branch labeled by 0 represents the left half of the tree and has 76 counts beneath it. The

branch labeled by 1 represents the right half of the tree and has 121 counts beneath it. The total number of counts is $N - D + 1 = 200 - 4 + 1 = 121 + 76 = 197$ so we have for the approximate branching probabilities $Pr(state\ A \rightarrow state\ A\ on\ symbol\ 0) \approx \frac{76}{197} = 0.385787$ and $Pr(state\ A \rightarrow state\ A\ on\ symbol\ 1) \approx \frac{121}{197} = 0.614213$ as shown in figure 4. The double circle in figure 4 corresponds to the start state in the graph and is given by the class containing the root node of the tree. The A labeling the vertex refers to the single class of tree nodes for this example. The edge labels correspond to the symbol on which the transition between states occurs, along with the probability of that transition relative to the other transitions leaving the same state.

Figure 4 The graph representing the machine for the biased coin simulation discussed in the text. Edges are labeled by the appropriate symbol and probability of occurrence of the transition labeled by that symbol. The A labeling the vertex refers to the single class of tree nodes for this example. The double circle represents the start state in the graph.

The Harmonic Oscillator

Next we consider the case of the simple harmonic oscillator with a single degree of freedom. Recall that the equation of motion is

$$\ddot{x} = -m\omega^2 x$$

where m is the mass of the oscillator and ω is its angular frequency. Instead of a second order differential equation we can write this as two first order equations in the variables $x_1(t) = x(t)$ and $x_2(t) = \dot{x}(t)$,

$$\dot{x_1} = x_2$$

$$\dot{x_2} = -x_1$$

where we have set the mass and frequency to unity. In these units the solutions of the differential equations are represented by parametrized circles in the x_1, x_2 plane (phase plane) as shown in figure 5. The radius of the circle is specified by the energy of the oscillator, which is in turn given by the initial velocity, $x_2(0)$, and displacement, $x_1(0)$, from the equilibrium position.

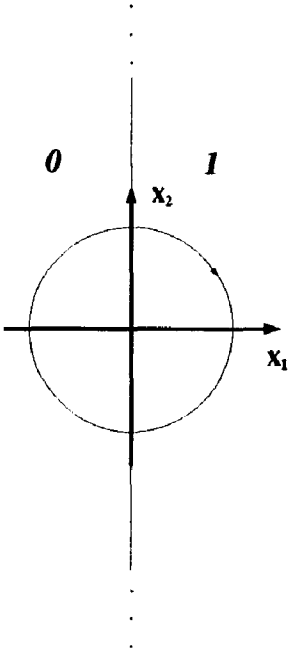


Figure 5 Phase plane of the harmonic oscillator with a particular solution. The phase plane is partioned into two partition elements, labeled 0 and 1, by the x_2-axis.

We construct a binary partition on the phase space of the oscillator by splitting the phase plane along the x_2-axis, i.e., for $x_1 = 0$. If we start the oscillator in the left half plane and "sample" the orbit at an interval $\tau = \frac{T}{2} = \frac{2\pi}{2\omega} = \frac{\pi}{\omega}$, recording which partition element the oscillator is in at each sample time, we obtain the following binary sequence[11]

$$\mathbf{S} = 0101010101010101010101010101010..$$

[11] Again, the effect of a particular choice of partition and sampling interval are obviously important issues. Here, however, we are simply attempting to demonstrate the techniques of machine reconstruction once these choices have been made.

If we construct a binary tree and a graph from this sequence in the same manner as we did in the previous example we obtain the $D = 4$ depth tree shown in figure 6.

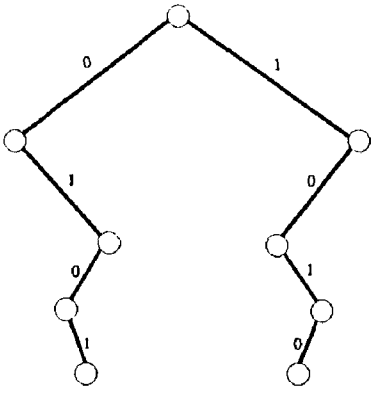


Figure 6 The binary tree for the linear harmonic oscillator with a binary partition.

For this tree we see that there are 3 different $L = 2$ subtrees and hence three classes of nodes in the tree. These are shown in figure 7. Note that the subtree with initial tree node as root has two branches. The fact that this subtree has branching corresponds to our initial ignorance of the phase of the oscillator. Once we have "observed" either a 0 or a 1 we have locked phases with the system and can predict with absolute certainty which sequence will follow. The two subtrees with no branching reflect this certainty and correspond to the two phases of the period 2 process.

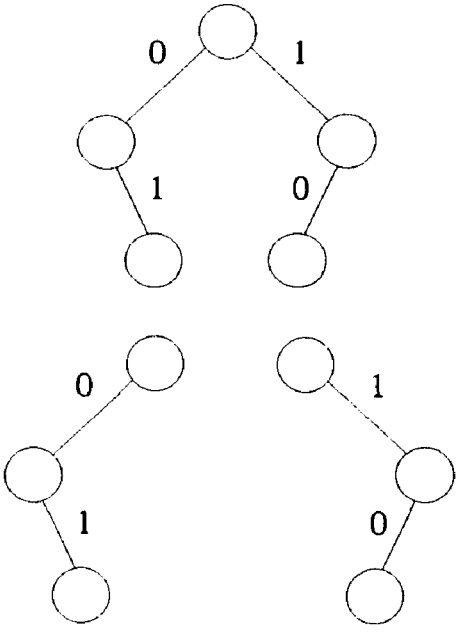


Figure 7 The three $L = 2$ subtrees for the harmonic oscillator with a binary partition.

The corresponding 3 vertex machine is shown in figure 8.

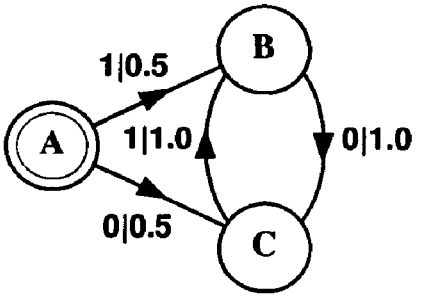


Figure 8 The machine for the linear harmonic oscillator with a binary partition. Recall that the graph vertex represented by a double circle denotes the start state

Notice in this example that there is a transient vertex (vertex A) in the graph. In this context transient means that if we follow an arrow leaving vertex A, the subsequent path obtained by

following arrows in the graph never returns to vertex A. In other words any path which leaves vertex A will subsequently only traverse vertices B and C, which constitute what we shall refer to as the recurrent part of the graph. Particularly important here is that there is no branching in the recurrent part of the graph. This should be expected in this case as we know that a simple deterministic description of the motion is available. The fact that vertex A has two branches leaving it represents our initial ignorance of the phase of the trajectory. Once we record a symbol, we know the phase of the oscillator and the subsequent behavior is deterministic and periodic, i.e. reflected by the lack of branching in the recurrent part of the graph. Note also that in our notation the vertex represented by a double circle denotes the start state.

Also note that the binary process given by the graph has period 2. Since our sampling rate was $\frac{T}{2}$, if we multiply the observed period by the sampling interval we recover the period of the observed process. This would also be the case if we used a four element partition, e.g., specified by cutting the plane into quadrants, and sampling at a rate of $\frac{T}{4}$.

Applying the techniques of machine reconstruction in this case gives us no new information. In fact using such a coarse grained partition of the state space gives us far less information than simply specifying the solution of the differential equation to some desired accuracy given a set of initial conditions and parameter values. We shall see in fact that the utility of machine reconstruction becomes most apparent when dealing with nonlinear systems which exhibit a phenomenon known as sensitive dependence on initial conditions (SDIC). For systems which exhibit SDIC, if the initial conditions of a trajectory are known only to finite accuracy, regardless of what that accuracy is, all information about the subsequent position of the phase point will eventually be lost. It is in such systems that a description balancing determinism and stochasticity will be found most useful.

The Logistic Map

Our final example is one that exhibits SDIC and is a more interesting example for illustrating the use of machine reconstruction. The process we obtain by applying the same techniques as

in the previous examples is stochastic, in that there is branching in the graph we obtain. The dynamics imposes structural constraints on the system, however. These constraints manifest themselves as nonbranching transitions in the graph.

The system we examine is a standard example from dynamical systems theory known as the logistic map. It was first introduced as a model of population growth [59] but has since been used to carefully study archetypal chaotic behavior, e.g., the period doubling route to chaos, that occurs in a large variety of physical systems. It is not our purpose here to discuss the logistic map in detail; for a nice review see [59].

The state space of the dynamics for the logistic map is the unit interval $x \in M = [0, 1]$. The dynamics on M is discrete in time and given by

$$x_{n+1} = F_r(x_n) = rx_n(1 - x_n)$$

where the parameter r specifies the height of the function $F_r(x)$. We partition the state space M into two pieces by splitting the unit interval at the midpoint $x_c = 0.5$ and label the left partition element with a 0 and the right one with a 1.[12] Figure 9 shows a schematic representation of $F_r(x)$, the partition on M, the graphical construction for obtaining the iterates of some initial condition x_0, and the symbol sequence obtained by labeling which partition elements the iterates fall in.

[12] The reason for choosing this particular partition is discussed in [17]

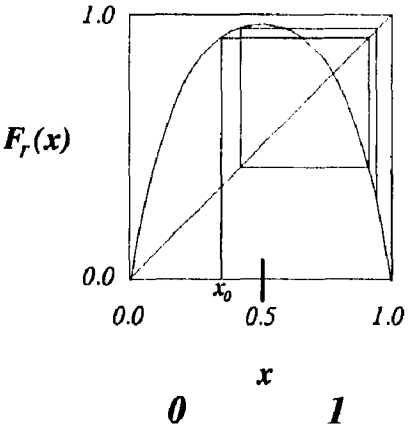


Figure 9 A schematic representation of the logistic map $F_r(x)$, the partition on M, the graphical construction for obtaining the iterates of x_0, and the symbol sequence obtained by labeling which partition elements the iterates fall in.

For our example we choose the r value $r \approx 3.9277370017867516$. This parameter value is known as a Misiurewicz parameter value and is frequently studied because it yields a chaotic system that has "nice" properties which make it easy to study.[13]. We obtain a binary sequence of 200 values by starting with the initial condition $x_0 = x_c = 0.5$, throwing away the first 300 iterations as transients and recording which partition element the next 200 iterations fall into. The sequence thus obtained is

S =10011100101011100111111010101100

10110101110010110011111100101001

11011011110111110101110011111111

00101111111010100111001011011100

11111001010101101101010010011110

01110110100101110011101010010001

01001001001111...

[13] This parameter value is found by requiring $F_r^1(x_c) = F_r^5(x_c)$

Obtaining the graph for this process, as will be discussed in parts IV and V, involves parsing the data stream with a longer window than one of length $D = 4$. For purposes of illustration, however, we show the tree that is reconstructed from the above data stream with a window of length $D = 4$ in figure above 10.

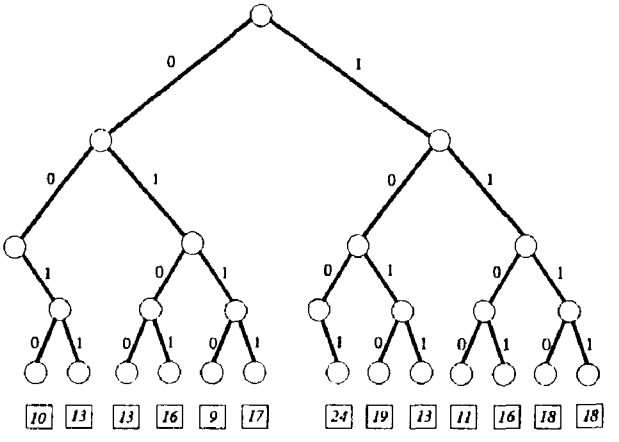


Figure 10 The binary tree for the logistic map at the parameter value $r \approx 3.9277370017867516$ with a binary partition on the unit interval.

We can obtain the graph for this process by going to a larger window size, and hence tree depth D. It shall be shown that any tree larger than the minimum depth tree to reconstruct the graph also yields the graph for the process. Remember, though, that our tree statistics are affected by the size of the initial data stream so we must exercise some care in the general case when choosing a tree and subtree size. For this example we obtained the graph in figure 11 by using a data stream of length $||S|| = N = 10^7$, a tree depth of $D = 32$, and a subtree depth of $L = 16$.

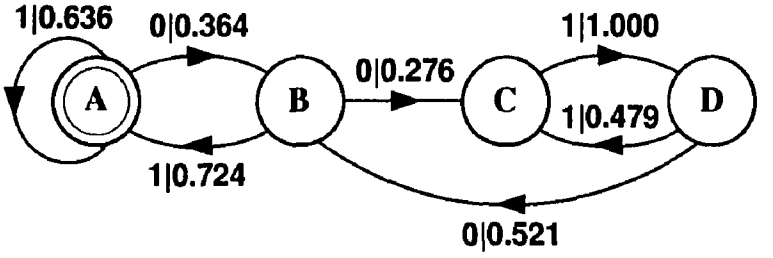


Figure 11 The graph for the logistic map at the parameter value $r \approx 3.9277370017867516$ discussed in the text with edges labeled by the appropriate symbol and probability of occurrence of that symbol.

Conclusion

It seems fairly intuitive that the larger the graph we construct using the above methods, the harder it is to represent the ensemble of possible sequences that the process can produce. The size of a reconstructed machine will in fact determine a first order measure of the complexity of a process modeled with this graph. For instance for an unbiased coin, yielding the smallest possible graph of one vertex as did our biased coin, the uncertainty of the result of the next flip is maximal but from the tree it is easy to construct the ensemble of possibilities. This ensemble is a uniform distribution on the set of all subsequences of a given length. We shall argue that quantities related to the size of such a graph representing a process are in fact useful measures of the complexity of that process.

Section 6 Computational Mechanics as Inference

After having discussed the scientific motivations for computational mechanics and provided some heuristic examples of it's main techniques, i.e., machine reconstruction, we briefly turn to some speculations about the relationship between computational mechanics and general questions regarding inductive inference.

Questions about the validity or efficacy of inductive inference constitute an area usually associated with the philosophy of science. Traditionally philosophers of science develop models, that they hope are consistent, for the process of obtaining scientific knowledge. Usually these models have little immediate effect on the methods of practicing scientists. We find, however, that philosophers debating the validity of inductive inference and scientists trying to learn the properties of complex systems have much in common and that advances in our understanding of either situation will help in understanding the other.

It is hoped that the classification techniques developed here will eventually lead to a consistent approach to the understanding of the basic elements of inductive inference, which, we would like to argue, underlie any attempt to classify and understand natural phenomenon. Despite the bad name induction has, perhaps justifiably, acquired among philosophers of science since Hume's devastating critiques, practicing scientists, most often naively, continue to rely upon it heavily in their day to day job of developing confidence in a given physical theory or predictive scheme. Philosophers of science might consider this as due to a lack of education. Here on the other hand we briefly explore why such ideas continue to be so stubbornly useful despite their apparent lack of a logical basis.

Traditionally clarification of the utility of inductive inference has been a rather academic exercise, rightly left to philosophers of science rather than practicing scientists. We feel however that the discovery and classification of patterns produced by complex systems entails just such a clarification. It must certainly be admitted nonetheless that Hume forever laid to rest the hope that the simplest model of induction, naive Baconian induction, could ever provide a useful and complete model of the growth of scientific knowledge. Hume's attack centered on the reliance of Bacon's theory on the ill defined notion of natural kinds. This idea might be simplistically characterized as the assumption that that elements of the world are classifiable in some straightforward and complete way. Current philosophical theories of the growth of scientific knowledge, steeped in Humean scepticism as they are, therefore seem to be an

inappropriate place to begin a search for an explanation of the efficacy of inductive inference and it's relationship to the understanding of complex physical phenomenon.

A prime example is Sir Karl Popper's theory, now known as naive falsificationism.[72] This theory states that to be a valid scientific theory a theory must make predictions that if shown not to be obeyed cause the theory to be rejected. This idea now seems so simple and obvious that it is hard to imagine that it was ever revolutionary. The problem is is that it is so far from describing the method in which science is actually practiced that it leaves one with a desire for a more detailed theory of the growth of scientific knowledge. It completely ignores the aforementioned role of inductive inference in building models of the world. The only place observation has in Popper's theory is in rejecting theories, which are regarded as free creations of the human mind. In this way Popper's theory can be considered deeply anti-inductive.[14]

A currently popular alternative to Popper's theory is Kuhn's theory of paradigm shifts [51] which is not only anti-inductive but also anti-rational. Kuhn deals with the problem that Popper's theory is not a good description of the practice of everyday scientists by acknowledging that theories are not generally rejected by scientists on the basis of a single piece of negative evidence as Popper suggests. A sort of "critical mass" of negative evidence must be amassed before the scientific community as a whole undergoes a "phase transition" , labeled a paradigm shift by Kuhn, to a new view of the world. Here again though, observation only plays a minor role in the determination of the models amongst which the scientific community shifts. Kuhn's theory is therefore also fundamentally anti-inductive. It is anti-rational as well in that there is no rational basis for determining what degree of negative evidence is necessary for a paradigm shift to occur. These theories and their variants therefore do not appear to provide a useful starting point for building an inductive model of the growth of scientific knowledge.

---

[14] Lakatos's revision, known as methodological falsificationism, which tries to modify Popper's theory to account for scientists stubborn use of induction still never seems to escape Popper's fundamentally anti-inductivist bias, treating as it does the inductive activity of practicing scientists as merely provisional.

We shall adopt, as a starting point for an inductivist based classification scheme a less rigorous, but at least intuitively more appealing Bayesian approach.[45],[79] The Bayesian approach essentially entails using whatever evidence is available to determine the degree of belief, in a particular model, that one is justified in assuming. The essential idea is to specify carefully the class of models from which one intends to infer the "correct" model of a set of data. We can thereby take advantage of whatever stationarity is evident in our environment, via the data, to pick an optimally predictive model from some class of models. But we usually want something more from a scientific theory than just predictability, e.g., we feel that there is something more satisfying about Newtonain mechanics than Ptolemaic epicycles although the latter was certainly a fine theory for celestial navigators to predict locations with. We argue in this thesis that a careful and systematic treatment of model determination is part of that something else that we expect from a scientific theory. It is felt that the constructive Bayesian approach advocated here seems more satisfying as a description of scientific activity than a Popperian reliance on the divine intuition of scientists.

We conclude this section with some comments about what a Bayesian approach entails and why we find it useful. Suppose we have some data S and a model of the process that produced that data M. A natural question we might ask in the context of inductive inference is what is the probability that given S, M is the best model of the system under study. To a "classical " statistician this is a meaningless question. All we can ask for are, for example, things like confidence intervals for a given M producing the data S. A Bayesian however would calculate the *posterior probability* $Pr(M|\mathrm{S})$ using Bayes' formula

$$Pr(M|\mathrm{S}) = \frac{\int Pr(\mathrm{S}|M')Pr(M')\delta(M - M')dM'}{\int Pr(\mathrm{S}|M')Pr(M')dM'}$$

and the *prior distribution* over models $Pr(M)$. It is the prior distribution that is the real bone of contention between classical and Bayesian statisticians. To a Bayesian the only honest way to state one's assumptions is by specifying a prior distribution. To fail to specify a prior distribution is to choose a uniform distribution implicitly.

A classicist would argue that the notion of a prior distribution is too vague to be useful. They would also argue that it is ill-defined in a frequency based theory of probability in that it does not represent the frequency of occurrence of some event from a well defined "collective" of events. It should be noted, however, that when a data set is "large enough" calculations done in the Bayesian framework agree with those done in the classical framework. These questions therefore are really only important as foundational issues in probability and in circumstances of having to make inferences with small or "incomplete" data sets, as is often the case in experimental nonlinear dynamics.

Our usage of the Bayesian framework is twofold, one legitimate (according to a Bayesian) the other perhaps not. Firstly we stress the initial choice of a model class in analyzing data produced by a nonlinear dynamical system. This choice provides (at least in part) a prior distribution over models in the chosen class. It is doubtful that a Bayesian would quibble with this. Secondly, a more questionable (but no less useful) technique we use is to construct "prior distributions" from previously obtained posterior distributions, reminiscent of the way the Chapman-Kolmogorov equation is used in statistical mechanics. While this certainly seems to be in keeping with the spirit of Bayesianism, in terms of using all available information to estimate a prior distribution it is most likely in violation of the letter of Bayesianism in not confining assumptions that go into choosing a prior distribution to the pre data gathering era.

Section 7 A Synopsis of the Thesis

In chapters 2, 3, and 4 of part I we give a brief outline of the elements of ergodic theory, information theory, and computation theory that provide the motivation and basis for the techniques developed in chapter 6 of part I and parts II, III, IV, and V. These are complex subjects with a rich and vast tradition; we shall provide only highly selective overviews of issues relevant to the subsequent discussion.

Chapter 5 of part I reviews some very basic elements of thermodynamics that are used in subsequent sections. In particular, notions from thermodynamics and the theory of phase transitions which allow us to demonstrate that the description of a system at a phase transition is most robustly characterized by measures derived from its stochastic formal language are discussed.

In chapter 6 of part I we summarize the relevant issues discussed in the chapters 2–5 of part I. We argue that the techniques introduced there provide necessary but insufficient tools for the construction of computational mechanics. We summarize what is needed in addition and then give a brief introduction to the formalism of computational mechanics.

Building on the techniques described in the previous chapters, parts II, III, IV, and V outline the basic program proposed in this thesis and contain reprints and preprints of the original research papers in which this program was first proposed.

Chapter 2 Elements of Ergodic Theory

Section 1 Introduction

Our review of ergodic theory will be rather cursory. For a more detailed treatment see [86], [16], [29], [53].

Consider the general class of deterministic dynamical systems described by

$$\vec{x}_{n+1} = \vec{F}_{\vec{\mu}}(\vec{x}_n)\,, \quad \vec{x}_0 \in \mathrm{M} \qquad \text{(discrete case)}$$

or

$$\dot{\vec{x}} = \vec{F}_{\vec{\mu}}(\vec{x})\,, \quad \vec{x}(0) \in \mathrm{M} \qquad \text{(continuous case)}$$

In both cases M is the m-dimensional space of states of the system. We shall be somewhat sloppy in this chapter and use the terms *state space* and *phase space* interchangeably as we do not need the complex geometrical apparatus for which such distinctions are important. $\vec{F}_{\vec{\mu}}$ is the *dynamic*, and $\vec{\mu}$ represents the set of *parameters* governing the dynamics. In the discrete case $\vec{x}_0 = (x_0^0, x_0^1, \ldots, x_0^{m-1})$ is the systems initial state and in the continuous case $\vec{x}(0) = (x^0(0), x^1(0), \ldots, x^{m-1}(0))$ is the systems initial state. For the rest of this chapter we shall also drop the vector notation since the context will determine the dimension of the relevant vector fields. We denote the transformation of the entire space M under the dynamic $\vec{F}_{\vec{\mu}}$ as $\{T_t(x) | t \in R\}$ in the continuous case and $\{T^n(x) | n \in Z\}$ in the discrete case or simply as T. Here R refers to the real numbers and Z to the integers. Whether we are discussing a continuous or discrete system will be clear from the context.

In this chapter we consider only the conservative or *Hamiltonian* systems of classical mechanics. The reason for this is that ergodic theory, which is based on a careful study of such systems, is a prototypical example of an attempt to develop a classification scheme for the description of regular and irregular motion. In fact irregular motion was first discovered in this

type of system by Hadamard in his study of the free motion of a particle on a surface of constant negative curvature and by Poincare in his work on celestial mechanics.[39],[71] Subsequent work by mathematicians and physicists yielded careful definitions of particular classification invariants. In particular we shall make extensive use of the Kolmogorov-Sinai entropy. We also note that the Kolmogorov-Arnold-Moser theorem, developed in the context of ergodic theory, demonstrates the existence of highly complicated systems which exhibit the type of mixture of regular and stochastic behavior upon which our definition of complexity is based.

To modern mathematicians ergodic theory is the study of the qualitative properties of actions of groups on spaces.[86] From this abstract vantage point it has been possible to prove many deep and far reaching theorems in various branches of mathematics from number theory to dynamical systems. But this abstract characterization has prevented the basic results of ergodic theory from being widely appreciated in the physics community where in fact ergodic theory was born. Boltzmann coined the term ergodic, an amalgamation of the Greek words ergon (work) and odos (path), to describe a hypothesis about the long term behavior of trajectories on constant energy surfaces of Hamiltonian flows arising in statistical mechanics. He was compelled to invent ergodic theory as a rigorous attempt to answer the physicists and mathematicians who vociferously attacked his claim to have demonstrated irreversibility in Hamiltonian systems via his H-Theorem. Boltzmann was using the H-Theorem to try to provide a microscopic basis for the second law of thermodynamics.

Consider the action $\{T_t(x)|t \in R\}$ that maps phase space into itself, specified by the n-degree of freedom Hamiltonian H, and given by Hamilton's equations

$$\dot{q}_i = \frac{\partial H}{\partial p_i} \quad , \quad \dot{p}_i = -\frac{\partial H}{\partial q_i} \quad , \quad i = 1, 2, ..., n$$

15

---

[15] This is certainly an example of the action of a group on a space. Here the group is specified by the Hamiltonian and it acts on phase space. That we have a group in this case is guaranteed by the facts that $T_t(T_s T_u) = (T_t T_s)T_u$, $T_0 = I$ (the identity), and $(T_t)^{-1} = T_{-t}$.

The phase space is $2n$ dimensional but since we shall only consider the case of time dependent Hamiltonians in this chapter, energy is conserved. With this constraint the motion of the system takes place on a $2n-1$ dimensional we refer to as the *energy surface.*

In the above notation

$$x_1 = q_1$$

$$x_2 = q_2$$

$$\vdots$$

$$x_n = q_n$$

$$x_{n+1} = p_1$$

$$x_{n+2} = p_2$$

$$\vdots$$

$$x_{2n} = p_n$$

and

$$\dot{x}_i = \left(\vec{F}_{\vec{H}}(\vec{x})\right)_i = \begin{cases} \frac{\partial H}{\partial x_{i+n}} & , \quad i = 1, 2, ..., n \\ -\frac{\partial H}{\partial x_{i-n}} & , \quad i = n+1, n+2, ..., 2n \end{cases}$$

Boltzmann's hypothesis was that each orbit of $\{T_t(x) | t \in R\}$ would cover the whole energy surface. The truth of this hypothesis was thought to be necessary in establishing the equality of time and phase space means which in turn is important in establishing many of the basic results in classical statistical mechanics, such as the H-Theorem. It turned out that the hypothesis as stated above is false. It was realized by Poincare and others that a more careful statement specifying the conditions under which the equality of time and phase space means holds, is necessary.

The properties a flow must satisfy so that time and phase space means of real valued functions of phase space are equal with respect to it, constitute what is now referred to as ergodicity. The first rigorous ergodic theorems, relying on the results of measure theory and establishing ergodicity as an important classification invariant, were proved in the early 1930's by Birkhoff and von Neumann.

In part as a result of this work it became apparent that general stochastic systems could be organized in a hierarchy of increasingly random behavior and characterized by various invariant properties. The types of system we shall discuss in terms of increasingly random behavior are called recurrent, ergodic, mixing, K, and Bernoulli.[53] The discrete version of the Bernoulli system, called the Bernoulli shift, is considered to be as random as a coin toss.[30]

An important development was Kolmogorov's introduction, in 1958, of entropy as a classification invariant. It was shown by Ornstein, in 1969, that entropy is in fact a complete invariant for Bernoulli shifts. We shall make much use of the various notions of entropy that first arose in the context of ergodic theory.

In addition to the statistical description of systems in terms of measure theoretic properties much recent work has focused on the topological properties of dynamical systems. Deep connections between these points of view have recently been established.[27] These connections serve to clarify the relationship between formal language theory, which we use to specify the topological properties of a system, and ergodic theory, which we use to specify it's asymptotic statistical properties. In what follows we give a brief heuristic description of the ergodic hierarchy.

Section 2 The Ergodic Hierarchy

Definitions

We now take a brief tour of the classical ergodic hierarchy. There are subtler gradations of the hierarchy, e.g., strong and weak mixing, than is necessary for us to discuss here. For

details see [86].

We note that any property that holds for a system in the hierarchy, e.g., equivalence of time and phase means, holds for systems higher in the hierarchy. This is important in interpreting systems further up the hierarchy as being more random.

We shall generally think of *phase space* or *state space* as a space M having the structure of a measure space $(\mathrm{M}, \mathfrak{B}, \mu)$ where $\mathfrak{B}$ is a σ-algebra of subsets of M. By a *σ-algebra* of subsets we mean a set of subsets $\mathfrak{B}$ of M obeying the following properties

$$i)\ \mathrm{M} \in \mathfrak{B}$$

$$ii)\ \text{if } B \in \mathfrak{B} \text{ then } \mathrm{M} - B \in \mathfrak{B}$$

$$iii)\ \text{if } B_n \in \mathfrak{B} \text{ for } n \geq 1 \text{ then } \bigcup_{n=1}^{\infty} B_n \in \mathfrak{B}$$

i.e., the entire set is a member as are complements and countable unions of members. σ-algebras are important in the definition of probability distributions on continuous spaces; the elements of the σ-algebra are the basic events for which probabilities are defined, i.e., they are the analogs of the subsets of single sample points for finite or countable stochastic processes such as coin flipping. μ is a probability measure on $(\mathrm{M}, \mathfrak{B})$ such that

$$\int_{\mathrm{M}} d\mu(x) = 1$$

Consistent with the above comment we note that the probability measure is defined on the set of elementary events, i.e., measurable sets, of the process.

A *measure preserving dynamical system* is a measure preserving map of the phase space into itself

$$T : \mathrm{M} \to \mathrm{M}$$

and by *measure preserving* we mean

$$\int_{T^{-1}\mathfrak{B}} d\mu(x) = \int_{\mathfrak{B}} d\mu(x)$$

or

$$\mu(T^{-1}\mathfrak{B}) = \mu(\mathfrak{B})$$

where $T^{-1}\mathfrak{B}$ is the set of points in M which are mapped by T into $\mathfrak{B}$. The n^{th} iterate of T will be denoted by T^n.

Recurrent Systems

The simplest notion of stochasticity in Hamiltonian systems is Poincare recurrence, which is a kind of continuous, temporal pigeon hole principle[16]. If a system is *recurrent* the trajectory returns to the neighborhood of a given point infinitely often. Poincare's theorem for Hamiltonian systems shows that recurrence is a simple consequence of area preserving motion in a bounded region. Hence recurrence only occurs if the motion on M is bounded. Poincare used the notion of recurrence to demonstrate that Boltzmann's H-Theorem is rigorously incorrect in the sense that a Hamiltonian system could return infinitesimally close to its initial condition infinitely often, thus violating any notion of simple irreversibility in Hamiltonian systems. Recurrence does not imply the next level of the hierarchy, ergodicity, because recurrent pieces of M can be disjoint which as we shall see cannot happen in ergodic systems.

Ergodic Systems

To discuss ergodicity we must first define time and space means of functions. The time mean of a square integrable, real valued function of phase space $f(x)$ is defined, in the discrete case

$$\widehat{f(x)} = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f(T^n x)$$

or in the continuous case

$$\widehat{f(x)} = \lim_{T\to\infty} \frac{\int_0^T f(T_t(x))dt}{\int_0^T dt}$$

[16] The pigeon hole principle says that if you have n pigeon holes and n+1 pigeons to put in the pigeon holes there will be at least one pigeon hole containing more than one pigeon.

In the following we shall just give the definitions for discrete systems.It can be shown [86] for "nice" dynamical systems that for "almost all" x, $\widehat{f(x)}$ is independent of x on a given orbit. The space mean of $f(x)$ is

$$\overline{f(x)} = \int_{\mathrm{M}} f(x)\, d\mu(x) \ ,$$

As stated above *ergodicity* means

$$\overline{f(x)} = \widehat{f(x)} \ ,$$

for almost all x.

An interesting alternative definition of ergodicity follows from the property that $\widehat{f(x)}$ is independent of the starting point x. From this and the fact that ergodicity implies the above property for any "nice" function we see that a system trajectory passes arbitrarily close to any phase point infinitely often, as expected. Thus, based on this property, M is said to be *indecomposable* or *metrically transitive*.

A classic example of a continuous ergodic dynamical system is given by rotation on a two dimensional torus with irrationally related frequencies given by

$$T: \quad \begin{aligned} x(t) &= 2\pi\omega_1 t \ + \ x(0)\,, \quad mod\, x = 1 \\ y(t) &= 2\pi\omega_2 t \ + \ y(0), \quad mod\, y = 1 \end{aligned}$$

and with $\frac{\omega_1}{\omega_2}$ irrational. Here M is a torus which we are representing in the above equations as the unit square with opposite sides identified, i.e.,

$$\mathrm{M} = (x, y) \ , \ 0 \leq x, y \leq 1$$

$$st \ \ (x, 0) \equiv (x, 1)$$

$$(0, y) \equiv (1, y)$$

A small blob of initial conditions winds densely around the torus, passing arbitrarily close to any point, without changing shape, as shown in figure 12.

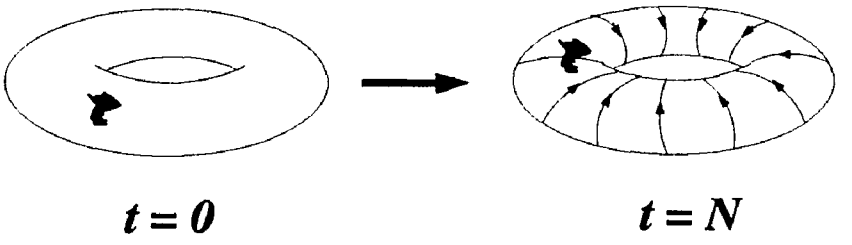


Figure 12 An initial blob of initial conditions winds around the torus. As t goes to infinity for irrational $\frac{\omega_1}{\omega_2}$ the windings densely cover the torus.

Mixing Systems

A flow is considered *mixing* if for all square integrable, real valued functions $f(x)$ and $g(x)$

$$\lim_{n\to\infty} \int_{\mathrm{M}} f(x)g(T^n x)d\mu(x) = \int_{\mathrm{M}} f(x)\, d\mu(x) \int_{\mathrm{M}} g(x)\, d\mu(x)$$

The primary importance of a system being mixing is that this guarantees that decay to equilibrium is attained while this is not the case for an ergodic system that is not mixing. A heuristic way to see this is to consider the special case

$$f(x) = g(x) = x$$

in the above expression. In this case we have

$$\lim_{n\to\infty} \int_{\mathrm{M}} x\, T^n x d\mu(x) = \left[\int_{\mathrm{M}} x\, d\mu(x)\right]^2$$

$$\lim_{n\to\infty} \langle xT^n x\rangle = \langle x\rangle^2$$

where we have used the notation

$$\int_{\mathrm{M}} f(x)d\mu(x) = \langle f(x)\rangle$$

for averages. Then the autocorrelation function

$$\langle (T^n x - \langle x\rangle)(x - \langle x\rangle)\rangle = \langle xT^n x\rangle - \langle x\rangle^2$$

using the above relation, decays to zero

$$\begin{aligned}\lim_{n\to\infty}\langle(T^n x - \langle x\rangle)(x - \langle x\rangle)\rangle &= \lim_{n\to\infty}\left(\langle xT^n x\rangle - \langle x\rangle^2\right)\\ &= \langle x\rangle^2 - \langle x\rangle^2\\ &= 0\end{aligned}$$

which implies that the system has relaxed to thermal equilibrium.[3]

As opposed to a system that is simply ergodic for which a blob of initial conditions retains its shape under iteration of the dynamics, for a mixing system a small blob of initial conditions spreads out into infinitesimally long, thin filaments that come arbitrarily close to any point in phase space. After a sufficient time, in any small region of phase space the ratio of points that have been iterated from the blob to points that have not is equal to the ratio of the initial volume of the blob to the volume of the entire phase space. This is suggested by the above expression in that the behavior of an iterate becomes completely decorrelated from its initial condition in the long time limit.

A subtlety, noted by Gibbs, is that the distribution of the blob only looks like an equilibrium distribution on a coarse grained level. Since for Hamiltonian systems phase fluid is conserved under the flow, the blob can never constitute a uniform distribution on phase space. The blob, non-blob fluid only appears uniformly mixed on a coarse grained scale.

A classic example is Arnold's cat map, a special case of the discrete systems known as toral automorphisms[17]. The cat map is given by

$$T: \begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} x_n \\ y_n \end{pmatrix} \mod 1$$

where M is a torus represented as the unit square as in the above case of rotation with irrationally related frequencies.

For a mixing system an initial blob of phase fluid under iteration of T becomes a stringy mass of filaments that wraps uniformly around the torus as shown in figure 13.

[17] The ubiquitous Axiom-A systems much discussed in the dynamics literature are in fact simple generalizations of the hyperbolic (read very random) toral automorphisms such as the cat map.[38]

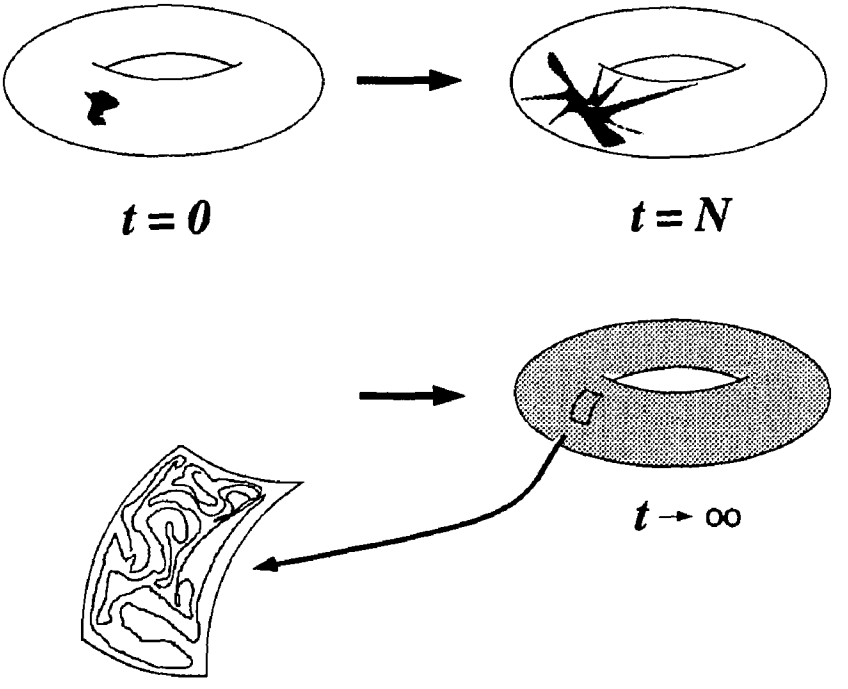


Figure 13 The blob of initial conditions stretches into a mass of stringy filaments that after a long enough time appear uniformly distributed on the torus. As shown in the blowup, on a microscopic scale the iterated points of the blob of initial conditions are distinguishable from the other points of the torus.

K-Systems

We introduce the next level of the hierarchy by noting that the spreading of the phase blob for a mixing systems need not be exponential, in the sense that the volume of the region containing phase fluid from the blob need not grow exponentially with time. When it does the system has what is referred to as positive Kolmogorov-Sinai or metric entropy (K-S entropy) and is reffered to as a *K-system.*

To define entropy consider a measure preserving dynamical system $(\mathrm{M}, \mathfrak{B}, \mu, T)$, e.g., Hamiltonian system as described above . Next we construct a finite state process, naturally associated with the dynamical system. This construction is important for the present purpose

of defining entropy but is also crucial in the reconstruction techniques central to this thesis. It explicitly connects the linguistic, or topological, properties of a dynamical system described in chapter 4 with the measure theoretic properties dealt with in ergodic theory.

The construction proceeds by considering a *partition* of M into a finite set of measurable subsets that cover M

$$P = \left\{ c_i : \bigcup_{i=0}^{k-1} c_i = \mathrm{M},\ c_i \cap c_j = \emptyset, i \neq j;\ i,j = 0,\ldots,k-1 \right\}$$

When we say c_i is measurable we simply mean that $c_i \in \mathfrak{B}$.

Note that any initial condition (technically "almost any" initial condition) x_0 is in a partition element. If we label the partition elements then for x_0 there exists a coded trajectory, where the n^{th} element of the coded trajectory is the partition label of the partition element that the trajectory is in on the n^{th} iterate of x_0 under T. We define the *entropy* of the partition P as

$$H(P) = -\sum_{i=0}^{k-1} \mu(c_i) \log_2 \mu(c_i)$$

where

$$\mu(c_i) = \int_{c_i} d\mu(x)$$

The motivation for this definition and its interpretation as a measure of information will be discussed in detail in chapter 3. Next we define the partition

$$T^{-1}P = \left\{ T^{-1}c_1, T^{-1}c_2, ..., T^{-1}c_{k-1} \right\}$$

where $T^{-1}c_i$ is the set of points in M that are mapped to c_i under T and hence $T^{-1}P$ is also a partition of M. Finally we define the partition $P \vee Q$ as

$$P \vee Q = \left\{ c_i \bigcap d_j \ :\ i = 1,...,k-1; j = 1,...,l-1;\ c_i \in P\ ,\ d_j \in Q \right\}$$

Then the entropy of T with respect to the partition P is

$$h(P,T) = h_\mu(P,T) = \lim_{L\to\infty} \frac{1}{L} H\left(P \vee T^{-1}P \vee T^{-2}P \vee \cdots \vee T^{-L+1}P\right)$$

If we truly wish to obtain a measure of the shuffling of points under T we do not want our measure to depend on the choice of partition so we define the *K-S entropy* of T as

$$h(T) = h_\mu(T) = \sup_P h(P,T)$$

As might be expected finding the supremum over partitions is in general a daunting task. In the special case that we can find a partition Q such that

$$h(Q,T) = \sup_P h(P,T)$$

we call Q a *generating partition* and we are done, in terms of calculating the metric entropy. Technically a generating partition is one that generates the σ-algebra $\mathfrak{B}$ of subsets of X in the following sense. Consider the set of partitions

$$P_{-i}^{j} = \bigvee_{l=-i}^{j} T^l P = \left(T^{-i}P \vee T^{-i+1}P \vee \cdots \vee T^j P\right)$$

A partition is said to be generating if

$$P_{-\infty}^{\infty} = \mathfrak{B}$$

Although abstract, this definition helps to describe the notion of a σ-algebra intuitively. We can think of the partition labels of P as representing the results of experiments performed on the dynamical system $(\mathrm{M}, \mathfrak{B}, \mu, T)$. Then the σ-algebra $\mathfrak{B}$ can be thought of as the set of possible measurement sequences that can be obtained, given a measuring apparatus P which defines a generating partition.

A theory of how to find a generating partition in the general case is currently lacking but there are special cases for which one can be found, e.g., the binary partition of the support of the logistic map.[21] It is the author's feeling that while rigorously determining that a partition is generating is an extremely hard mathematical problem it seems that one can do "well enough", in terms of determining relevant quantities like dimensions, entropies and Lyapunov

exponents, for many cases with a rather crude partition. An example of a good approximation to a generating partition for the dissipative Henon map is given in [36]. There an intuitively plausible "necessary" condition for a partition to be generating was given.[18] It may, however, be even easier to determine approximate generating partitions from experimental or simulated data sets than from the detailed consideration of stable and unstable manifolds in [36]. The search to define appropriate criteria for the existence and determination of generating partitions both from the structure of known dynamical systems and directly from data sets is currently an active field of research.[25],[36],[31]

The importance of the K-S, or metric entropy $h(T)$ is as a measure of our average uncertainty about where T moves the points of M. If $h(T) > 0$ we call $(\mathrm{M}, \mathfrak{B}, \mu, T)$ a K-system. $h(T)$ measures the average rate at which information about the state of the dynamical system is lost and hence is a measure of the randomness associated with the system. A system that is mixing but not K has $h(T) = 0$. For K-systems the volume of initial phase blobs grows exponentially. For non–K mixing systems, despite the fact that phase blobs eventually permeate the entire phase space, their volume does not grow exponentially.

The exponential growth of phase blobs suggests a relationship between the metric entropy and the *Lyapunov exponents* $\lambda_i(x)$ of $(\mathrm{M}, \mathfrak{B}, \mu, T)$. The $\lambda_i(x)$ are defined as

$$\lambda_i(x) = \lim_{n\to\infty} \frac{1}{n} \log\left(\Lambda_{n,i}(x)\right)$$

. $\Lambda_{n,i}(x)$ is the i^{th} eigenvalue of the matrix $\prod_{i=0}^{n} J\left(T^i x\right)$ with $J\left(T^i x\right)$ the Jacobian of the transformation $T^i x$. The Lyapunov exponents measure the average separation rate of nearby phase points under iteration of T. In fact Pesin [67] has shown that

$$h(T) = \int_{\mathrm{M}} \left(\sum_{\lambda_i(x)>0} \lambda_i(x) \right) d\mu(x)$$

[18] For those familiar with dynamical systems jargon, this criterion is that if a hyperbolic fixed point and a homoclinic point given by the transversal intersection of this fixed points stable and unstable manifolds exist, then either the fixed point or some preiterate of it and the homoclinic point or some iterate or preiterate of it are required to be in different partition elements.

where the sum in the integral is over positive Lyapunov exponents. This relationship underlines the interpretation of $h(T)$ as a measure of the average deformation of phase blobs.

Another interpretation of $h(T)$ that is important in applications, is that it is inversely proportional to the time interval over which the state of a chaotic system can be predicted.[82]

As an example we note that Arnold's cat map is K as well as mixing. We shall demonstrate that it has positive K-S entropy via an application of Pesin's relation. To use Pesin's relation we first calculate the Lyapunov exponents. Usually we must calculate the spectrum of the Jacobian of the linearized dynamics about some point to calculate the local Lyapunov spectrum about that point. In the case of the cat map our task is simpler in two ways. First the spreading rates and Lyapunov exponents are uniform throughout the phase space (torus), i.e., the slope is constant. Secondly the map is piecewise linear, the "nonlinearity" being introduced via the mod operator. Thus we need only calculate the global exponents specified by the linear operator

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$$

As can be found in many dynamics textbooks [38] the Lyapunov exponents are found in this case to be

$$\lambda_{1,2} = \ln \sigma_{1,2}$$

where $\sigma_{1,2}$ are the eigenvalues of A. Noting the global constancy of

$$\lambda_{1,2} = \ln \left(\frac{3 \pm \sqrt{5}}{2} \right)$$

we have by Pesin's relation

$$\begin{aligned} h(T) &= \int_{\mathbf{M}} \left(\sum_{\lambda_i(x)>0} \lambda_i(x) \right) d\mu(x) \\ &= \sum_{\lambda_i>0} \lambda_i \\ &= \lambda_1 \\ &= \ln \left(\frac{3+\sqrt{5}}{2} \right) \end{aligned}$$

The two values $\lambda_{1,2}$ correspond to the stretching and folding of regions under the cat map. λ_1 corresponds to exponential growth along the direction given by the associated eigenvector

$$\vec{\zeta_1} = \sqrt{\frac{2}{5-\sqrt{5}}}\left(\frac{\sqrt{5}-1}{2}\hat{x} + \hat{y}\right)$$

if we map the torus onto the unit square with doubly periodic boundary conditions. Similarly λ_2 corresponds to exponential decay along the direction

$$\vec{\zeta_2} = \sqrt{\frac{2}{5+\sqrt{5}}}\left(-\frac{\sqrt{5}+1}{2}\hat{x} + \hat{y}\right)$$

In this way it becomes clear how an initial phase blob is stretched into filaments and wrapped around the torus.

Markov Shifts

The cat map provides a nice way of describing the next level of the hierarchy, Markov shifts, as well. Adler and Weiss [1] in fact invented a special type of partition called a Markov partition to show that the K-S entropy is a complete invariant, i.e., a unique description, of two dimensional toral automorphisms like the cat map.

A *Markov partition* P of M with respect to T is a partition

$$P = \left\{ c_i : \bigcup_{i=0}^{k-1} c_i = \mathrm{M},\ c_i \cap c_j = \emptyset, i \neq j;\ i,j = 0,\ldots,k-1 \right\}$$

such that the pieces of the elements c_i of P are shuffled amongst each other as in figure 14. In particular for a Markov partition we require

$$c_i = \left(\bigcup_j T c_j\right) \bigcap c_i \ , \ \forall i$$

When we can define a Markov partition on M and since P is by definition finite in this case, we have defined a finite state Markov process with the elements of P as states as shown in figure 14. In general we have a Markov matrix which determines the one step probability of going from any partition element to any other. Since a Markov matrix has elements that are

either positive or zero we can use the Frobenius-Peron theorem [86] to find an eigenvector associated with the largest eigenvalue of the Markov matrix. The elements of this eigenvector give the asymptotic probability distribution over the elements of P. As indicated in figure 14 and described by Adler and Weiss [1] the general technique for building a Markov partition is to align the partition boundaries along the stretching and folding directions of the map.

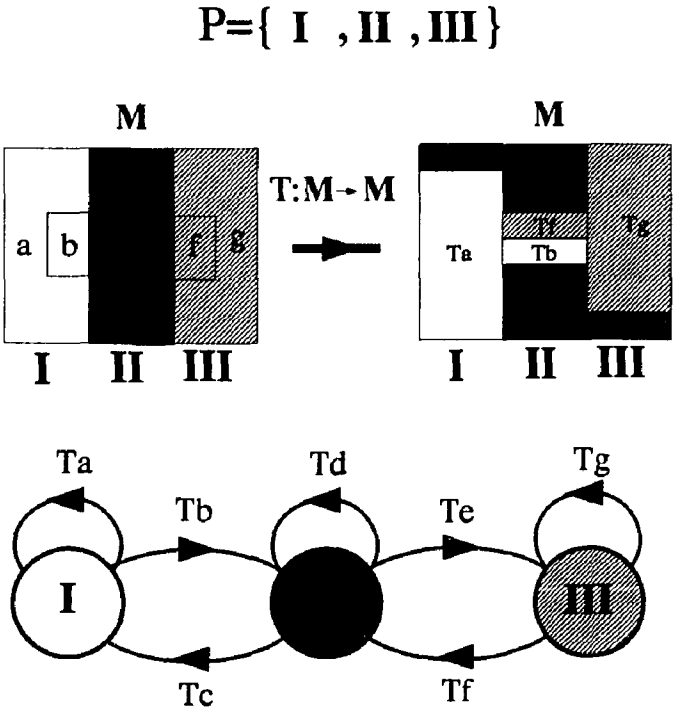


Figure 14 A state space M, a transformation T of M and a Markov partition P of M. Note that the definition of P depends on T as well as M. Below is the graph of the associated Markov process, where the branching probabilities are proportional to the areas of the pieces of the partition that are mapped into each other under T.

That this class of system, which includes the cat map, is more random than a general K-system is seen by the fact that the spreading of phase blobs over M is uniform. This was utilized in the case of the cat map to avoid the spatial integration when we used Pesin's relation to calculate $h(T)$. A system can therefore have $h(T) > 0$, indicating exponential spreading of phase blobs, without the spreading being uniform as in the case of Markov shifts.

Bernoulli Shifts

Finally we reach the pinnacle of the ergodic hierarchy, the systems that are described as being as random as a coin toss.[62],[30],[53] A Bernoulli shift is essentially just an alphabet of symbols such that the probability of observing any particular symbol is independent of the previous sequence of symbols. In the language of stochastic processes a Bernoulli shift is just an independent, identically distributed (i.i.d.) process on the given set of symbols. Examples are coin tosses (with two states), roulette (with the number of states given by the number of partition elements of the wheel) and successive positions in a random walk (with the number of states determined by the dimensionality of the underlying space)

In a measure theoretic sense this is not a huge departure from the class of Markov shifts as Ornstein showed that these classes are in fact isomorphic.[61] This is perhaps not surprising in that trajectories in both types of system can be described probabilistically by a product measure. In the case of Markov shifts this description is in terms of products of the asymptotic state probabilities defined above. A Bernoulli shift can be thought of as a Markov shift whose state transition matrix consists of n identical rows where n is the number of states. So from an asymptotic statistical point of view there is little difference between these Markov shifts and Bernoulli shifts. In fact Ornstein's proof that $h(T)$ is a complete invariant for all Bernoulli shifts has been extended to Markov shifts.[61] Put another way a Markov shift can be recoded, generally in terms of a larger alphabet, such that the new system is a Bernoulli shift.

We shall discuss finite time systems in what follows and hence wish to emphasize the distinction between Markov shifts and Bernoulli shifts perhaps more than an ergodic theorist would. From the point of view of identifying particular sequences there is a tremendous difference between these two classes of process. We shall see that the most natural description of a process in terms of observation states is as a Markov shift on a set of observation states. Penrose [66] argues convincingly that such a description is just what is necessary to provide an observational basis for statistical mechanics.

We have outlined the ergodic hierarchy in an attempt to describe an archetypal classification scheme, here as a way of directly quantifying the notion of randomness. We turn next to recent developments in ergodic theory that will help us make the connection with the material of chapter 4.

Section 3 Symbolic Dynamics, Subshifts of Finite Type, and Sofic Systems

We now take a slightly more abstract point of view and directly consider dynamics on sets of symbols. Underlying this point of view in terms of applications in physics and dynamical systems, is the idea that the set of symbols serve as the labels of partition elements for a coarse graining of some state space. But as has been amply documented [2], [38] much can be learned about a system simply from its representation as a symbol set and a dynamics on the symbol set. There are also many interesting applications of symbolic dynamics in the coding and information theory of discrete channels and sources as well as to the spectral theory of nonnegative matrices.[68]

In many cases a one to one, or almost one to one, correspondence between the trajectories of a dynamical system and the symbol sequences of a symbolic dynamical system can be demonstrated. We shall assume this to be the case in this section. One of the most important examples of this occurs for hyperbolic dynamical systems, i.e., systems having no nonhyperbolic fixed points. If a system with positive metric entropy is hyperbolic it has homoclinic points, i.e., points of transversal intersection between the stable and unstable manifolds of a hyperbolic fixed point. If it has homoclinic points there are horseshoes, i.e., regions that are stretched and folded back onto themselves, embedded in it. Finally it has been proven that there is a one to one mapping from the trajectories of a horseshoe to the full shift on two symbols, i.e., the topological equivalent of the Bernoulli shift.[38] By topological equivalent we simply mean that for every binary sequence there is a trajectory and vice versa where we are ignoring the probabilities of symbols.

Another common example is the mapping of the trajectories of unimodal one dimensional maps onto symbol sequences of symbols representing the partition elements of the known generating partitions in these cases.[15], [21] Perhaps more important than the consideration of the ideal mathematical case of Axiom-A systems or one dimensional unimodal maps is that generic damped, driven systems exhibit behavior very much like the ideal cases in which horseshoes are observed. This suggests that symbolic dynamic techniques are quite generally useful.[5]

Although important in a broad sense, considerations regarding one to one mappings of trajectories onto symbol sequences, in situations for which the dynamics is known, are not the main reasons we discuss symbolic dynamics. Our primary motivation is that symbolic dynamics is simply the most appropriate representation for the reconstruction of discrete dynamical systems from the necessarily discrete output of measurement channels.

The symbolic part of symbolic dynamics simply refers to the fact that we are working with a finite alphabet of symbols A. In fact, in symbolic dynamics, we generally work with infinite strings of symbols from A. The dynamics part of symbolic dynamics consists simply of the operation of shifting an infinite symbol sequence one place to the left

$$\sigma(x) = y$$

where

$$y_i = x_{i+1} \quad \forall i \in Z$$

Z is the integers and x and y are infinite strings of symbols from A. σ is reffered to as the shift map. With these definitions in place we denote the correspondence between the trajectories of a dynamical system $(\mathrm{M}, \mathfrak{B}, \mu, T)$ and the symbol sequences of a symbolic dynamical system as

$$\begin{array}{ccc} \mathrm{M} & \xrightarrow{T} & \mathrm{M} \\ \pi \downarrow & & \downarrow \pi \\ \mathrm{A}^* & \underset{\sigma}{\rightarrow} & \mathrm{A}^* \end{array}$$

where π maps a point of M into the sequence of labels of partition elements that it appears in under repeated iteration of T. We generally consider T^{-1} as well, for maps onto bi-infinite strings in $\mathbf{A}^*$ where

$$\mathbf{A}^* = \prod_{i=-\infty}^{\infty} (\mathbf{A})_i$$

The above commuting diagram simply states that projecting a point of M onto $\mathbf{A}^*$ and shifting the infinite sequence thus obtained, one unit to the left, yields the same infinite sequence as first iterating that point under T then projecting onto $\mathbf{A}^*$. Henceforth we shall simply concern ourselves with the symbolic dynamics.

The simplest type of symbolic system we can define is the topological sibling of the Bernoulli shift defined in the last section. In fact if we have a set of symbols A then the full shift is the set of all sequences of symbols from A

$$\mathbf{A}^* = \prod_{i=-\infty}^{\infty} (\mathbf{A})_i$$

and the shift operator σ which shifts sequences one place to the right.

In general we would like to consider subshifts, i.e., subsets of the set of all infinite sequences. First we define a block as a finite sequence of symbols. A *subshift* Λ is defined as a subset of $\mathbf{A}^*$ such that there exists a set of blocks of symbols $\mathcal{F}$ such that $x \in \Lambda$ if and only if x contains no blocks in $\mathcal{F}$. Two systems we shall encounter frequently, called the golden mean and even systems can be characterized as subshifts on a binary alphabet. In the case of the golden mean system $\mathcal{F}$ consists of one block, the sequence 00. In the case of the even system all blocks containing odd runs of $1's$ sandwiched by $0's$ are excluded and cannot be built from smaller members of this same set, hence $\mathcal{F}$ is infinite for the even system. Therein lies a crucial property of the subset of subshifts that contains the even system and which will be discussed below.

An important property of subshifts is that they are shift invariant, i.e.,

$$\forall x \in \Lambda \, , \, \sigma(x) \in \Lambda$$

This simple fact often allows us to recognize subsets of the full shift which are not subshifts. A nice example, given by Marcus [58], is the subset of the full 2-shift consisting of all the binary sequences containing only one 1. Assume this set is a subshift. Since these sequences contain arbitrarily long finite sequences of $0's$ no block of $0's$ can be forbidden. Therefore $0^\infty = ...000.000...$ must belong to the set. But clearly, by definition, it doesn't, therefore we have a contradiction and this set is not a subshift.

To further emphasize the connection between symbolic dynamics and formal language theory which is discussed in chapter 4, we next define the language associated with a shift. First we denote the set of all blocks of length L which appear in elements of a given subshift Λ, as $\mathbf{B}_L(\mathbf{\Lambda})$. Then the language associated with a given shift is

$$\mathcal{L}_\Lambda = \bigcup_{L=1}^{\infty} \mathbf{B}_L(\mathbf{\Lambda})$$

We note that as defined, the class of subshifts is a very large class, in fact an uncountably large class. It constitutes a quite general characterization of languages. In formal language theory the class equivalent to the class of subshifts is the most general class of languages, called the recursively enumerable languages, and is discussed in chapter 4. In what follows we shall discuss severely restricted classes, namely we shall discuss the subset of subshifts associated with the simplest class of languages, called the regular languages. We speculate, however, that as the techniques of computational mechanics become more sophisticated and able to classify more complex systems, more sophisticated symbolic dynamic analyses, in terms of more complex sets of subshifts, will also become possible.

As promised, we now restrict our attention to simple classes of subshifts. We begin by defining the sibling of the Markov shifts from the last section. We refer to these as topological Markov shifts or subshifts of finite type (SFT). These are also closely related to the discrete noiseless channels defined by Shannon in information theory.[83] The simplest characterization of an SFT is simply that it has a finite list of forbidden blocks $\mathcal{F}$, the canonical example being

the golden mean system with $\mathcal{F} = \{00\}$. As we saw above $\mathcal{F}$ is not finite for the even system and hence it is not a SFT.

Another characterization of SFT, introduced by Parry [65] is the reason for the name topological Markov shift. This characterization is given by any JxJ matrix, T_0, having only 0 or 1 for elements, where $J = |\mathbf{A}|$.[19] This defines a shift invariant subset Λ of $\mathbf{A}^*$ given by

$$\Lambda = \{y = (y)_i : (T_0)_{ii+1} = 1 \ \forall i\}$$

T_0 determines which symbols of $\mathbf{A}$ are allowed to follow each other. For the golden mean system as an SFT on two symbols we have

$$T_0 = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$$

These shifts are reffered to as subshifts of finite type, because of the topological analogy with the probabilistic Markov property of finite memory. That is, the set of symbols that it is possible to observe on the next shift is completely determined by the previously observed symbol. Thinking in terms of machines, we can test membership in $\mathcal{L}_\Lambda$ with a machine that requires only a finite amount of memory. Consider a tape containing some symbol sequence y. Feed this tape into our machine, symbol by symbol to test membership in $\mathcal{L}_\Lambda$. The machine need only remember the last symbol, i.e., have one memory register, to tell whether or not the next symbol is legal given the last symbol read.

An easy way to visualize a subshift of finite type is as a graph with the elements of $\mathbf{A}$ as vertices, and an edge between vertices $a_i, a_j \in \mathbf{A}$ if and only if $(T_0)_{ij} = 1$. Then all the elements of Λ can be obtained as infinite paths in this graph. In an obvious way we can obtain a Markov shift from a subshift of finite type by defining a stochastic matrix of transition probabilities T_1 with $(T_1)_{ij} = 0$ iff $(T_0)_{ij} = 0$ and $\sum_j (T_1)_{ij} = 1 \ , \ \forall i$.

Finally to make complete contact with the finite automata and regular languages defined and discussed in chapter 4, we consider symbolic systems which are represented graphically, like the

[19] The reason for the notation T_0 will become clear below.

subshifts of finite type, except with the elements of A as edge labels instead of vertex labels. It might seem that there would be little difference in moving from a graphical description of symbolic processes in which elements of A are represented as vertex labels to one in which the elements of A are represented as edge labels. This is true in the probabilistic case. Quantities we would like to measure, like metric entropy are insensitive to differences in edge labeling and are dependent only on the topology of a graph and the probabilities defined on its edges.[20] In the parlance of ergodic theorists this is made possible by the fact that there always exists a finite to one factor map of an edge labeled system to a SFT, i.e., a system representable as a vertex labeled graph. The finite to oneness refers to the mapping of trajectories from the edge labeled system to the SFT under the factor map.

The situation in the topological case however is entirely different. There are systems representable as edge labeled graphs that do not have the topological Markov property. That is, knowing the previous symbol does not, in these cases, uniquely determine the set of symbols that are possible on the next iteration. By introducing such systems into symbolic dynamics a simple form of infinite memory has been introduced. Weiss [87] was the first to describe this type of system in symbolic dynamical terms and named them *strictly sofic systems* (SSS). We call the larger class of systems, including both SFT and SSS, *sofic systems*. Due to the memory properties alluded to above we refer to them as finitary rather than finite although the defining (finite) graph certainly provides a finite representation of the given process. This is the reason Weiss named them sofic after the Hebrew word "sof", meaning finite.

---

[20] Convergence to asymptopia however is in part governed by edge labeling as we shall see in chapter 6.

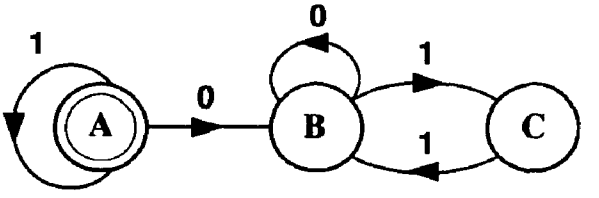


Figure 15 The even system discussed in the text.

The canonical example of a SSS is the even system shown in figure 15 It can be thought of as a parity recognizing machine. In this case the set of forbidden blocks $\mathcal{F}$ is infinite. We can see that infinite memory is required of our string recognizing machine for certain legal strings of the even system. Say that we send it a tape with some string y and our machine reads a 0 from the tape. Now if it reads a sequence of $1's$ it must keep counting, keeping track of the number of 1's, until it sees a zero, i.e., for an arbitrarily large sequence of $1's$. The finite graph representation of the even system indicates that this is not the kind of infinite memory that is required to characterize certain "high level" or more complex languages that we shall encounter in chapter 4, and why sofic systems in fact generate the simplest class of languages, regular languages.

We find that discussions of resource bounds, such as amount of memory, and language complexity are more easily discussed from the standpoint of computation theory where they were introduced. The techniques we develop in parts II, III, IV, and V are of necessity an amalgamation of the statistical descriptions of ergodic theory and information theory with the linguistic descriptions of computation theory.

Chapter 3 Elements of Information Theory

Section 1 Introduction; Information in Dynamical Systems

In chapters 2 and 4 we discuss hierarchical descriptions of systems exhibiting increased levels of randomness and complexity respectively. One of the goals of the thesis is to argue that the location of a given physical system in only one or the other of these hierarchies constitutes an incomplete description of that system. In other words the information gained by learning a systems location in these hierarchies is complimentary. This is in distinction to what is argued in the theory of algorithmic complexity, that measures of randomness are measures of complexity.

To quantify these ideas it is necessary to develop measures that distinguish between various levels in the hierarchies. The metric entropy described in chapter 2 is one example of such a measure. We shall go into more detail in this chapter. In particular the measures we develop are based on the notion of information and hence the present chapter is a highly selective introduction to those ideas from information theory that we find most useful.

An inkling of the importance in physics of what we now think of as information was glimpsed by the founding fathers of statistical mechanics Boltzmann and Gibbs [33] and later discussed by Szilard and Brillouin [10]. It is generally acknowledged however that Shannon [83] was the first to define information carefully. The present view of the use of information theory to characterize the relationship between natural phenomena, measurement channels and observers owes a great debt to the ideas of the Santa Cruz dynamical systems collective, in particular as discussed in [85] and [17] as well as to the ideas of Oliver Penrose [66]. We shall first consider information via a discussion from [66].

We can think of the state of a dynamical system we are trying to observe as a message we wish to obtain through our experimental apparatus considered as a measurement channel. In Shannon's language this would be a communication channel. For purposes of simplicity

we shall consider the channel to be what Shannon described as noiseless. This is clearly an approximation and a more careful treatment would state in what limits it is a good one. A schematic representation of how one might view a typical physics experiment in this context is given in figure 16

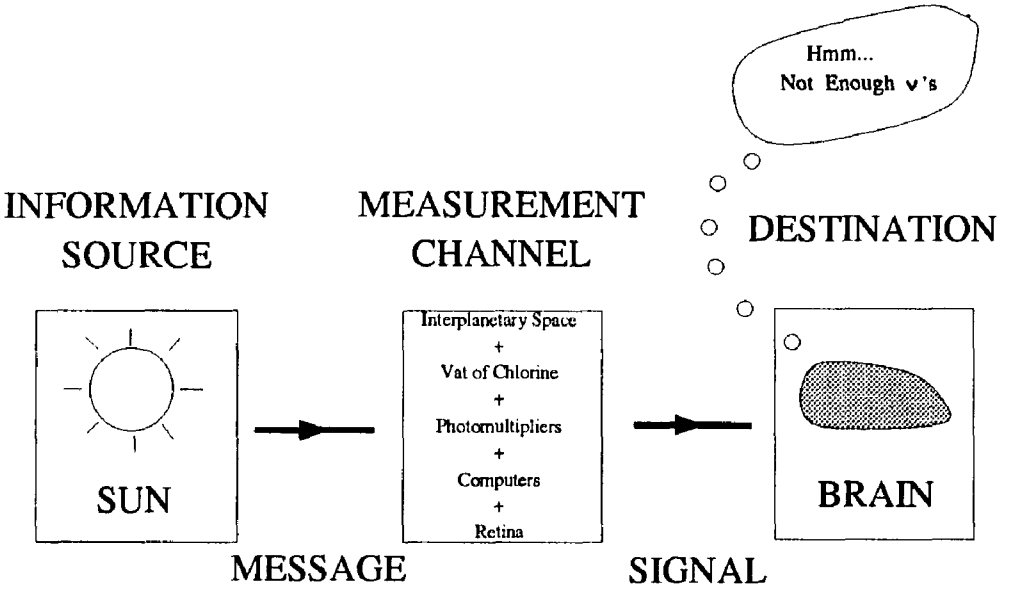


Figure 16 An experiment on a dynamical system viewed as a communication system.

Information theory allows us to quantify the amount of information in a message, i.e., a measurement. This quantity will be seen to be roughly proportional to the number of letters required to encode the message using the alphabet available to us via our measuring apparatus. This alphabet in turn determines the channel capacity of our measurement channel. The *channel capacity* is an upper bound on the rate at which the channel can transmit information. We note here that in any discussion of the channel capacity of a particular measurement channel it is important to discuss the sampling rate and spatial resolution of the experimental apparatus.

Our experimental apparatus has a finite resolution[21] and sampling rate which determine a limit on the number of observational states available to us in the characterization of the states of the underlying dynamical system.[22] Therefore we can ask the simple question: With a particular channel capacity, can we uniquely identify states of the dynamical system? Obviously if the underlying system is defined on a continuous phase space our finite resolution forces us to answer the above question in the negative but this is not what we're driving at. Lets assume that the specification of a dynamical state to some number of decimal places, say ten, gives a "complete enough" description. Lets further assume that we want to specify the dynamical state of a system of some "reasonable" number of constituents, say 10^{23}. Therefore we need approximately 10^{24} decimal digits to "uniquely" describe a state of the system. Obviously, there is no one to one map between states of the dynamical system and the observational states of any imaginable experimental apparatus so a statistical description of the information obtained in a given measurement is necessary. This was all understood clearly by the founders of statistical mechanics. The subtlety arises in determining how to coarse grain the dynamical states into observational states such that the information obtained in a measurement is maximal over the set of such coarse grainings. We hinted at at least a theoretical approach to this question in chapter 2 via the use of the metric entropy. An elaboration of these ideas will be discussed in the chapters on computational mechanics using the information quantities described in this chapter.

In the dynamical systems literature [17], [40] a connection between the dimension of an object and its entropy, an information theory based quantity, is frequently noted.[23] This connection is based on the definition of a spectrum of information quantities first discussed

---

[21] We shall use ϵ to denote the magnitude of the resolution in any particular case and will call the automata we construct via computational mechanics, ϵ-machines.

[22] That is if in fact at the crudest level we wish our observational states to be in one to one correspondence with the states of our measuring apparatus. We shall see that for certain combinations of natural phenomena and measurement channel a much simpler approach is available, but here we wish to simply discuss bounds on observation.

[23] By object in the above statement we typically mean the invariant distribution of a dynamical system but there are other examples so we shall keep the discussion general.

by Renyi [75] so we defer a detailed discussion of this connection until we have defined the Renyi spectrum.

The use of the idea of information in physics is frequently objected to on the grounds that it is vaguely defined in terms of physical quantities. Renyi's response to this type of objection was to compare the development of the concepts of energy and information.[75] He recounts the fact that it took a long time for the abstract concept of energy to develop and become useful. This was due to the multitude of forms in which it appeared , i.e., mechanical, thermal, chemical, electromagnetic and others. It didn't occur to physicists to characterize these in the same units and hence develop the concept of energy until long after the physical effects of these different forms were first understood. A situation similar to this exists now in terms of the various definitions of information and entropy. The obvious utility of the these different notions of information and entropy in their respective contexts indicates that a general, abstract definition of information will most likely constitute a significant advance.

We note that these comments apply even more to the as yet vaguely defined idea of complexity in physics. The motivation for a clear definition of complexity is widespread in the scientific community. Universal agreement on the particulars of such a definition is, however, far from the present condition. In this thesis our definition of complexity is closely linked with the concept of information, hence a precise formulation of both concepts is interdependent. Before examining our formulation it seems appropriate, for heuristic purposes, to examine the historical development of the abstract notion of information and its relation to physics. This should help develop some intuition about how to apply it in physically interesting situations.

Section 2 Historical Development of the Idea of Information

The Gospel According to Hartley

The earliest example, known to the author, of an attempt to quantify the idea of information was Hartley's [42] logarithmic measure of information. This measure can be thought of as the

information obtained when picking an element from a set $\mathcal{S}$ with $\| \mathcal{S} \| = N$ elements and where picking any of the N elements is equally likely, i.e.,

$$p = \frac{1}{N}$$

Hartley's measure of information is then defined to be

$$\begin{aligned} H_0 &\equiv -\log_2 p \\ &= \log_2 N \end{aligned}$$

where the reason for the subscript 0 will become apparent below.

If we wish to characterize information as a measure of the surprise we experience at selecting a particular element from a set, we expect any such measure to be monotonic in the size of the set. This is certainly true of Hartley's definition. This definition of information is identical to the definition of the thermodynamic entropy, to within a choice of units, of a system which is specified by a microcanonical ensemble. It is the special case of Shannon's entropy when the underlying probability distribution is uniform. It serves as the analog of the logarithm of the volume of available phase space, here simply in terms of a set of equally likely choices. We shall therefore simply refer to Hartley's measure as well as Shannon's generalized version as entropy.We also note that in anticipation of the close analogy between information theory and thermodynamics, this measure of information, like thermodynamic entropy, is additive when combining systems. As we shall see, this is a crucial property for information measures. If we consider two distinct systems with N_1 and N_2 elements respectively the information obtained by first picking one of the N_1 elements and then one of the N_2 elements is

$$\begin{aligned} H_0^T &= \log_2 N_1 N_2 \\ &= \log_2 N_1 + \log_2 N_2 \\ &= H_0^1 + H_0^2 \end{aligned}$$

The Gospel According to Shannon

The Shannon Entropy

It is generally acknowledged that Shannon, in 1948, initiated the formal study of information in terms of information transmission. The key assumptions in Shannon's formulation are: 1) that the amount of information transmitted through a communication channel can be measured by the amount of uncertainty removed upon receipt of a message and 2) that this uncertainty is proportional to the average value of the probability of symbols received.[83],[68]

Shannon generalized Hartley's information by considering the probabilities associated with subsets $\mathcal{S}_i \subset \mathcal{S}$ of a given set $\mathcal{S}$. He defined the information associated with a subset as

$$I_i = -\log_2 p_i$$

In particular, if the probability of a subset is simply given by the relative size of the subset

$$p_i = \frac{N_i}{N}$$

and

$$\begin{aligned} I_i &= -\log_2 p_i \\ &= \log_2 N - \log_2 N_i \end{aligned}$$

The total entropy associated with the set is taken to be the weighted average of the information of the subsets, i.e., we have Shannon's ubiquitous formula

$$\begin{aligned} H_1(\mathcal{S}) &= \sum_i p_i I_i \\ &= -\sum_i p_i \log_2 p_i \end{aligned}$$

This is the *Shannon entropy*. Put another way $H_1(\mathcal{S})$ is the expectation value of the quantity $\log_2 \frac{1}{p_i}$.

When it is clear what set and distribution we are reffering to we shall denote the Shannon information simply as $H_1 = H_1(\mathcal{S})$. We define $0 \log_2 0 \equiv 0$ in light of the fact that $\lim_{x \to 0} x \log_2 x = 0$

Shannon interpreted H_1 as an average code length. For example consider an alphabet A having $J = |\mathbf{A}|$ symbols a_i, $i = 0, 1, ..., J-1$ and a distribution for the probability of seeing a given symbol for a particular source $p_i = Pr(a_i)$. We can then interpret H_1 as the average information per symbol in a message given in units of bits/symbol. This interpretation of H_1 corresponds with our intuition that for the "most random" sources, i.e., those that yield uniform distributions on the symbols, we need longer code words on the average because we have to "account" for more source words. For less uniform distributions we can code the more probable words with shorter code sequences and use longer codes for the least likely sequences. When we take the average length by multiplying the codelength for the words by their probabilities and summing we get a smaller codelength for less uniform distributions.

As a simple example of the above ideas consider coin flipping. We can parametrize the process by specifying the probability p of obtaining heads, i.e.,

$$p_1 = Pr(Heads) = p$$

$$p_2 = Pr(Tails) = 1 - p$$

Then

$$\begin{aligned} H_1 &= -\sum_i p_i \log_2 p_i \\ &= -p \log_2 p - (1-p) \log_2 (1-p) \end{aligned}$$

As shown in figure 17 the maximum value of the information as a function of the parameter p occurs when $p = \frac{1}{2}$, i.e., a uniform distribution on heads and tails. In this case we need 1 bit to code the results heads or tails on the average, as the outcome of each toss is completely uncertain. If the coin is biased we don't need to specify the same amount of information on the average, i.e., over a large number of tosses. At the other extreme if $p = 0$ or $p = 1$ there is no uncertainty involved and we need 0 bits on the average to code the results.

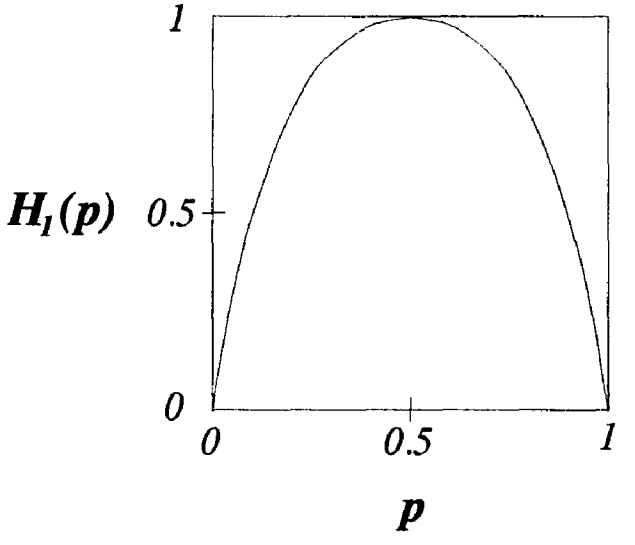


Figure 17 Schematic representation of the Shannon entropy vs. the probability of heads for coin tossing.

Other Measures of Information

Shannon defined other quantities we find useful and so define here. Henceforth we shall consider the sets over which our distributions are defined to be finite alphabets. Consider the alphabet **A** as above and another alphabet **B** having $K = |\mathbf{B}|$ symbols b_l, $l = 0, 1, ..., K-1$. Then along with the marginal distributions $p_i = Pr(a_i)$ and $p_l = Pr(b_l)$ we have the joint distribution $p_{il} = Pr(a_i, b_l)$ and the conditional distributions $p_{i|l} = Pr(a_i|b_l)$ and $p_{l|i} = Pr(b_l|a_i)$. From these distributions we define the following quantities: the *joint entropy*

$$H_1(\mathbf{A}, \mathbf{B}) = -\sum_{i,l} p_{il} \log_2 p_{il}$$

the *conditional entropy*

$$H_1(\mathbf{A}|\mathbf{B}) = -\sum_i p_i \sum_l p_{l|i} \log_2 p_{l|i}$$

and the *mutual information*

$$I_1(\mathbf{A}; \mathbf{B}) = \sum_i \sum_l p_{li} \log_2 \frac{p_{li}}{p_i p_l}$$

If we have two distributions defined on $\mathbf{A}$, $p_i = Pr_1(a_i)$ and $q_i = Pr_2(a_i)$ we define the *relative entropy*

$$D_{q,p} = \sum_i p_i \log_2 \frac{p_i}{q_i}$$

The relative entropy is also known as the *discrimination* or *Kullback-Leibler distance*. It can be thought of as a measure of distance between the distributions, although it is not strictly a metric because it is not symmetric in its arguments and does not obey the triangle equality.

Finally, if we consider that symbols from $\mathbf{A}$ are input to some communication channel as a message and symbols from $\mathbf{B}$ are read as a signal, then the *channel capacity* of the communication channel is given by

$$C = \sup_{p_i} I_1(\mathbf{A};\mathbf{B})$$

In the context of making measurements on a dynamical system the channel capacity is essentially the topological entropy that one obtains from a given data set.

The following important relationships for these quantities can be easily verified [9]:

$$\begin{aligned} I_1(\mathbf{A};\mathbf{B}) &= H_1(\mathbf{A}) - H_1(\mathbf{A}|\mathbf{B}) \\ &= H_1(\mathbf{B}) - H_1(\mathbf{B}|\mathbf{A}) \\ &= H_1(\mathbf{A}) + H_1(\mathbf{B}) - H_1(\mathbf{A},\mathbf{B}) \end{aligned}$$

$$D_{q,p} \geq 0 \ \ with \ \ D_{q,p} = 0 \ \ iff \ \ q_i = p_i \ \forall i$$

$$I_1(\mathbf{A};\mathbf{B}) = D_{p_{il},p_ip_l} \geq 0 \ ,$$

$$D_{p_{il},p_ip_l} = 0 \ \ iff \ \ p_{il} = p_ip_l \ \ \forall i,l$$

As well as in the current work, these quantities and their relationships are used and discussed, in the context of dynamical systems in [85], [17], and [31].

The Gospel According to Renyi

With the goal of defining information as an element of a general statistical mechanical theory we turn to Renyi's definition of a "generalized" measure of information.[75]

Renyi devised a measure of information which could be related directly to statistical mechanics by providing a parametrized measure of information for which the parameter can be thought of as an inverse temperature. This allows for the definition of a partition function, in strict analogy to that defined with respect to the canonical distribution in statistical mechanics. As we shall demonstrate a direct analogy to the definition of free energy in terms of the canonical distribution is the definition of generalized mean used by Renyi in his definition of information measure.

He argued that two fundamental, and intuitively compelling, postulates were involved in the above definitions of information. The first postulate was the requirement alluded to previously that information be additive. The second postulate is that if different amounts of information occur with different probabilities then the total information should be the weighted average of the individual informations. Shannon's measure of information obeys both postulates and uses the standard definition of mean, what Renyi calls the linear average. Renyi then considered the average in the second postulate in terms of a generalized mean [48] of real numbers $x_1, \ldots, x_n$ with weights $p_1, \ldots, p_n$ given by

$$\varphi^{-1}\left(\sum_{i=1}^{n} p_i \varphi(x_i)\right)$$

where $\varphi(x)$ is any real valued, continuous, monotone function. In particular for the mean of a set of weighted informations

$$H = \varphi^{-1}\left(\sum_{i=1}^{n} p_i \varphi(I_i)\right)$$

He then proved that the only types of function $\varphi(x)$ in the above definition which also satisfy the additivity postulate are linear and exponential functions. With $\varphi(x)$ linear we recover Shannon's

definition of information. In the case of an exponential Renyi considered the parameterized form

$$\varphi(x) = 2^{(1-\alpha)x}$$

Then we define the α-order information for a given distribution

$$\begin{aligned} H_\alpha &= \frac{1}{1-\alpha}\log_2\left(\sum_{i=1}^{n} p_i \cdot 2^{(1-\alpha)I_i}\right) \\ &= \frac{1}{1-\alpha}\log_2\left(\sum_{i=1}^{n} p_i^\alpha\right) \end{aligned}$$

With p_i defined on the alphabet A, and consistent with the notation in the previous section $H_\alpha = H_\alpha(\mathrm{A}).H_\alpha$ has a form reminiscent of the equation for the free energy in terms of the partition function

$$\begin{aligned} F_\beta &= \varphi^{-1}\left(\sum_{i=1}^{n} g_i\varphi(E_i)\right) \\ &= -\frac{1}{\beta}\log\left(\sum_{i=1}^{n} g_i e^{-\beta E_i}\right) \\ &= -\frac{1}{\beta}\log Z_\beta \end{aligned}$$

$\varphi(E_i) = e^{-\beta E_i}$ is the Boltzmann factor of the i^{th} state, g_i it's degeneracy, and Z_β the partition function. We shall make extensive use of the Renyi information of probability distributions over symbol sequences representing trajectories in dynamical systems.

It is also useful to note that if we pick this parametrized version of $\varphi(x)$ we in fact recover our previous definitions of information at specific parameter values, namely when $\alpha = 0$ we recover the Hartley information

$$H_0 = \log_2 N$$

and when $\alpha = 1$ the Shannon entropy, via L'Hospitals rule

$$\begin{aligned}
\lim_{\alpha \to 1} H_\alpha &= \lim_{\alpha \to 1} \left(\frac{1}{1-\alpha} \log_2 \left(\sum_{i=1}^{n} p_i^\alpha \right) \right) \\
&= \frac{\lim_{\alpha \to 1} \frac{d}{d\alpha} \left(\log_2 \left(\sum_{i=1}^{n} p_i^\alpha \right) \right)}{\lim_{\alpha \to 1} \frac{d}{d\alpha}(1-\alpha)} \\
&= \lim_{\alpha \to 1} - \frac{\left(\sum_{i=1}^{n} p_i^\alpha \log_2 p_i \right)}{\left(\sum_{i=1}^{n} p_i^\alpha \right)} \\
&= - \sum_{i=1}^{n} p_i \log_2 p_i
\end{aligned}$$

We now make a slight digression to mention a consequence of the above additivity postulate, a consequence which leads to one of the fundamental questions of computational mechanics. The additivity postulate for a function combined with the postulate that for closed systems this function is nondecreasing provides a definition of the thermodynamic entropy. It can be shown, however [66] that in the statistical mechanics of finite systems no quantity can be defined which satisfies both of the postulates satisfied by the thermodynamic entropy. It is only in the thermodynamic limit (volume and particle number going to infinity with density finite) that such a quantity can be defined. So only in the thermodynamic limit is it appropriate to think of statistical mechanics as providing a basis for thermodynamics. We shall characterize this problem in a slightly generalized form using the information spectra defined above: In what sense (i.e., of what magnitude are the corrections) can an abstractly defined set of macroscopic variables, or observation states, appropriate to the analysis of a particular problem be defined in terms of the statistical properties of a "more fundamental" , and perhaps mechanically more easily defined, set of microscopic variables? The attempt to answer this question results in the technique we shall refer to in subsequent chapters as determining the spectra of ϵ-machines.

As mentioned in the introduction the Renyi spectrum is often used to discuss the connection between information quantities and dimension quantities for a given system.[37],[17] A basic

technique in dynamical systems theory is to first consider a partition of the phase space of a dynamical system with largest partition element of size $\sim \epsilon$. Upon iteration of the dynamical system for arbitrarily long times (provided the system is ergodic) a distribution on the elements of the partition is obtained and one can calculate the Renyi spectrum of this distribution. The connection with a spectrum of dimensions is obtained by dividing the information quantities by $\log \epsilon$ and taking the limit $\epsilon \to 0$, i.e.,

$$D_\alpha = -\lim_{\epsilon \to 0} \frac{(\alpha - 1)^{-1} \log \sum_{i=0}^{N(\epsilon)-1} p_i^\alpha}{\log \epsilon}$$

where $N(\epsilon)$ is the number of partition elements at resolution ϵ. In other words the dimensions are the scaling exponents of the distribution with respect to volume in the support space.[24]

In particular we see that the above expression yields the box dimension for $\alpha = 0$, i.e.,

$$\begin{aligned} D_0 &= -\lim_{\epsilon \to 0} \frac{(0 - 1)^{-1} \log \sum_{i=0}^{N(\epsilon)-1} p_i^0}{\log \epsilon} \\ &= \lim_{\epsilon \to 0} \frac{\log N(\epsilon)}{\log \epsilon} \end{aligned}$$

Similarly we can obtain the information dimension for $\alpha = 1$ and the much discussed [37] correlation dimension for $\alpha = 2$.

The Gospel According to Rissanen

Finally, we briefly mention an information quantity defined by Rissanen [76] and used as a measure of what he calls stochastic complexity. We feel it is an important quantity in that Rissanen defined it with respect to the model class one is using to characterize a process, an important and by now familiar theme of this thesis. He argues that the Shannon entropy is defined with respect to some distribution of which, in the general case of modeling data,

[24] We note that the standard notation in the current literature replaces α in the above expression with q and uses α as the Legendre transform companion of q. We use the above notation to maintain consistency with our earlier discussion.

we have no knowledge, and this fact limits the practical utility of the Shannon entropy as a measure of information.

Rissanen uses the notion of complexity developed by Solomonoff, Kolmogorov, and Chaitin [55] in the context of computation theory, to extend Shannon's measure of information to a more practically useful form. This notion of complexity, called algorithmic complexity, is defined with respect to binary strings and is essentially just given by the length of the shortest program, implemented on a universal computer, necessary for reproducing a particular string. Rissanen makes use of this notion by replacing the universal computer with a class of probabilistic models. This, according to Rissanen, provides the missing component of Shannon's information, that which identifies the uncertainty inherent in the estimation of models. He further argues that interpreted as a code length his measure of information provides a model dependent definition of probability and relegates any subjectivity to the choice of model classes. Combined with a hunting license for determining model optimality, i.e., his minimum description length principle, it would appear that we then have a universal procedure for model estimation. We shall argue, however, that in fact the determination of model classes is a subtle and important issue and is no way specified by an after the fact criterion for optimal model selection. The problem is something like searching for a global minimum in a fractal mountain range. Consequently a constructive procedure for determining model classes is essential.

The information quantity he defined, the stochastic complexity, is an extension of Shannon's notion of the average code length associated with a source and is directly related to the algorithmic complexity as indicated above. We first define the model class

$$M_k = \{P(\mathrm{S}|\theta), \pi(\theta)\}$$

where: S is a data set, θ is a k vector of parameters, $\pi(\theta)$ is a prior distribution over the parameters, and $P(\mathrm{S}|\theta)$ is the probability of S given θ. Then the stochastic complexity of

S given M_k is

$$\begin{aligned} I(\mathrm{S}|M_k) &= -\log_2 P(\mathrm{S}) \\ &= -\log_2 \int P(\mathrm{S}|\theta)\, d\pi(\theta) \end{aligned}$$

which resembles the above definitions of information quantities.

Consider the maximum likelihood estimates $\hat{\theta}$ of the parameters θ obtained from a data set S. Using the minimum description length principle [76] we expand $I(\mathrm{S} \mid M_k)$ in the fluctuations about the maximum likelihood estimate, i.e., in the quantity $\left\| \left(\theta - \hat{\theta} \right) \right\|$. If we keep only first order terms we obtain

$$I(\mathrm{S}|M_k) \approx -\log_2 P\left(\mathrm{S}|\hat{\theta}\right) + \frac{k}{2} \log_2 n$$

where n is the size of the data set. We see that there are two terms here, the first $-\log_2 P\left(\mathrm{S}|\hat{\theta}\right)$ is the Shannon term and corresponds to the prediction error in a modeling interpretation. The second term $\frac{k}{2} \log_2 n$ is the model dependent term as promised.

As an example consider a Bernoulli shift, as discussed in chapter 2, with alphabet size $J = |\mathrm{A}|$ yielding a data string S of length n. Assume the i^{th} symbol of A occurs n_i times in S and we note that $\sum_{i=0}^{J-1} n_i = n$. Choose, as our parameters the J probability estimates for the symbols of A, θ_i. Then we have

$$P(\mathrm{S}|\theta) = \prod_{i=0}^{J-1} \theta_i^{n_i}$$

The maximum likelihood estimates of the θ_i given our data string S are

$$\hat{\theta}_i = \frac{n_i}{n}$$

and hence to lowest order

$$\begin{aligned} I(\mathbf{S}|\theta) &= -\log_2 P(\mathbf{S}) \\ &\approx -\log_2 P\left(\mathbf{S}|\hat{\theta}\right) + \frac{J}{2}\log_2 n \\ &= -\log_2 \prod_{i=0}^{J-1} \hat{\theta}_i^{n_i} + \frac{J}{2}\log_2 n \\ &= -\log_2 \prod_{i=0}^{J-1} \left(\frac{n_i}{n}\right)^{n_i} + \frac{J}{2}\log_2 n \\ &= -n\sum_{i=o}^{J-1} \left(\frac{n_i}{n}\right)\log_2\left(\frac{n_i}{n}\right) + \frac{J}{2}\log_2 n \end{aligned}$$

where we see clearly, in this example, the division into a Shannon type term that becomes dominant for large n and a model dependent term. A more careful approach including higher order terms and using Stirlings approximation can be found in [76]. We shall make use of a similar analysis in computational mechanics when we derive generic thermodynamic properties from the combinatorics of the Markov shifts that arise naturally from our reconstruction techniques.

Chapter 4 Elements of Computation Theory

Section 1 Introduction

Apart from the practical utility of numerical computation in physics we shall argue in this chapter that the structures defined in theoretical computer science provide useful models for describing the complex behavior of nonlinear systems. In dynamics a description of complex behavior generally consists of two major components. The first is referred to as topological and is a structural description of the set of possible behaviors of the system. For example the specification of a strange attractor gives a topological description of a systems behavior; it specifies the locus of phase points for all initial conditions in some basin of attraction. The second major component of a dynamical description is probabilistic. It provides information on how likely any of the behaviors specified in a topological description are to occur. In the example of a strange attractor, probabilistic information would be provided by a probability distribution defined on the strange attractor.

We assert in this thesis that the structures of theoretical computer science provide an appropriate framework for the topological description of the behavior of complex nonlinear systems. This is similar to the way in which differential equations provide a useful framework for the description of the behavior of linear systems.

In Chapter 1 we argued that the explicit specification of the class of models one is using in the analysis of data is crucial in trying to classify and organize descriptions of complex physical behavior. Generally, however, model specification has not been of much concern in 20th century theoretical physics. Our conception of fundamental entities, i.e., particles and fields, and their interactions have been drastically modified by the developments of relativistic quantum theory. The basic models of physical processes, however, have not. The quantum mechanical hydrogen atom is a straightforward variant of the Kepler problem of orbital motion. The calculational

tour-de-forces of quantum field theory sit atop a model of weakly coupled harmonic oscillators. In searching for models and methods of classification for complicated systems with strong nonlinear couplings, one of the goals of the work in this thesis was to search for examples in which the careful specification of model classes is important.

One of the most obvious examples is computation theory in which the idea of a model of computation arises. A computation or task can be thought of as a sequence of operations or an algorithm. If we code the algorithm in terms of some set of symbols we can think of the algorithm as a legal sequence, or *word* of some language. For instance consider the task of printing the first 20 billion digits of π. We can code this task in the symbols of some computer language and hence can consider this task as one of the words recognized by the compiler of the computer language used. We note that we shall use the idea of a language in a slightly more abstract way than this example indicates. The way we shall think of a language is as a set of sequences characterized by the intrinsic difficulty of recognizing the entire set. Used in this way most of the common computer languages we would use to accomplish the above task would be considered equivalent. A particular task coded in one language could be "trivially" recoded into another.

Finally a *model of computation*, as illustrated in figure 18, is a representation of a language, where we use language in the abstract sense as a set characterized by the difficulty of determining membership. A simple example of a model of a language is a direct listing of the words of the language, i.e., a "dictionary". Another model is some mechanical procedure, i.e., a *machine*, for recognizing or producing words. This machine, in its recognizing mode, can be thought of as a function . Whereas a function on real numbers, gives a real number when fed a real number a machine gives a yes or no, i.e., in or not in the language, when fed a word. Another model we shall discuss is the *grammar* of a language, that is a set of rules for producing words. In this context a better synonym for legal sequence might be sentence. As we shall see there are equivalent models for languages of a given capability. For example the simplest or least capable

languages, called regular languages, are characterized by machines known as finite automata as well as by grammars known as linear grammars.

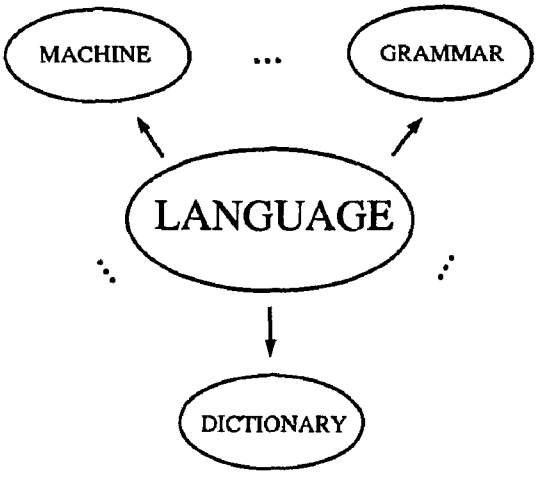


Figure 18 A language and some models of the language.

A careful specification of the model of the language of a given computation or task helps to determine whether or not the task can be accomplished and if so what resources, e.g., time and memory, will be necessary. We have found that the models and classes of models described by computation theorists provide a useful framework for the definition of classification invariants for complex data sets, in particular data sets produced by the observation of nonlinear dynamical systems.

An intuitive motivation for the use of computation theory in physics is as the natural language for the description of constraints. Textbooks in statistical mechanics discuss constraints in the context of thermodynamic entropy. The relaxation of a constraint leads to a system with higher entropy or randomness. We shall argue that a quantitative characterization of the constraints, as provided by the tools of computation theory, yields more information about a system than one would have had without such a description. An example is the case of a spin

glass for which a complete description of the constraints would yield a detailed description of the architecture of the set of degenerate ground states. The tendency in physics has been to isolate a set of states with equal energy, e.g., a set of degenerate ground states, and invoke the assumption of equal a priori probabilities for the elements of such a set. A modification of such procedures is indicated in the case of a system with complicated constraints. While the standard procedures of determining Boltzmann factors and constructing partion functions yield the appropriate probabilistic information about the system there is useful structural information as well, information contained in a knowledge of the constraints.

As mentioned in chapter 1, computation theory had its beginnings in the foundations of mathematics and logic; more recent contributions have come from workers in diverse fields. Biologists studying nets of neurons and morphogenisis of plant structure, engineers developing switching theory and linguists attempting to model natural language have all provided models that have become cornerstones of the modern theory of computation. What is important about these models is that they quantify the resources necessary to accomplish a specific task. For example one of the main uses of switching theory is to quantify the loads and consequent resource requirements of telephone networks. With the advent of computing machinery, computer scientists began to think of these resources in terms of time and memory and developed theories of the computability and complexity of certain tasks with respect to these quantities.

A nice example of the interplay of ideas from computer science and physics is the development of the notion of simulated annealing for systems like spin glasses with "rugged energy landscapes".[46] An analogy between finding the shortest path in the traveling salesman problem and finding the true ground state of a spin glass was important in this development. It can be shown that finding the true ground state of a moderately sized spin glass with any conceivable computational resources is in general not possible in a "reasonable amount of time".[25] Given this fact, simulated annealing allows one to find a good approximation to the

[25] Computer scientists refer to the class of "impossible" problems to which the traveling salesman belongs as the class of NP-complete problems.[32]

ground state, i.e., a very deep local minimum with energy close to the ground state energy, by "starting" the system at high temperature and lowering the temperature slowly. At high temperature the time scale for hopping between local energy minima is small and the system explores a large region of state space. As the temperature is slowly lowered the system eventually "freezes" in a deep local minimum.

As another simple illustration of the use of computational ideas in physics consider a system described by a differential equation. If this equation is discretized the difference operators corresponding to the differential operators in the original equation have more terms, at a larger number of points, according to the degree of the derivatives in the original equation. Hence differential equations containing higher derivatives require more "memory", e.g., stored variables, to calculate a value at any given point.

Before describing Chomsky's hierarchy as a prototype of classification based on complexity we need some basic definitions which will be given in the next section.

Section 2 Definitions

Equivalence Relations

We shall see that a fundamental operation used in parts II, III, IV, and V is to partition the set of trajectories of a dynamical system into equivalence classes based on an equivalence relation designed to produce optimally predictable states, i.e., sets of trajectories. Hence we define equivalence relations in this section.

Consider a set S. A *binary relation* R on S is a set of pairs of elements of S denoted aRb, i.e., $aRb = \{(a,b) : a,b \in S\}$. A relation is called an *equivalence relation* if the following conditions hold $\forall a,b,c \in S$

$i)$ reflexivity $\quad aRa$

$ii)$ transitivity $\quad aRb$ and $bRc \Rightarrow aRc$

$iii)$ symmetry $\quad aRb \Rightarrow bRa$

The importance of an equivalence relation is that it partitions the set S into disjoint nonempty subsets S_i referred to as *equivalence classes*, obeying the following properties

$a)$ $S_i \cap S_j = \phi$

$b)$ $\forall a, b \in S_i$, aRb is true

$c)$ $\forall a \in S_i$ and $\forall b \in S_j$, aRb is false

Strings and Languages

We next define some primitive concepts which will be used specifically in this chapter but also generally in computational mechanics and in parts II, III, IV, and V. The notion of a symbol occurs naturally in dynamics as the label of a partition element in phase space. In this chapter however, a *symbol* is simply an undefined primitive quantity, like a point in geometry. A set of symbols will be called an *alphabet* $\mathbf{A}$. A *string* ω is a finite sequence of symbols $s \in \mathbf{A}$. A *concatenation* of two strings is the string obtained by writing the first string followed by the second. For example consider the alphabet $\mathbf{A} = \{a, e, j, o, s, t\}$. Then eat, at, $joes$ are strings of symbols of $\mathbf{A}$. $eatat$ is a concatenation of the first two strings and $eatatjoes$. We shall use starred notation on a given string to represent the infinite concatenation of that string, i.e., ω^* denotes $\omega\omega\omega\omega\omega\omega$....

A *formal language*, or simply *language* $\mathcal{L}$, is a set of strings from a particular alphabet. One of the simplest formal languages to specify is the set of all strings on a given alphabet denoted $\mathbf{A}^*$, e.g., the set of binary strings, with $\mathbf{A} = \{0, 1\}$ and $\mathbf{A}^* = \{\epsilon, 0, 1, 00, 01, 10, 11, ..\}$ where ϵ is the empty string. Note that the words of this language correspond to the possible sequences of coin tosses from the example in chapter 1.

Another useful characterization of languages is in terms of the set of *irreducible forbidden sequences* $\mathcal{F}$ of $\mathcal{L}$. This is the set of words in $\mathbf{A}^*$ which are not in $\mathcal{L}$ and which are not concatenations of smaller forbidden sequences.

Graphs

A *graph* $G = (\mathbf{V}, \mathbf{E})$ is a set of *vertices* $\mathbf{V}$ and a set of *edges* $\mathbf{E}$ connecting the vertices. The edges can be thought of as pairs of vertices that the edge connects

$$e = \{v, v'\} \quad e \in \mathbf{E}, \quad v, v' \in \mathbf{V}$$

A finite graph is one for which $\|\mathbf{V}\| < \infty$ where $\|\mathbf{V}\|$ denotes the number of vertices in the graph. A *path* in a graph is a sequence of vertices such that there exist an edge for each pair of consecutive vertices in the path. A *connected graph* is a graph for which there are paths connecting any two vertices in the graph. A *cycle* is a path whose initial and final vertices are the same.

A *directed graph* or *digraph* is a graph for which the edges are given a direction, i.e., the edges are ordered pairs, $e = (v, v')$, rather than simply pairs. The *in-edges* of a particular vertex are the edges entering that vertex and the *out-edges* of a vertex are those leaving the vertex. A *labeled graph* is one for which each edge of the graph is labeled with a symbol and a *labeled digraph* is defined in an obvious way. A *strongly connected graph* is a digraph for which there are paths of directed edges connecting any two vertices in the graph. In chapter 1 figure 8 gives an example of a labeled digraph and figure 4 gives an example of a strongly connected labeled digraph.

A *tree* is a special kind of digraph with the following properties: a) there is a special vertex called the *root* from which there is a path to every other vertex and to which there is no path from any vertex, b) there are no cycles, and c) each vertex, besides the root has only one edge entering it. A *terminal vertex* is a vertex in the tree with no out-edges. A *parse tree* is a tree all of whose paths from the root to the terminal vertices have the same length. Examples of parse trees are given in figures 2 and 6 of chapter 1.

Machines

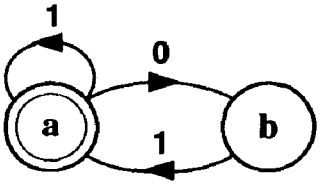
As discussed in the introduction to this section one particular model of computation consists

essentially of a language over some alphabet of symbols and a machine that reads the symbols from a tape and performs some action. In most cases this action consists of simply producing other symbols although this representation is by no means unique. We use the term machine in the sense of Turing, i.e., a conceptual entity that performs a computation and not necessarily as the physical instantiation of any particular computational model. Formally a simple model of a *machine*, or *automaton*, is a quintuple $M = \{V, \mathbf{A}, T, v_0, V_F\}$ where V is a set of states. $\mathbf{A}$ is an alphabet of *tape symbols*. $T : V \times \mathbf{A} \rightarrow V$ is the transition rule, taking the current state and a symbol read from a tape into a new state i.e., $v' = T(v, s)$ with $v, v' \in V$ and $s \in \mathbf{A}$. v_0 is the initial state, and V_F is a subset of V and is the set of final or accepting states. We define *accepting states* as follows. If after having read some sequence of symbols from a tape, the machine is in an accepting state $v \in V_F$, that sequence of symbols is an element of the language $\mathcal{L}$ recognized by the machine M. Therefore by the language $\mathcal{L}$ *recognized* by the machine M we mean that set of symbol sequences that when read by M starting in v_0, leave M in an accepting state $v \in V_F$. If the set V is finite we call the automaton a *finite automaton.*

A convenient representation of a finite automaton is as a graph with the set V corresponding to the vertices of the graph and an edge corresponding to each transition rule, labeled by the appropriate symbol from $\mathbf{A}$. Figure 19 shows a graph and the corresponding machine and alphabet[26].

---

[26] Note that in the figure there is no transition from state **B** on the symbol 0. An alternate method of representing these "forbidden transitions" is commonly encountered in the computation theory literature.[43]. This method adds a state $\mathbf{C} \notin V_F$, i.e., a *nonaccepting state* such that this state "absorbs" all forbidden transitions. Using our method, encountering a nonexistent transition while reading a string indicates that this string is not part of the language associated with the given system. Using this alternate method, entering the nonaccepting state represents the same situation.



$M=\{V, \mathbf{A}, T, v_0, V_F\}$

$V=V_F=\{\mathbf{a}, \mathbf{b}\}$

$v_0=\mathbf{a}$

$\mathbf{A}=\{0,1\}$

T:

s \ v	a	b
0	b	X
1	a	a

Figure 19 The labeled digraph, machine, state set, start state, alphabet and transition function for a system called the golden mean system. The X in the upper right entry of the table for the transition function represents the fact that there is no transition from state b on the symbol 0.

If , as is the case for the machine in figure 19, there is a only one state in each entry of the transition table, we say that all the transitions are unique. This means that being in a state and reading a symbol determines a unique state as the result of a transition. Interpreted in terms of the graph this means that there cannot be two ore more edges, labeled with a particular symbol, leaving a vertex and going to distinct vertices. We call machines that have this property *deterministic.*

At this point we note the distinction between generating and recognizing a language. We define *parsing* a sequence for a machine as reading that sequence from a tape. We can use a machine to parse a sequence of symbols and determine whether or not it is a legal word of the language recognized by the machine, as described above. On the other hand we can use a given machine to generate all the legal sequences of the corresponding language by following all the paths in the digraph representing the machine.

We finish with a slight digression from the formal language theory of machines for the purpose of introducing a common model class of stochastic processes, the class of Markov chains. These are strongly related to the types of machine defined above. A *Markov chain* is a set of states $\mathcal{V}$, and a set of transitions $\mathcal{E}$ between these states. Each element of $\mathcal{E}$ is an ordered pair of states of $\mathcal{V}$, i.e., $\mathcal{E} = \{(\nu_i, \nu_j) : \nu_i, \nu_j \in \mathcal{V}\}$. Associated with each element of $\mathcal{E}$ is a real number giving the probability $Pr(\nu_i \rightarrow \nu_j)$ of a transition from state ν_i to state ν_j with the restriction that $\sum_i Pr(\nu_i \rightarrow \nu_j) = 1$, i.e., the probability of making some transition from any given state is unity. The standard graphical representation of a Markov chain, in terms of states as vertices and transitions as edges suggests a strong structural similarity with the machines we have defined in this section. There are important differences however.

If we regard the states of the chain as an alphabet then the vertices in the graphical representation are labeled but the edges are not. Although in the pictorial representation of Markov chains the edges are labeled with the transition probabilities, these are usually not thought of as formal symbols. So the set of sequences of states of the chain determined by paths in its graphical representation correspond to a language $\mathcal{L}$ defined on the symbols representing the states of the chain. The set of such languages, discussed in chapter 2, and corresponding to the set of finite Markov chains is referred to by ergodic theorists as the set of subshifts of finite type (SFT). One might think, due to the structural similarities noted above, that this set of languages is the same set as the set of languages recognized by finite automata.

This is in fact not the case and in the context of ergodic theory the set of languages recognized by finite automata was called the set of sofic systems (SS), by Weiss [87]. The set of SFT are a subset of the set of SS. The set of SS which are not SFT turn out to have interesting topological properties exemplified by their semigroup structure as we shall show in part V

The upshot of these remarks is that a stochastic generalization of finite automata is not a trivial mapping onto the set of Markov chains as we shall discuss in more detail in parts II, III, IV, and V.

Grammars

Another convenient class of models of formal languages is the class of grammars. Grammars define production rules for generating legal sequences of the language. They were introduced by Chomsky in his attempt to classify natural languages.[14] Grammars have provided convenient linguistic descriptions of spatio-temporal pattern generation, most notably in the the use of L-systems for the description of plant growth.[56]

Formally a *grammar* is a quadruple $\Gamma = (\sum, \mathbf{T}, \mathbf{P}, \alpha)$ where $\sum$ is a set of variables. $\mathbf{T}$ is a set of terminal variables; legal strings ω are composed of sequences of symbols of $\mathbf{T}$. $\mathbf{P}$ is a set of production rules. Each production rule $p \in \mathbf{P}$ is a rule for replacing variables with strings of symbols. Productions have the form $\sigma \rightarrow t$ where $\sigma \in \Sigma$ and $t \in (\Sigma \cup \mathbf{T})^*$. α is a particular variable called the start symbol. We "derive" legal strings in the language by using the production rules, which give the legal substitutions for variables in terms of variables and terminals. Starting with the start symbol and performing successive substitutions, any string of terminals we obtain in this manner is a legal string in the language.

Section 3 The Chomsky Hierarchy

Introduction

We now turn to a description of Chomsky's hierarchy of increasingly capable models of computation.[14] This hierarchy consists of four basic classes. It is by no means exhaustive in describing the various models of computation that have been suggested in the literature and is meant here to serve only as an indication of the kinds of question that arise in the attempt to classify systems based on complexity. For a detailed treatment see [43].

We argue that the higher levels of this hierarchy are extremely useful in describing the structure and behavior of complex nonlinear systems. As an example the highly correlated behavior exhibited by certain systems at phase transitions is well characterized by treating the behavior of the system as a computation and attempting to classify the computational power

inherent in this observed behavior. Such a classification scheme uncovers subtle structures in such behavior that are missed by statistical measures, such as correlation functions, which are traditionally used to indicate the onset of a phase transition. We demonstrate this in part III by showing that the complex behavior of a system at the onset of chaos is well characterized by a variation of a context free grammar, which we define later in this section.

Despite this we shall not describe the higher levels of the hierarchy in much detail as the bulk of this thesis is concerned with finitary processes, or those that can be modeled within the lowest level of the hierarchy, i.e., as finite automata. Although much remains to be done in terms of characterizing complex systems as computations and locating their position in the hierarchy, there is already a tremendously rich structure associated with the set of stochastic finite automata. Therefore aside from the work in part III mentioned above we concern ourselves in this thesis primarily with stochastic finite automata.

Finite Automata and Regular Languages

Regular languages constitute the lowest rung of the Chomsky hierarchy. They represent the easiest class of languages to describe in a sense to be made precise below. For our purposes the most important characteristic of regular languages is that they are the languages recognized or generated by *finite automata* (FA), where by FA we mean machines as described in the introduction and for which

$$||V|| < \infty$$

$||V||$ for an FA is generally thought of in computer science, and here as well, as the amount of memory associated with that FA. More specifically the number of bits of memory of the machine M is $\log_2 ||V||$.

We now define some special classes of FA. If the FA is deterministic it is called a *deterministic finite automaton* (DFA). If it is not deterministic it is called a *nondeterministic finite automaton* (NFA).

Any regular language $\mathcal{L}$ can be recognized or generated by many FA. To each regular language $\mathcal{L}$, however corresponds a unique minimal DFA. By a *minimal DFA* we mean the DFA with the fewest vertices which recognizes $\mathcal{L}$. That a unique minimal DFA exists for any regular language $\mathcal{L}$ is an extremely important point that we shall make frequent use of.[43]

The word deterministic is used differently here than the way the it is used in physics. Here nondeterminism stems from the machine not in general admitting a unique transition given a state and input symbol. The machine, in some state, has a choice of transitions on a particular symbol. Even if the machine is a DFA, however, there may be branching. This means that, unlike the solution of a differential equation in phase space, the current state does not uniquely determine the subsequent state. There are, for DFA's with branching, different choices of subsequent states based on which symbol is read or produced. Determinism in physics, based on existence and uniqueness theorems of differential equations, is much more restrictive. A succession of physical states, as determined by the solution of some differential equation, is uniquely determined by the initial state; no choice is possible. We emphasize that if there is branching in the DFA its states can't correspond to physical states determined by physical laws written as differential equations. The DFA states in this case are those of a stochastic or statistical model. Research in dynamical systems has shown however that even in cases where the deterministic dynamics can be written down explicitly, stochastic models arise when any level of coarse graining of the underlying state space is required,e.g., for making measurements.[22]

We wish to make a couple of final comments about regular languages. The first is that there are useful ways of proving that a language is nonregular. A frequently used technique is called the pumping lemma.[43] This lemma basically states that accepted words in a regular language can be "pumped" to generate larger accepted words. By *pumping* we mean simply repeating cyclically. If one can produce words of a language that cannot be built by pumping then one has shown that the language is not regular. For a string $z \in \mathcal{L}$ we denote its length by $|z|$. The statement of the pumping lemma is: For any regular language, $\mathcal{L}$, there exists a constant $n(\mathcal{L})$

such that for any $z \in \mathcal{L}$, $|z| \geq n(\mathcal{L})$, we can write $z = uvw$ such that:

$$a)\ |v| \geq 1$$

$$b)\ |uv| \leq n(\mathcal{L}),\ \text{and}$$

$$c)\ \forall i \geq 0\ ,\ (uv^i w) \in \mathcal{L}$$

and $n(\mathcal{L}) \leq \|V\|$, where $\|V\|$ is the number of states in the minimal DFA recognizing $\mathcal{L}$.

Finally the class of grammars, equivalent to FA in the aforementioned sense, are called regular grammars. Regular grammars are characterized by productions of the form

$$A \rightarrow uB$$

$$A \rightarrow B$$

where A and B are variables and u is a string of terminals. If the production rules have this form the grammar is also called right linear and if they have the form

$$A \rightarrow Bu$$

$$A \rightarrow B$$

it is also called left linear.

Examples of Finite Automata

We would now like to consider some examples of finite automata or regular languages that will come up often in this thesis because they are particularly simple and structurally interesting. For all the examples we shall use the binary alphabet $\mathbf{A} = \{0, 1\}$.

Example: Coin Flipping

The first example is just the machine corresponding to the process of coin flipping (biased or unbiased in this case, as we shall only by concerned with legal sequences and not their relative frequency of occurrence in this chapter). The language was given above and simply consists of the set of all binary strings $\mathcal{L} = \mathbf{A}^*$. The graph representing the DFA in this case is the graph with one vertex and two edges shown in figure 20. In our coin flipping example $\mathcal{F} = \phi$, i.e., the empty set.

Figure 20 The labeled digraph for the coin flipping process.

Example: The Period Three Process

Our second example is that of a non-stochastic model we call the period three system. We show the graph representing its DFA in figure 21 and

$$\mathcal{L} = \{\epsilon, 0, 1, 01, 10, 11, (011)^n, (101)^n, (110)^n\}$$

$$\mathcal{F} = \{00, 001, 010, 100, 111, ...\}$$

Notice that despite the fact that this process is not stochastic, there is branching in the transient part of the graph. This phenomenon represents the fact that to predict the next symbol in the sequence requires that we know which phase of the period three sequence we are in. So we first have to "phase lock" before we can make optimal predictions. For example consider a period three process that produces the binary string ...011011011011... or $(011)^*$. If a first observation of the process yields a 1 the phase of the process is not determined in the sense that future observations (including the observed 1) could yield either $(110)^*$ or $(101)^*$. Since this type of "inferential phase locking" is important in stochastic systems as well as periodic ones we shall refer to it in the general case as *stochastic phase locking*.

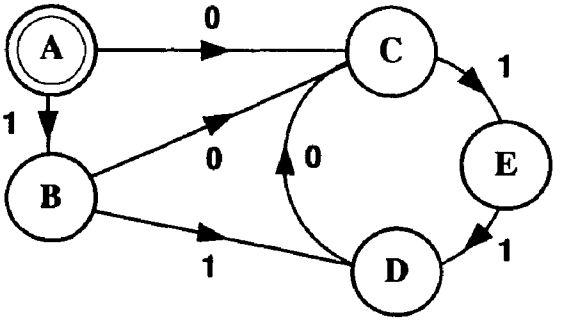


Figure 21 The labeled digraph for the period three process.

Example: The Golden Mean System

The third example, called the golden mean system, is an example of what ergodic theorists refer to as a subshift of finite type.[87] The language for the golden mean system is easy to describe, it is just the set of binary strings containing isolated $0's$, i.e., no strings of the form ...00... are allowed. We write the language for the golden mean system as

$$\mathcal{L} = \{\epsilon, 0, 1, 01, 10, 11, 010, 011, 101, 110, 111, ..\}$$

and

$$\mathcal{F} = \{00\}$$

The graph representation of its minimal DFA was shown in 19. This system is the topological equivalent of a Markov or finite memory process in the sense that for any legal string only a finite substring is necessary for the complete and unique specification of the set of symbols that are allowed on the next transition. It is therefore called a *topological Markov chain* or equivalently a *subshift of finite type* (SFT). For example any time a 0 occurs it must be followed by a 1 regardless of what preceded it. Any time a 1 occurs it may be followed by either a 0 or a 1, regardless of what preceded it.

Example: The Even System

The last example, called the even system, is an example of the more general class of systems that ergodic theorists call sofic systems and which is not a SFT.[87] It is a simple parity recognizing system; it only allows strings having even sequences of $1's$ bounded by $0's$. The graph representation of its minimal DFA is shown in figure 22 We have

$$\mathcal{L} = \{\epsilon, 0, 1, 00, 01, 10, 11, 000, 001, 011, 100, 101, 110, 111,$$
$$0000, 0001, 0011, 0110, 0111, 1000, 1001, 1011, 1100, 1101,$$
$$1110, 1111, ...\}$$

and

$$\mathcal{F} = \{01^{2n+1}0\}$$

The reason that the even system cannot be regarded as a topological Markov chain is that there is one "bad string" such that no matter how long a subsequence is observed the set of symbols which are allowed on the next transition cannot be uniquely determined. The "bad string", in this case, is the string of consecutive $1's$. There is no initial 0 in this string to determine whether or not any string of $1's$ has to be even. Therefore, whether or not a 0 is allowed on the next transition is not determined. This can be seen by noting that there are two paths in the graph corresponding to a string of consecutive $1's$ for which we are unaware of the preceding symbol. For one of the paths either a 0 or 1 is allowed but for the other only a 1 is allowed.

Although the even system has a simple and finite representation as a DFA there is a sense in which a minimal sort of infinite memory is required to uniquely determine its state in all cases, given some arbitrary string. A way of thinking about this property is in terms of the stochastic phase locking discussed above. For a topological Markov chain, or SFT, there always exists some minimal length of string (for any string of the language) that one needs to have observed, to determine uniquely which state one is in, i.e., to phase lock in the sense of being able to make optimal predictions. For a sofic system there exists strings which violate this criterion, e.g., the string of $1's$ for the even system. In fact one of the hallmarks of sofic systems, such as the even

system, is that they contain cycles in the transient part of their machines, which is equivalent to the statement that there exists strings that don't phase lock to recurrent modes.

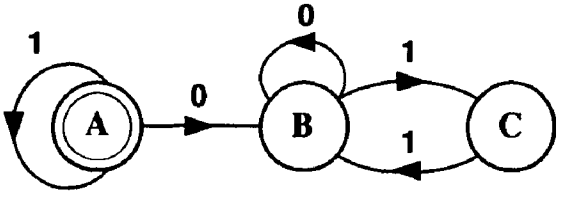


Figure 22 The labeled digraph for the even system.

This notion of stochastic phase locking determines what we can infer about the behavior of the process given what we have seen. As we have seen it reduces to our standard notion of phase locking in the case of a periodic system.

Pushdown Automata and Context Free Languages

Consider the hypothetical task of trying to decide whether some polymer of arbitrary length, but with specified length in any given instance, is reflection symmetric about the midpoint of the chain. Assume that when somebody approaches you with their polymer they specify the length L but that your machine must be able to check polymers of arbitrary length. Perhaps the most efficient type of machine one could devise to accomplish this task would be one that reads the first L/2 molecules coded as symbols and pushes (puts) the appropriate symbol on a stack. Then as it reads each of the second set of L/2 symbols it compares that symbol to the symbol on the top of the stack. If the symbols match it pops (pulls) that symbol off the stack and proceeds to read the next molecule/symbol in the polymer chain. If, having read the last symbol in the polymer chain, the stack is empty, the polymer is reflection symmetric about its midpoint, otherwise it is not. This procedure in fact corresponds to one of the common methods of determining the legality of strings in a context free language ,i.e., checking for the empty

stack condition. Note that since the machine must be able to give an answer for polymers of arbitrary length a finite automaton is not capable of this task.

Context free languages (CFL) were first developed by Chomsky [14] and their practical importance for use as lexical analyzers was soon realized by computer scientists. The reason for this is that they are perhaps the simplest languages allowing for a primitive form of recursion, i.e., nesting. It is clear that for the implementation of arbitrarily nested structures some form of infinite memory is required. However direct access to any memory element is not required, so the form of infinite memory in this case is more restricted than in the general case of random access memory. In fact the memory structures associated with the recognition of context free languages are stacks and the associated machines are called pushdown automata, which we shall define carefully below. A schematic representation of a pushdown automaton is shown in figure 23, in which a finite automaton reads (writes) from (to) an input (output) tape and has access to a stack tape.

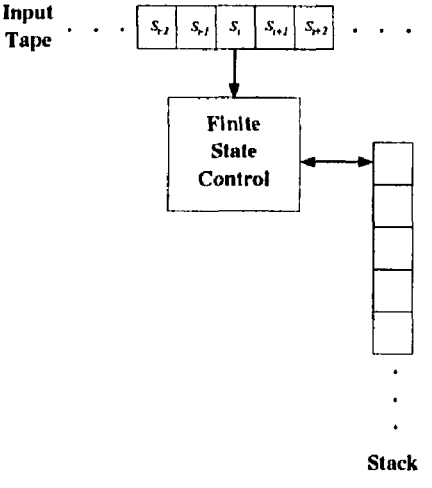


Figure 23 Schematic representation of a pushdown automaton with finite state control, input tape, and stack.

A *stack* is a type of memory for which only the last element stored is currently accessible, or last-in-first-out in computer science jargon. A commonly cited mechanical analog of a stack

is the spring supported plate stacker found in restaurants, where the last plate put on the stack is the only one available until it is removed at which point the second to the last plate is the only one available, etc. This is a natural model for physicists in that obtaining and storing information is usually thought of as occurring via local interactions, i.e., global information is processed piece by piece but, barring a radical quantum mechanical interpretation, global structure is maintained whether or not we have access to arbitrary elements of it. Here we are thinking of persistence in structure as memory.

We find it convenient to define context free languages in terms of the grammars that generate them. A *context-free grammar* $\Gamma = (\sum, \mathbf{T}, \mathbf{P}, \alpha)$ is a grammar as defined above with $\mathbf{P}$ a finite set of productions. The productions are applied to legal strings to obtain new, longer legal strings. The term context free refers to the fact that the application of the productions is independent of the position, i.e., context, in the string that they are applied to.

A *context-free language* $\mathcal{L}$ is a set of sequences consisting solely of terminals $t \in \mathbf{T}$ which can be obtained by the application of any set of productions to the start symbol α.

A simple example of a context free language is the set of all legal arithmetic expressions involving the operations of addition and multiplication.[43] The grammar is given by

$$\begin{aligned}\Gamma &= \left(\sum, \mathbf{T}, \mathbf{P}, \alpha\right) \\ &= (\{\gamma\}, \{+, *, \#, (,)\}, \mathbf{P}, \gamma)\end{aligned}$$

where # stands for any real number, or more generally any member of some field. The variable γ is just a placeholder and start symbol for the "derivation" of the expressions. The productions $\mathbf{P}$ are given by

$$1)\ \gamma \rightarrow \gamma + \gamma$$

$$2)\ \gamma \rightarrow \gamma * \gamma$$

$$3)\ \gamma \rightarrow (\gamma)$$

$$4)\ \gamma \rightarrow \#$$

By continued application of these productions such that the result is a set of terminals any legal expression involving addition and multiplication over the reals ,or any other field, can be derived. Consider an example of a derivation

$$
\begin{aligned}
&\gamma \\
&\Rightarrow \gamma * \gamma && \text{by 2)} \\
&\Rightarrow (\gamma) * (\gamma) && \text{by 3) twice} \\
&\Rightarrow (\gamma + \gamma) * (\gamma * \gamma) && \text{by 1) and 2)} \\
&\Rightarrow (\# + \#) * (\# * \#) && \text{by 4)}
\end{aligned}
$$

is a legal expression of this arithmetic language, where any four real numbers can be used for the $\#'s$.

Formally a *pushdown automaton* (PDA), the machine "dual" to a context free grammar, is a machine $M = \{V, \mathbf{A}, \Xi, T, v_0, Z_0, V_F\}$ where we have introduced the stack alphabet Ξ, i.e., the alphabet of symbols we push and pop from the stack. $Z_0 \in \Xi$ is the start symbol, or first state on the stack. T is a transition function $T : V \times \mathbf{A} \times \Xi \rightarrow V \times \Xi^*$. In other words given a current state, a tape symbol, and the current symbol on the top of the stack T gives a new state, possibly the same, and either adds a symbol to the stack from Ξ, removes a symbol from the stack, or does nothing to the stack, any of which result in some configuration of the stack.

We can define deterministic and nondeterministic PDA in a matter analogous to the way in which we defined DFA's and NFA's but unfortunately the class of context free languages recognized by deterministic PDA's is not the same as the class recognized by nondeterministic PDA's and the question of establishing minimality is a subtler issue than in the case of regular languages.

Just as for regular languages there is a pumping lemma to determine whether or not a given language $\mathcal{L}$ is context free. In this case two short substrings of a string that are close together can be pumped an arbitrary number of times and the resulting string must also be a member of the context free language. To be precise, the lemma [43] is: For any CFL, $\mathcal{L}$, there exists a

constant $n(\mathcal{L})$ such that for any $z \in \mathcal{L}$, $|z| \geq n(\mathcal{L})$ we can write $z = uvwxy$ such that:

$$a)\ |vx| \geq 1$$

$$b)\ |vwx| \leq n(\mathcal{L}),\ \text{and}$$

$$c)\ \forall i \geq 0\ ,\ (uv^iwx^iy)\ \in \mathcal{L}$$

Additional structure can be added to the grammars producing context free languages by adding, for example, a set of indices I to Γ, i.e., $\Gamma = (\sum, \mathbf{T}, I, \mathbf{P}, \alpha)$. This adds a mild form of context sensitivity to the languages produced by such grammars and the grammars themselves are reffered to as indexed grammars. The generalized productions $\mathbf{P}$ can be of one of three forms

$$a)\ \ W \rightarrow \alpha$$

$$b)\ \ W \rightarrow Yi$$

$$c)\ \ Wi \rightarrow \alpha$$

where $W, Y \in \sum$, $i \in I$ and $\alpha \in (\sum \cup \mathbf{T})^*$. When the productions are applied we may attach identical index sets to widely separated substrings of a larger string providing the aforementioned context sensitivity.

The utility of this additional structure will become apparent in part III when we prove that the symbolic dynamic language produced by the logistic map at the Feigenbaum point is in fact a language produced by an indexed grammar, specifically a type called a restricted indexed context free grammar (RICFG).

Linear Bounded Automata and Context Sensitive Languages

For completeness we mention briefly that intermediate in Chomsky's hierarchy between PDA's and Turing machines, to be defined below, lie a class of machine which is a restricted form of Turing machine. The restriction amounts to bounding the input tape to the machine, i.e., we add two symbols, left and right end markers. When reading the tape the machine cannot move past either of the end markers nor can it write over them. The machine, grammar and language in this case are referred to as linear bounded automata (LBA), context sensitive

grammars (CSG) and context sensitive languages (CSL). As we shall make no use of this class of automata or language in this thesis we turn now to a discussion of Turing machines.

Turing Machines and Recursively Enumerable Languages

The Turing machine is perhaps the simplest mathematical model of a general purpose computer. We discuss it briefly, mainly because it represents the pinnacle of the Chomsky hierarchy and because of its elegant representation of the deep notion of an effective procedure. It has been clear to mathematicians and computer scientists for a large part of the twentieth century that the notion of an effective procedure outlines the boundaries of the Platonic universe, i.e., that area of mathematics accessible to mathematicians, as discussed in chapter 1 in the context of Hilbert's program to formalize mathematics.[43] Subsequent to this work in the foundations of mathematics, physicists, in the context of quantum mechanics, began to discuss similar issues in terms of determining the inferential boundaries of the physical universe. [88] This thesis can in fact be understood as an alternate approach to such questions.

A particularly important type of Turing machine, sometimes referred to as the universal Turing machine, exemplifies the property of universal computation, i.e., anything that can be computed can be computed by a universal Turing machine. This property is called computation universal.

An early demonstration of the importance of Turing machines was in using them to demonstrate the existence of noncomputable functions. Put another way there are more functions than there are algorithms for computing functions. The proof of this statement is a simple generalization of Cantor's diagonalization argument showing that the real numbers are uncountable. This proof is a canonical example of the way in which Turing machines are used. It can be stated simply, without the full formalism necessary for the general use of Turing machines. Therefore we repeat the argument as given in [43].

Specifically note that for each computable function there is by definition a computer program that computes it. But computer programs are finite sets of symbols, so the set of computer

programs is countable. In particular consider all the functions from the integers into the set $\{0,1\}$. Assume this set of functions is countable, then we can label them with the integers, i.e., f_i. Now consider the function

$$f(n) = \begin{cases} 0 & if \quad f_n(n) = 1 \\ 1 & otherwise \end{cases}$$

But this function doesn't correspond to any integer, hence we have a contradiction, i.e., we can't assume that this set of functions is countable.

An important aspect of Turing's model is the identification of the notion of computable function with the previously described class of functions called the partial recursive functions. That this identification can be made is called the Church-Turing thesis and its importance lies in yielding an intuitive notion of what the term computable means. This intuitive notion is simply that given an unlimited amount of space (memory) to perform a computation, the computation will terminate, surely a reasonable expectation for a computable function. We stress however that the Church Turing thesis is just that, a thesis, and the connection between computable functions and partial recursive functions does not have a rigorous basis.

One of the beauties of the Turing machine model is that despite its computational power the general model class is relatively easy to specify. Formally a *Turing machine* (TM) is given by $M = \{V, \mathbf{A}, \tau, T, v_0, B, V_F\}$. The definitions correspond to those for FA's and PDA's except that τ now represents a set of tape symbols that the finite state machine can read. $B \in \tau$ is a special tape symbol called the blank symbol. $\mathbf{A} \subset \tau - B$ is the set of input symbols. $T : V \times \tau \rightarrow V \times \tau \times \{L, R\}$ is the next move function where L and R specify the direction a tape head moves along the tape. In particular a move of the Turing machine can be described in terms of the three actions: 1) change the state, 2) replace (or not) the tape symbol that the head is presently above, 3) move the head to the right, R, or left, L, one cell. We give a schematic representation of a Turing machine in figure 24.

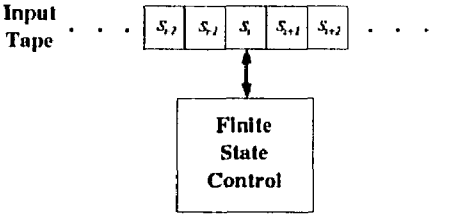


Figure 24 Schematic representation of a Turing machine with finite state control and input tape. Note that a Turing machines finite state control is more powerful than that of a pushdown automaton in that it can scan the input tape in either direction and write onto it as well as read it..

The set of languages recognized by Turing machines are referred to as recursively enumerable (re). The grammars that produce these languages are called unrestricted grammars. They are unrestricted in the sense that productions of the form $\eta \rightarrow \zeta$ where η and ζ are arbitrary strings of grammar symbols, with $\eta \neq \epsilon$, are permitted. We only mention re languages and unrestricted grammars for completeness. We shall not use them further.

Stochastic Automata

For the analysis of the behavior of nonlinear dynamical systems the topological behavior, i.e., the structure of the support of the dynamics in phase space, is well described, in conjunction with symbolic dynamics, by various of the automata we have introduced in this chapter. However, as we have discussed, the asymptotic statistical behavior, as well as finite time statistics and convergence rates are important in trying to obtain a complete characterization of the behavior of a system. Therefore in parts II, III, IV, and V we shall add probabilistic structure to the machines introduced in this chapter, primarily finite automata, by labeling edges in the graph with branching probabilities. We shall thus be able to calculate various statistical and informational quantities associated with the process characterized by the machine. We refer to these "generalized" machines as stochastic automata.

In part III we also introduce, as a conceptual tool, a slight variant of the Turing machine called the Bernoulli Turing machine (BTM). The primary purpose of introducing the BTM

is to characterize the generation of random numbers in a simple way by adding a register that generates them, and which we refer to as a random register. One could accomplish the same task by generating pseudo random numbers via a complex deterministic process, e.g., some linear congruential algorithm, for a standard universal Turing machine. BTM's therefore reflect a change in emphasis rather than a change in substance. We wish to emphasize that the cost of inferring nothing about a system, i.e., of drawing a random number, is relatively small in our scheme.

Conclusion

It is hoped that this brief survey of the Chomsky hierarchy has served to convey the way that increasing levels of complexity of particular systems can be characterized in terms of the resources, e.g., time and memory, necessary to recognize or simulate them.

Chapter 5 Thermodynamics and Phase Transitions

Section 1 Introduction

As we wish to exploit the relationship between dynamics, observation states and the thermodynamic limit we briefly review some of the basic elements of equilibrium thermodynamics that are crucial to an understanding of the arguments developed later. In particular the relationships we shall discuss are those between large volume and finite densities in thermodynamics and the corresponding quantities in dynamics: long time, and finite entropy and complexity rates.

As stated in the introduction, the need for this chapter stems from the authors perception that thermodynamics is often given short shrift in the modern undergraduate physics curriculum. Those with a good understanding of elementary thermodynamics, at the level of Callen's book [12], may certainly skip this chapter.

The great beauty of thermodynamics, and also its great mystery on first encounter, is its generality. The utility of a thermodynamic description was illustrated by the example given in chapter 1. As was noted there, not having to specify the co-ordinates and velocities of 10^{23} molecules appears to offer some advantage when identifying a macroscopic sample of fluid. To one weaned on the atomic view of matter it is certainly an amazing fact that for practical purposes a relatively few quantities need be specified to identify a macroscopic body and it's equilibrium state. Perhaps even more amazing is that based on a few, observationally motivated, postulates one fundamental equation relating extensive, i.e., volume dependent, quantities can be defined from which all other thermodynamic relationships flow. The properties of the fundamental equation lead to a geometric characterization of the principles obeyed by macroscopic bodies in equilibrium in terms of maximum entropy, minimum energy principles.

Of course, the price one pays for such generality is the lack of knowledge of microscopic detail. This is the main reason for the current view amongst scientists searching for the "fundamental laws of nature" that thermodynamics is of mainly engineering or historic value. It is currently felt that a complete description of nature in terms of microscopic detail at some "fundamental scale", e.g., the Planck scale, is almost in hand and that nothing new can be learned from an examination of macroscopic quantities. It is felt here, on the contrary, that there is in fact much that is useful in thermodynamic characterizations. Recent results in nonlinear dynamics and condensed matter theory suggest that there is semantic information in a system's rules of organization at any scale, that cannot be obtained from an examination of the "basic" interactions of the system's constituents. These rules of organization manifest themselves as constraints on the system's behavior as observed at some scale. We contend in this thesis that picking a scale of observation sets the level of averaging that one does in observing a given system. The averaging is over behavior at both larger and smaller scales. As a simple example think of a set of observations of barometric pressure. The averaging is over the effects of both large scale weather patterns and small scale molecular motions. In a standard thermodynamic analysis large scale averaging involves things like boundary conditions and stationarity assumptions while small scale averaging involves things like the assumption of molecular chaos and the exchange symmetry of the constituents.

In computational mechanics we argue that a choice of observation scale implicitly identifies an appropriate representation of the behavior of the system in terms of observational states and a stochastic process describing the dynamics on those states. Computational mechanics offers a way of determining this representation and the dynamics on the observational states thus obtained. For a simple system in thermodynamic equilibrium, averaging involves the determination of a single state, the equilibrium state, and the dynamics is is trivially governed by this fixed point. Computational mechanics allows us to generalize this situation and analyze systems which exhibit nontrivial dynamics on some set of observational states.

In section 2 we state the basic definitions of equilibrium thermodynamics that will be used later. In section 3 we discuss statistical mechanics from Boltzmann's original perspective, discussed in more detail in chapter 2, of trying to use it as the mechanical theory underlying thermodynamics. In particular we focus on the topics of ensembles, equilibrium fluctuations, and phase transitions.

Finally in section 4 there is a hueristic discussion of the renormalization group. The main reason for its inclusion is as a means of motivating some of the later discussions of scaling in dynamical systems. Almost all modern discussions of the attempt to characterize the set of trajectories of a dynamical system involves an examination of the scaling properties of this set of trajectories, i.e., what symmetries, if any, are obeyed by the set under changes of scale. The techniques used in these discussions are those originally developed by Kadanoff, Wilson [69] and others in the context of trying to understand phase transitions and referred to as renormalization group techniques.

Section 2 Basic Thermodynamic Definitions

As suggested by Callen [12] the underlying problem of equilibrium thermodynamics is the determination of the equilibrium state for a closed macroscopic system with a given set of parameter values. This can be accomplished by first assuming a minimal set of observation based postulates by means of which a *fundamental relation* relating the extensive, i.e., volume dependent, quantities can be derived. The form of fundamental relation we find most useful is in terms of the extensive thermodynamic entropy as a function of the extensive energy, volume and particle number

$$S = S(U, V, N)$$

Historically, the road to the postulates and the solution of the underlying problem were fairly tortuous. In retrospect however the postulates can be stated simply and abstractly. The gist of the postulates is that: 1) the entropy can in fact be characterized by the energy,

volume and particle number, 2) that special (equilibrium) states of S exist, 3) that the values assumed by unconstrained extensive parameters in these special states is such as to maximize S, and 4) that S is additive over subsystems. If these postulates are assumed, the functional form of the fundamental relation is determined and all else about a macroscopic system in equilibrium can be obtained from the fundamental relation. In particular a set of intensive parameters can be defined as the partial derivatives of the fundamental relation with respect to the extensive parameters. With the fundamental equation given as above, such that the entropy is the independent variable, we have

$$\begin{aligned} \frac{1}{T} = \beta &= \frac{\partial S}{\partial U} \\ -\frac{P}{T} = -\beta P &= \frac{\partial S}{\partial V} \\ -\frac{\mu}{T} = -\beta \mu &= \frac{\partial S}{\partial N} \end{aligned}$$

where T is the *temperature*, P is the *pressure*, and μ is the *chemical potential.*

The extensive quantities appear here as independent variables and the intensive quantities as derived quantities but there is a certain symmetry that underlies the relationship between the extensive and intensive quantities that is hidden by this approach. Also, it is hard not to think of temperature as a fundamental quantity as its definition is virtually inseparable from that of thermodynamic equilibrium. In fact experimentally it is the intensive parameters, e.g., temperature and pressure, that are often independent and under the control of the experimenter so in general it is desirable to have a way of transforming between various representation in terms of which the various intensive and extensive quantities appear as independent variables. This is accomplished by means of Legendre transforms, familiar from mechanics in terms of going from a Lagrangian representation to a Hamiltonian one. But perhaps even more useful, in our context, than the utility of these transforms is that they lead to quantities that are easily interpretable in very general settings. These quantities are typically referred to as *thermodynamic potentials*. The most useful for us are the Helmholtz and Gibbs free energies. When starting

with the fundamental relation

$$S = S(U, V, N)$$

the functions we obtain by transforming from S as a function of the extensive parameters to S as a function of the intensive parameters are known as Massieu functions. These can be written in terms of the basic thermodynamic potentials as long as one is careful to keep track of functional relationships. In particular we have

$$S(\beta) = S(U) - \beta U = -\beta F$$

$$S(\beta, \beta P) = S(U, V) - \beta U - \beta PV = -\beta G$$

where F and G are the standard *Helmholtz* and *Gibbs free energies* respectively. The above postulates and the use of Legendre transforms [12] determine that

$$\beta = \frac{\partial S}{\partial U}$$
$$\beta P = \frac{\partial S}{\partial V}$$

If we rewrite the expression for the Helmholtz free energy in the more familiar form

$$F = U - \frac{1}{\beta} S$$
$$= U - TS$$

we can generalize the standard thermodynamic interpretation of F. Thermodynamically F measures the amount of work that may be extracted from a macroscopic system and reflects the experimental fact that this quantity decreases with temperature. It characterizes the trade off between the need a system feels to minimize its energy and maximize its entropy. But one can make a more general interpretation in terms of a tradeoff between minimizing complexity and maximizing randomness. We shall see that in these terms the notion of a generalized free energy serves as a useful concept in computational mechanics. This more general notion is already hinted at in information theory in Rissanen's definition of stochastic complexity, and in dynamical systems by Legendre transforms of the various entropies, dynamical entropies, and dimensions.

Section 3 Thermodynamics from Statistical Mechanics

Introduction

When the atomic view of matter began to become entrenched, in the latter part of the nineteenth century, Maxwell, Boltzmann, Gibbs and others invented statistical mechanics in an attempt to derive the grand principles of thermodynamics from the basic mechanical laws governing the motions of atoms. The success of this program has been controversial since its inception, and according to some the various controversies have not yet been resolved.[50] Statistical mechanics, however, has certainly proved to be an extremely fruitful theory in its own right, in both the classical and quantum domains. As one of the primary concerns of computational mechanics is to study the transition from microscopic to macroscopic for general nonlinear systems we turn next to a brief discussion of the basic definitions of statistical mechanics.

Ensembles and Partition Functions

It is common in modern treatments of statistical mechanics to begin by deriving (or just stating) the expression for the canonical ensemble. This is a sensible approach as the canonical ensemble describes a macroscopic system in contact with a heat bath, a ubiquitous situation known as a closed system. But given the slightly historical approach we are taking we wish to first describe the situation of an isolated system, i.e., one that is thermally as well as mechanically isolated.

If, as Boltzmann and Gibbs did, we start by considering our microscopic system to be Hamiltonian we note that in general the only integral of the motion we can generally assume is the total energy. Therefore our Hamiltonian system, containing say $\sim 10^{23}$ point particles moves on a $6 \times 10^{23} - 1$ dimensional hypersurface in phase space. We wish to first determine the equilibrium probability distribution $\mu(x)$ on the energy surface from which we can obtain equilibrium values for all other quantities of interest. Taking his cue from thermodynamics

Gibbs obtained $\mu(x)$ by defining an additive, positive definite function on the energy surface that is required to be a maximum in equilibrium, i.e., an entropy

$$S(U) = \int_{U<H(X)<U+\Delta U} \mu(x) \log \mu(x)\, d\mu(x)$$

where, for convenience, we actually integrate in a thin energy shell $U \rightarrow U + \Delta U$ in phase space. Maximizing $S(U)$ subject to the constraint

$$\int_{U<H(X)<U+\Delta U} d\mu(x) = 1$$

via the technique of Lagrange multipliers [73] yields

$$\mu(x) = \begin{cases} \frac{1}{\Omega(U,\Delta U)} & U < H(x) < U + \Delta U \\ 0 & otherwise \end{cases}$$

where $\Omega(U, \Delta U)$ is the volume of the energy shell, i.e., we obtain a uniform distribution on the energy shell.

Next we consider closed systems, i.e., those that are mechanically but not thermally isolated. For convenience and later use we switch to a description in terms of systems with a discrete set of states. We note that the energy is in general not constant for a system in thermal contact with a heat bath and the system has a range of energy fluctuations. The standard procedure for obtaining the canonical distribution, as it is called for closed systems, is the same as was used for the microcanonical distribution. We begin with Boltzmann's definition of the statistical mechanical entropy, maximize it under the proper constraints using Lagrange multipliers, and obtain as the probability of the i^{th} state

$$p_i = \frac{e^{-\beta E_i}}{\sum_j e^{-\beta E_j}}$$

where E_i is the energy of the i^{th} state, and $e^{-\beta E_i}$ is reffered to as the Boltzmann factor of the i^{th} state. The all important quantity

$$Z = \sum_j e^{-\beta E_j}$$

is the *partition function* and from it important thermodynamic quantities can be calculated. Of particular importance is the relation which can be derived, between the partition function and the statistical mechanical Helmholtz free energy

$$-\beta F = \ln Z$$

Recalling the discussion, in the last section, of the interpretation of the free energy we see that the above relation gives us a means of obtaining macroscopic, extensive quantities from microscopic state descriptions, at least if we are in thermal equilibrium and assume that the thermodynamic and statistical mechanical free energies approximately specify the same quantity. For a careful discussion of this last point see [6].

We shall utilize the statistical mechanical relationship between the free energy and partition function in combination with the thermodynamic relationship between free energy, energy and entropy in computational mechanics. We shall do so by regarding $-\beta F$ as a Legendre transform of the entropy, with entropy as the independent variable in a fundamental relation. From this we shall find that the conditions for a phase transition, stated in the next section, apply to statistical mechanical as well as thermodynamic quantities. As noted in that section there is no guarantee that statistical mechanical derivations lead to concave fundamental relations, i.e., fundamental relations that describe stable equilibria. In fact until the advent of renormalization group treatments in the 1970's the theory of phase transitions was strictly thermodynamic and hence phenomenological. There was no statistical mechanical interpretation of concave or nonconcave fundamental relations. Computational mechanics also indicates that nonconcave relations represent "interesting" behavior in nonlinear systems.

For completeness we mention that a similar set of relationships can be derived [12] in the case of open systems, i.e., those that may exchange particles as well as energy with their surroundings. In this case the ensemble is called the grand canonical ensemble and the grand

partition sum is given by

$$\mathcal{Z} = \sum_j e^{-\beta(E_j - \mu N_j)}$$

where μ is the chemical potential and N_j is the occupation number of the j^{th} state. It should also be noted that the three representations can be shown to be equivalent in the thermodynamic limit. That is in the limit as volume and particle number become infinite, and such that the density remains finite.[12],[73]

Fluctuations and Large Deviations About Equilibrium

While the fundamental problem of equilibrium thermodynamics may be to determine the equilibrium state of a system, an equally important problem is to determine the magnitude and frequency of fluctuations about the equilibrium state. In fact the very notion of equilibrium is linked to the idea of the stability of the equilibrium state as we shall discuss in more detail in the next section. For now it is sufficient to note that the magnitude of the fluctuations in fact determine the utility of knowing the equilibrium state. If the fluctuations are large and often, one will find the system so rarely in it's equilibrium state that a determination of the particular equilibrium state is not terribly useful. From a statistical point of view this is the case for which higher order moments of the distribution on phase space, whose mean gives the equilibrium state, are important.

One of the first and most simple theories of fluctuations about equilibrium is due to Einstein.[73] To discuss his theory consider a closed system with energy U. The entropy of this system in the thermodynamic limit (recall the equivalence of the microcanonical and canonical distributions in the thermodynamic limit) is approximately $S(U) \approx \log \Omega(U)$ where $\Omega(U)$ is the number of available microstates at energy U and where we have set Boltzmann's constant equal to unity. Now if we make the approximation that that $\sum_U \Omega(U) \approx \Omega\left(\hat{U}\right)$ where $\hat{U}$ is the energy

at maximum entropy, i.e., equilibrium, then the relative probability of any state is given by

$$\begin{aligned} p(U) &\propto \frac{\Omega(U)}{\sum\limits_{U} \Omega(U)} \\ &\approx \frac{\Omega(U)}{\Omega\left(\hat{U}\right)} \\ &= e^{\left(S(U)-S\left(\hat{U}\right)\right)} \\ &= e^{\Delta S(U)} \end{aligned}$$

Note that since $S\left(\hat{U}\right)$ is the maximum of $S(U)$, $\Delta S(U)$ is negative. The point here is that the probability of a fluctuation about the equilibrium energy $\hat{U}$ decreases as an exponential function of the distance from equilibrium.

The above argument can be made rigorous using the theory of large deviations.[11],[28] One can determine under what assumptions an exponential decay of the amplitude of fluctuations about some mean value will occur. We can restate the above result in the more general notation of large deviation theory. Consider some sequence of random variables $\vec{x} = (x_0, x_1, ..., x_{n-1})$, e.g., a time series of values drawn from some distribution. Consider also some statistic Y_n associated with this sequence. Usually Y_n is a partial sum, such as the mean $Y_n = \frac{1}{n}\sum\limits_{i=0}^{n-1} x_i$. Assume that the series Y_n is convergent, i.e., $Y_n \underset{n\to\infty}{\to} m$. Then, given appropriate assumptions about $\vec{x}$, the distribution(s) its values are drawn from, and Y_n the main result of large deviation theory can be stated as

$$Pr(Y_n > |x - m|) \underset{n\to\infty}{\simeq} K(m,n)e^{-nI(x-m)}$$

This says that the probability of a fluctuation about the mean decays as an exponential function of the distance from the mean. $I(x)$ is called the large deviation rate function and is the equivalent of a thermodynamic density. When multiplied by n we obtain the equivalent of an extensive thermodynamic quantity, e.g., the thermodynamic entropy or a difference of entropies. Taking the large n limit is equivalent to taking the thermodynamic limit. Therefore the function $\Delta S(U)$, considered in the thermodynamic limit, is the thermodynamic version of the "extensive"

quantity $nI(x)$ in the large n limit. It is probably clear by now that the term large deviation theory results from the fact that we use it to estimate the likelihood of large deviations from the mean of some process.

Another important large deviation rate function is the spectrum of singularities $f(\alpha)$ studied recently in the geophysics, fluid dynamics, and dynamical systems literature.[57], [40] This function is usually described as the fractal dimension of a point set comprising the support of that part of a distribution function that diverges with index α. It can however, be derived as the scaling function for an entropy–like quantity in a manner completely analogous to the way in which we obtain the Renyi dimensions from the Renyi entropies in chapter 3. Put another way $f(\alpha)$ measures how likely it is for a system to be found in the regions of state space for which the invariant distribution diverges with index α. We shall in fact see that a very general characterization of the large deviation rate functions associated with a dynamical system can be obtained from a symbolic dynamical treatment combined with the reconstruction techniques which form the core of computational mechanics.

We note that the conditions under which a large deviation rate function fails to exist are just those conditions under which a system is dominated by large fluctuations.[28] This happens for example at a second order phase transition and as a result the systematic explication of the conditions under which rate functions fail to exist gives a useful characterization of the onset of phase transitions. We shall use this characterization to describe what we refer to as "typical" behavior in dynamical systems.

Phase Transitions

Later we shall make the analogy between the coexistence of different phases of a material, e.g., at a phase transition, and various coexistent states of automata describing dynamical systems so we briefly review the phenomenology of equilibrium, unstable as well as stable, and phase transitions.

For an observable system to be in equilibrium it is necessary that this equilibrium be stable. Stated another way, in terms of Le Chatelier's principle, the equilibrium state of a system should be such that equilibrium is restored if the system undergoes a spontaneous or induced fluctuation away from equilibrium. This simple requirement places an enormously useful geometric constraint on the fundamental relation describing a thermodynamic system.[12] This constraint is that in the entropy representation the entropy must be a concave function of the other extensive parameters. Any deviation of the fundamental equation from concavity is in fact an indication of the instability of a particular phase of the system. In fact nonconcavity implies the existence of multiple energy states with the same entropy; if this entropy is the maximum entropy then nonconcavity implies the existence of multiple equilibrium states. This is what occurs at a phase transition. It is interesting to note that "fundamental like" relations can be obtained empirically, or from statistical mechanics via combinatorial analysis and that, in general these are not concave. An equivalent equilibrium function can be obtained by considering the concave hull of the "fundamental like" relation.

In the context of computational mechanics we shall find that an indication of complexity in a system is given by nonconcavity of some fundamental relation. This is intuitively plausible, as one would expect a system which has a multiple set of energetically equivalent equilibrium states available to it, to be more complex than one with a single equilibrium state available to it. Common examples of this in physics are solids or spin glasses with multiple metastable states, ground states or vacuum states separated by large potential barriers.

A convenient way to think about phase transitions is in terms of the thermodynamic potentials, in analogy with the mechanical potential. The different phases of a system are associated with the local minima of the thermodynamic potential under consideration. For example, if we consider a piece of a fluid as a system, then we can consider the Gibbs potential of this piece of fluid, as it can exchange both matter and energy with the surrounding fluid. As a function of volume, different local minima correspond to phases of different density.

As we vary the parameters, e.g., temperature, of a system it can occur that two or more local minima occur simultaneously as ground states, i.e., both have the same energy and this is the lowest energy for the system. The phases associated with those minima can coexist as stable states and the system is at a phase transition with respect to the parameters which have been varied. If the local minima are asymmetric in the thermodynamic potential, away from the transition point, the transition is a first order transition and the different phases correspond to distinct regions of parameter space. The fact that the different phases correspond to different regions of phase space leads to a discontinuity across the transition in various thermodynamic quantities such as the heat capacity and compressibility. These differences are referred to as latent quantities. We shall define a latent complexity based on the analogy with standard thermodynamic phase transitions.

If, on the other hand the separate local minima are born together with the same value of the thermodynamic potential the phase transition is of second order and the appearance of a particular phase occurs via spontaneous symmetry breaking.

As defined above, the temperature is the slope of the entropy. It is a function of the energy, in the entropy representation of the fundamental equation. From the above discussion it is clear that a minimal requirement for the coexistence of two phases in equilibrium be that they have identical values of this slope, or

$$\beta_1 = \beta_2$$

Similarly the two states must have the same value of the free energy, or interpreting the statistical mechanical free energy as essentially a thermodynamic potential, this is equivalent to

$$\begin{aligned} p_1 &= \ln Z_1 \\ &= \ln Z_2 \\ &= p_2 \end{aligned}$$

and hence

$$-\beta_1 \ln Z_1 = -\beta_2 \ln Z_2$$

We shall make extensive use of this relation in our discussion of "inferential equilibria" and complexity.

Section 4 The Renormalization Group Treatment of Phase Transitions

We turn now to a brief discussion of the renormalization group. First we describe some of the important phenomenology of second order phase transitions which the renormalization group was able to explain and constituted the primary reason for its development.

One of the striking characteristics of phase transitions, at least those of second order, is the power law divergence of various quantities, e.g., specific heats and magnetization, at the critical parameter value. Just as important is the power law decay of the correlation function of a system at the critical point, indicating that in terms of calculating averages the central limit theorem does not hold. It only holds in situations of exponential decay of correlations. Power law decay of the correlation function in turn implies that the system is dominated by fluctuations on all scales. By the 1970's these power laws had been well catalogued experimentally. It was a quite remarkable fact that systems with widely different microscopic dynamics exhibited the same critical exponents, i.e., the powers in the power laws, and it was said that systems with the same critical exponents belonged to the same universality class. Although phenomenological laws relating the various critical exponents had been discovered and could be derived using the assumption that the underlying thermodynamic potentials were homogeneous functions of their arguments, a "fundamental" theory of phase transitions was still lacking.[69]

Wilson, aided by a heuristic scheme of Kadanoff reffered to as decimation and blocking transformations, was perhaps the first to rigorously apply renormalization group techniques to the study of phase transitions. However similar techniques had originally been developed, in what Wilson referred to as the "ancient version" , by quantum field theorists and scientists working on the theory of the Kondo effect, which is a description of magnetic impurities in metals at low temperature.

The primary difference between the "ancient" and Wilson's modern version of the renormalization group lies in Wilson's use of Kadanoff's intuitive picture of the transformation of a discrete spin system by block spin averaging and decimation, rather than in the abstract attempt to determine how to take the infinite limit of some cutoff parameter, e.g., the reciprocal of a minimum length scale in the problem. Wilson turned Kadanoff's picture into the well defined renormalization group operation[27].

We will also find the technique of block averaging over a discrete set of elements a useful one in computational mechanics. There we deal with a finite set of partition elements of some phase space observed at a discrete set of measurement intervals. The language of spin systems is easily utilized here by making the analogy between the lattice of a spin system and the discrete space-time points of the dynamical system plus measuring apparatus. Also analogous are the set of possible spin values and the set of partition element labels. Considering limits of the block averaging operation helps define questions about when averaging of a microscopic description is useful, questions that are central to computational mechanics.

The intuitive basis for the notion of spin blocking and decimation as applied to the critical phenomena of spin systems comes about from a consideration of the spatial correlations of the spin systems. To determine spatial correlations we usually define an order parameter which is a quantity that characterizes the phases of a system. Examples are the magnetization of a spin system and the density of a fluid. For a system undergoing a second order phase transition it is well known, theoretically and experimentally, that fluctuations in the order parameter are correlated over a large set of spatio-temporal scales. A lovely demonstration of this is the critical opalescence of certain fluids at a phase transition where light scatters off of density fluctuations on many scales. The measured and theoretical correlation lengths of such systems diverge as a power law in the distance from the critical parameter, e.g., temperature, as described above.

---

[27] Since averaging is involved, the operators actually comprise a semigroup as opposed to a group.

Since the correlation length is infinite for a given system at a critical point then it should remain unchanged under a blocking transformation, i.e., it should remain infinite under such a blocking. This is the case for Wilson's renormalization group operator. The correlation length of a system at its critical point is invariant under a renormalization group transformation.

What Wilson showed formally was that if you start with some discrete spin system described by a parametrized Hamiltonian, e.g., parameterized by the temperature and interaction strengths between a spin and the other spins of the lattice, then apply the spin blocking and decimation technique what you obtain is another Hamiltonian. This new Hamiltonian is described by some new set of parameters, not necessarily the same number of parameters as the original set. For example if you begin with a Hamiltonian containing only nearest neighbor interactions, the renormalization group operator does not in general yield a new Hamiltonian with only nearest neighbor interactions; general sets of interactions must be considered. It is easy to see, intuitively, that application of the renormalization group operator thus describes a path through the space of all Hamiltonians. This is not to say that this path is easy to describe in any particular situation!

In the space of all Hamiltonians the Hamiltonian describing a system at its critical point is a fixed point under the renormalization group operation, as we have just argued. But consider systems with a finite correlation length, including systems in the neighborhood of the critical Hamiltonian. These get renormalized to a system with zero correlation length under continued blocking and decimation of the lattice so the fixed point is an unstable fixed point[28] For arbitrary systems there is no reason to believe that only single fixed points exist and it is expected that complicated sets of fixed points are in general the case.

We also note that a hypersurface in the space of Hamiltonian systems is described by any fixed value of the temperature and in particular, at the critical temperature, the surface is called the critical surface, on which, obviously, the critical point lies, By adjusting the temperature of the system we can describe another important path through the space of Hamiltonians called

[28] This is as opposed to descriptions of self organized systems in the recent dynamics literature [89] for which it is claimed that the critical point is stable or attracting in some sense.

the physical line, for the obvious reason that this is an idealized version of the path that an experimentalist would follow. The importance of the critical surface is that it is what nonlinear dynamicists would refer to as the stable manifold of the critical point; once you are on the critical surface all renormalization group transformations lead to the critical point, which is consistent with the fact that at the critical temperature, i.e., on the critical surface, all renormalization group operations leave the correlation length infinite. So in general the fixed point(s) of a Hamiltonian spin system are what is (are) reffered to as hyperbolic, which was briefly described in the last chapter and means that the fixed point(s) has (have) a stable and unstable manifold. If you begin on the stable manifold, i.e., the critical surface, as stated above, renormalization moves you toward the fixed point, otherwise it moves you away, asymptotically on the unstable manifold.

As we discussed above the general utility of the renormalization group is in terms of describing universality classes in critical phenomena. The technique itself offers the possibility of transforming, via the renormalization group, from a system under study that is not solvable to one that is. For the study of phase transitions this usually involves studying small perturbations about the critical point.

In computational mechanics we attempt to extend the renormalization scheme to describe not only the divergence of correlation functions but to try to demonstrate the scaling structure of the particular blocking schemes associated with particular nonlinear systems. In part III we use these techniques to show explicitly that the "grammar" of the orbit structure of the logistic map at the period doubling accumulation point is what formal language theorists refer to as restricted context free.

Finally we describe the decimation technique, related to Kadanoff's technique, which we use on the graphs that we construct. The technique consists simply of merging graph vertices that lie along a deterministic chain, i.e., a chain with no branches. The term used in the graph theory literature for this operation is dedecoration. The detailed specification of decimation is as follows: 1) if a vertex has only one edge leaving it, that vertex is merged with the vertex

to which this single outedge leads, 2) The label of the single edge that was deleted in step 1) is concatenated onto the symbol or string of symbols labeling every edge that entered the edge deleted in step 1), 3) having begun the process at any deterministic transition, we continue the process until there are no or only one deterministic transitions left. A simple, one step, example of this operation is shown in figure 25 on a system, called the golden mean system discussed in chapters 2 and 4.

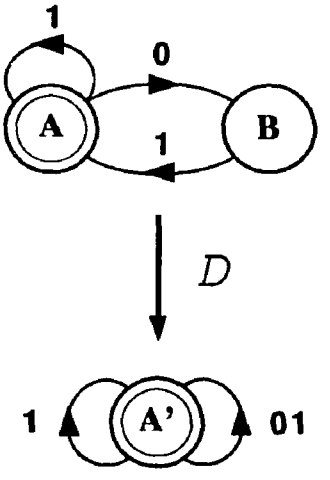


Figure 25 Decimation, in one step, for the golden mean system.

Under this operation periodic graphs, of period p, become graphs with a single vertex and a single edge labeled by the p length sequence constituting the basic periodic sequence. Similarly graphs with branching reduce to simpler forms. In particular there is a large class that reduce to a graph with one vertex and a set of edges corresponding to a set of sequences which can be thought of as being drawn randomly on consecutive trials. So, similar to renormalization group transformations on spin systems, unless there is an irreducibly infinite set of vertices, corresponding to spin systems with infinite correlation length, successive decimations drives us towards simpler and simpler systems. This technique is used in the computational mechanical analysis of the logistic map in part III.

Chapter 6 Summary of Part I and a Look Ahead

Section 1 Introduction

In part I of this thesis we have argued that new tools are necessary for modeling and predicting the behavior of complex nonlinear physical systems. A careful analysis of what is required of these tools suggests a statistical, in particular Bayesian, classification based approach which we called computational mechanics.

We tried to demonstrate that various techniques currently used in physics, e.g., renormalization group techniques, as well as techniques from ergodic, information and computation theories provide a natural basis for the development of such tools. However these techniques are neither complete, in and of themselves, nor well integrated in terms of the task at hand. In the next section we review why it is felt that these techniques are useful but incomplete for our purposes and how we intend to augment and integrate them. Next we briefly summarize the results of parts II, III, IV, and V. In the final section we outline the current view of what remains to be done.

Section 2 Integration of the Various Techniques

Ergodic Theory

Ergodic theory provides the dynamicist with tools for the asymptotic statistical description of a system's behavior. In reviewing the classical ergodic hierarchy we found that some of the standard quantities, e.g., the entropies, used in ergodic theory, when they can be obtained, provide useful means of characterizing how random a given dynamical system is. These quantities tell us on the average how unstable a system is to perturbations, i.e., the greater the average instability the more random the system appears.

If one is trying to infer asymptotic quantities, like the entropies, directly from data many assumptions must be made about stationarity, ergodicity, and other properties of the source

in the absence of any specific model. This is essentially because the asymptotic quantities are only rigorously defined for infinite trajectories as opposed to the finite trajectories that constitute a data set.

On the other hand if a model of the source can be inferred from the data, one may in turn infer something about the convergence properties of statistical measures of the behavior of the source, based on the this model. The problem is that ergodic theory provides no obvious techniques for inferring a model from data. This is the essential reason that we turn to the methods of computation theory to augment the statistical methods of ergodic and information theories.

A question related to the inference of models from data arises; how hard is it to infer an appropriate model of the given source from the data? The answer to this question leads directly to a measure of the complexity of the source that is in some sense "orthogonal" to the entropy measures that quantify the randomness of the source. That the entropies do not provide measures of this type of complexity is indicated by the fact that if we are given a model, entropies can be calculated without any consideration of how hard that model would be to infer from a data set.

The above discussion may be phrased in dynamical systems language by noting that knowing the entropy or dimension spectrum of a given system in general says nothing about the topology of its solution manifold.

Computation Theory

Given the above limitations of an asymptotic statistical description we turn to computation theory in an attempt to obtain models from data as well as a quantitative measures of the difficulty of inferring a model from a data set.

Treating the combination of source and measurement as a computation allows us to ask what the complexity of such a computation is in terms of the resources necessary for carrying it out. A more suggestive way of characterizing the situation is to treat the set of observed

sequences produced by a source plus measurement instrument as a language. Then we ask how hard is it to learn the grammar of this language?

Chomsky's hierarchy categorizes a set of increasingly hard to characterize languages. To model a given data set one specifies a class of models from Chomsky's hierarchy. If this class is insufficient for a given source this is indicated by the divergence of some quantity that measures the resources necessary for characterizing a language. One then turns to the next level of the hierarchy and so on. Computational mechanics specifies just such quantities as we shall see in parts II, III, IV, and V.

There is a problem with this procedure however. To describe the problem we first note that the relative frequency of occurrence of a given trajectory or sequence is a direct result of the dynamics of the system under study. By quantifying only how hard it is to specify a particular set of observed sequences and ignoring their frequency of occurrence one thus ignores information about the dynamics. We therefore need to augment a purely computational description of the data set with statistical information. This we do by introducing various information measures on the models which are inferred from the data. In fact it is the full "spectrum" of information measures associated with a model, from a particular class, which we argue most completely characterizes the complexity of a source and measurement instrument.

In dynamical systems language, computation theory thus provides useful measures of the complexity of the topology of a system's solution manifold. This description must at least be augmented by the probability distribution over this manifold for a complete description of the system's behavior.

Computational Mechanics

Introduction

The above discussion implies that a combination of computational and statistical techniques yields the most useful description of a complex nonlinear dynamical system's behavior.

Therefore the remainder of this section consists of a fairly succinct and terse description of the methodology of computational mechanics, the attempt to integrate and extend the techniques that were introduced from ergodic, information and computation theories in the previous chapters. The reading in this section will probably be a little dry as we try to extract and record only the formal essence of parts II, III, IV, and V here. The reader may, at this point, wish to skip ahead to the those parts, as the formalism is there fleshed out with examples, then refer to this section for summary and/or review. The reader may also wish to look back at the examples in chapter 1, which are simple examples of the way in which these techniques are applied.

The Basic Structure of ϵ-machines

The first step is to obtain a data stream. We assume that the underlying process is governed by a noisy discrete-time dynamical system

$$\vec{x}_{n+1} = \vec{F}_{\vec{\mu}}(\vec{x}_n) + \vec{\xi}_n, \quad \vec{x}_0 \in \mathrm{M}$$

where M is the m-dimensional space of states, $\vec{x}_0 = \left(x_0^0, x_0^1, \ldots, x_0^{m-1}\right)$ is the system's initial state, $\vec{F}$ is the dynamic, $\vec{\xi}_n$ represents external time-dependent fluctuations, and $\vec{\mu}$ represents the set of parameters governing the dynamics. The exact states of the observed system are translated into a sequence of symbols via a measurement channel.[17] This process is described by a parametrized partition

$$P_\epsilon = \left\{ c_i : \bigcup_{i=0}^{k-1} c_i = \mathrm{M},\ c_i \cap c_j = \emptyset, i \neq j;\ i,j = 0,\ldots,k-1 \right\}$$

of the state space M, consisting of cells c_i of volume ϵ^m that are sampled every τ time units. A measurement sequence consists of the labels from the successive elements of P_ϵ visited over time by the system's state. This way a sequence of states $\{\vec{x}_n\}$ is mapped into a sequence of symbols $\{s_n : s_n \in \mathrm{A}\}$, where $\mathrm{A} = \{0,\ldots,k-1\}$ is the alphabet of labels for the k ($\approx \epsilon^{-m}$) partition elements.[19] The computational models reconstructed from such data are referred to as *ϵ-machines*.

Given the data stream in the form of a long *measurement sequence* $\mathbf{s} = \{s_0 s_1 s_2 \cdots : s_i \in \mathbf{A}\}$, the second step in machine inference is the construction of a *parse tree* $T = \{\mathbf{n}, \mathbf{l}\}$. The links $\mathbf{l}$ are labeled by the measurement symbols $s \in \mathbf{A}$. An *L-level subtree* T_n^L is a tree that starts at node n and contains all nodes below n that can be reached within L links. To construct a tree from a measurement sequence we simply parse the latter for all length L sequences and from this construct the tree with links up to level L that are labeled with individual symbols up to that time. We refer to length L subsequences $s^L = \left\{ s_i \cdots s_j \cdots s_{i+L-1} : s_j = (\mathbf{s})_j \right\}$ as *L-cylinders*. Hence an L level tree has a length L path corresponding to each distinct observed L-cylinder. Probabilistic structure is added to the tree by recording for each node n_i the number $N_i(L)$ of occurrences of the associated L-cylinder relative to the total number *N(L)* observed,

$$p_{n_i}^T(L) = \frac{N_i(L)}{N(L)}$$

At the lowest computational level ϵ-machines are represented by probabilistic labeled digraphs. Their topological structure is described by a *labeled digraph* $G = \{\mathbf{V}, \mathbf{E}\}$ that consists of vertices $\mathbf{V} = \{v_i\}$ and directed edges $\mathbf{E} = \{e_i\}$ connecting them, each of the latter is labeled by a symbol $s \in \mathbf{A}$.

To reconstruct a topological ϵ-machine we define an equivalence relation, subtree similarity, denoted $\sim$, on the nodes of the tree T by the condition that the L-subtrees are identical:

$$n \sim n' \text{ if and only if } T_n^L = T_{n'}^L$$

Subtree equivalence means that the link structure is identical as was shown trivially in the example in chapter 1. This equivalence relation induces on T, and so on the measurement sequence s, a set of equivalence classes $\left\{ \mathbf{C}_m^L : m = 1, \ldots, K \right\}$ given by

$$\mathbf{C}_l^L = \left\{ n \in \mathbf{n} : n \in \mathbf{C}_l^L \text{ and } n' \in \mathbf{C}_l^L \text{ iff } n \sim n' \right\}$$

A labeled digraph G is then constructed by associating a vertex to each tree node L-level equivalence class; that is, $\mathbf{V} = \left\{ \mathbf{C}_m^L \right\}$. Two vertices v_k and v_l are connected by a directed edge

$e = (v_k \rightarrow v_l)$ if the transition exists in T between nodes in the equivalence classes,

$$n \rightarrow n' : n \in \mathbf{C}_k^L, n' \in \mathbf{C}_l^L$$

The corresponding edge is labeled by the symbol(s) $s \in \mathbf{A}$ associated with the tree links connecting the tree nodes in the two equivalence classes

$$\mathbf{E} = \left\{ e = (v_k, v_l; s) : \; v_k \underset{s}{\rightarrow} v_l \text{ iff } n \underset{s}{\rightarrow} n'; \; n \in \mathbf{C}_k^L, n' \in \mathbf{C}_l^L, s \in \mathbf{A} \right\}$$

Transition probabilities $p(v|v';s)$ also are assigned to the graph's edges, where v and v' represent vertices and s represents the symbol labeling the edge that connects them. These probabilities may be assigned as follows. Label the tree nodes n^d with their distance d from the root node, i.e., the number of links traversed in getting from the root to n^d. For a given vertex v in the graph, search the corresponding equivalence class $\mathbf{C}_k^L$ for the node closest to the root, i.e., highest in the tree

$$\min_d n^d : n^d \in \mathbf{C}_k^L$$

There may be more than one. If there is one use it; if there are more than one choose one. Each branch in the tree leaving $\min_d n^d$ corresponds to a labeled edge in the graph leaving v. The transition probability of this edge is then assigned the branching probability of the corresponding branch in the tree. This branching probability is determined, as above by treating the node $\min_d n^d$ as the root of the subtree $T^{L-d}_{\min_d n^d}$ and using the above histogram technique.

The statistical structure of an ϵ-machine is given by a parametrized *stochastic connection matrix*

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

that is the sum over each symbol $s \in \mathbf{A}$ in the alphabet $\mathbf{A} = \{i : \; i = 0, \ldots, k-1; \; k = \mathcal{O}(\epsilon^{-m})\}$ of the *state transition matrices*

$$T_\alpha^{(s)} = \left\{ e^{\alpha \log p(v_i|v_j;s)} \right\}$$

for the vertices $v_i \in \mathbf{V}$. We shall distinguish two subsets of vertices. The first $\mathbf{V}_t$ consists of those associated with transient states; the second $\mathbf{V}_r$, consists of recurrent states. T_0 is called the *connection matrix* of G and describes the degree of connectivity (i.e., number of edges) between vertices. T_1 is called the *Markov matrix* and describes the probabilities of transitions between the states represented by the vertices $v_i \in \mathbf{V}$.

The *α-order total Renyi entropy*,[74] of the measurement sequence up to L-cylinders is given by

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

where the *partition function* is

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

with the probabilities $p(s^L)$ defined on the L-cylinders $\{s^L\}$. The *Renyi specific entropy*, i.e., entropy per measurement, is approximated [20] from the L-cylinder distribution by

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

$$\text{or } h'_\alpha(L) = H_\alpha(L) - H_\alpha(L-1)$$

and is given asymptotically by

$$h_\alpha = \lim_{L \to \infty} h_\alpha(L)$$

An ϵ-machine's structure determines several key quantities. The first is the stochastic DFA measure of complexity. The *α-order graph complexity* is defined as

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

where the probabilities p_v are defined on the vertices $v \in V$ of the ϵ-machine's l-digraph and for the ϵ-machines considered in this thesis are uniquely determined by the set of transition probabilities $p(v|v';s)$.

The entropies and complexities are "orthogonal" in the sense described above. The entropy is determined by the principal eigenvalue λ_α of T_α,

$$h_\alpha = (1-\alpha)^{-1} \log_2 \lambda_\alpha$$

and the complexity by the associated left eigenvector of T_α

$$\vec{p}_\alpha = \{p_v^\alpha : v \in \mathbf{V}\}$$

that gives the asymptotic vertex probabilities.

The most familiar special cases occur for the parameter values $\alpha = 0$ and $\alpha = 1$. For $\alpha = 0$ we have

$$H_0(L) = \log_2 N(L)$$

and

$$\begin{aligned} h_0 &= \lim_{L\to\infty} \frac{\log N(L)}{L} \\ &= \log_2 \lambda_{max} \end{aligned}$$

is the topological entropy, where λ_{max} is the principal eigenvalue of the connection matrix T_0. The natural measure of complexity in the topological case is given by

$$C_0 = \log |\mathbf{V}|$$

where $|\mathbf{V}|$ is the total number of vertices of the ϵ-machine.

For $\alpha = 1$ we have

$$H_1(L) = -\sum_{s^L \in \{s^L\}} p(s^L) \log p(s^L)$$

The metric entropy of an ϵ-machine is a measure of the asymptotic rate of change of information in the cylinder distributions and is given in the present context by

$$\begin{aligned} h_\mu &= \lim_{L\to\infty} \frac{H_1(L)}{L} \\ &= \sum_{v\in V} p_v \sum_{v'\in V} p\Big(v|v';s\Big) \log p\Big(v|v';s\Big) \end{aligned}$$

where the edge probabilities $p(v|v';s)$ and vertex probabilities p_v are as defined above. The natural measure of complexity in the metric case is given by

$$C_\mu = -\sum_{v \in \mathbf{V}_r} p_v \log p_v$$

The complexities introduced above can be used as a diagnostic for determining adequacy of model class. If, for example,

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

vanishes, then the noisy dynamical system has been identified. If it does not vanish, then c_α is a measure of the rate of divergence of the model size and so quantifies a higher level of computational complexity. In this case we attempt to find a finite description at some higher level, e.g., in the Chomsky hierarchy.

The Semigroup Structure of ϵ-machines

Additional structure is given by the minimal semigroup associated with an ϵ-machine. This structure allows one to distinguish between finite models which can be characterized as Markov processes, i.e., subshifts of finite type, and those that cannot, i.s. strictly sofic systems. A *semigroup* Γ is a set $\{g_i\}$ endowed with a binary operation which is associative, i.e., preserves translation invariance of the operation. In distinction to a group, a semigroup contains an absorbing element a such that

$$ag_i = g_i a = a \quad \forall g_i \in \{g_i\}$$

and

$$\exists g_i, g_j \text{ distinct from } a \text{ such that } g_i g_j = a$$

It is the absorbing element which serves to determine illegal sequences for an ϵ-machine. A semigroup for which one of the elements is an identity element e such that

$$eg_i = g_i e = g_i \quad \forall g_i \in \{g_i\}$$

is called a *monoid.* A semigroup is said to be *finitely generated* if there is a finite subset of semigroup elements

$$\{g_j\} \subset \{g_i\}$$

such that any element of the semigroup Γ is given by a product of elements of the subset $\{g_j\}$ The elements of $\{g_j\}$ are called *generators.* The semigroup naturally associated with an ϵ-machine is the matrix semigroup generated via matrix multiplication by the symbol transition matrices (or state transition matrices of order 0)

$$T_0^{(s)}$$

That the symbol transition matrices serve as generators of the semigroup makes it clear that we can represent the structure of the semigroup in terms of a semigroup graph G_Γ for which each semigroup element represents a vertex

$$v_i = g_i$$

and there is an edge between vertices v_i and v_j labeled with an s if and only if

$$g_i T_o^{(s)} = g_j$$

As in the general case for directed graphs G_Γ contains a transient set of vertices as well as a recurrent set of vertices and we shall see that the structure of the transient set is crucial for the complete classification of ϵ-machines.We note also that each element of the semigroup represents a set of legal sequences for the ϵ-machine, grouped according to unique sets of preceding and succeeding sequences.

The Fluctuation Spectrum of ϵ-machines

We also find it useful to determine the fluctuation spectrum of processes modeled by particular ϵ-machines. The central idea is to obtain useful "macroscopic" properties by various

averaging techniques on the population of "microscopic" sequences. The main method is to examine and make use of the scaling structure of sequence distributions as a function of the sequence length L both for finite L and as $L \rightarrow \infty$.

We would like to examine various quantities, such as the length dependent cylinder energy $\mathbf{U}_{s^L}$ of a particular cylinder s^L defined as $\mathbf{U}_{s^L} = \log_2 \frac{1}{Pr(s^L)}$ or the entropy of a set of cylinders with a given energyU, i.e.,$S(\mathbf{U}, L)$. $S(\mathbf{U}, L)$ is defined in terms of the distribution of energy over the "microstates" s^L which is summarized by grouping them into energy macrostates, or energy level sets,

$$\mathcal{S}^L_{\mathbf{U}} = \{\omega : \mathbf{U}_\omega = \mathbf{U}, \omega \in \mathcal{S}^L\}$$

Then

$$S(\mathbf{U}, L) = \log_2 \left\| \mathcal{S}^L_{\mathbf{U}} \right\|$$

$\mathbf{U}_{s^L}$ and $S(\mathbf{U}, L)$ are useful quantities in the case of finite L but generally diverge as $L \rightarrow \infty$ like extensive thermodynamic quantities or the "bare" quantities of quantum field theory. In part V we develop a thermodynamic formalism that allows us to average these quantities, taking account of significant fluctuations. This leads to finite quantities, which are the analogues of thermodynamic densities or the "dressed" quantities of quantum field theory.The asymptotic fluctuation spectrum is obtained from the fundamental relation for the intensive or "dressed" quantities

$$s = s(\mathcal{U}) = \lim_{L \rightarrow \infty} s(\mathcal{U}, L)$$

with

$$\mathcal{U} = \lim_{L \rightarrow \infty} \mathcal{U}_{s^L}$$

where we define the intermediate or "transtensive" quantities

$$s(\mathcal{U}, L) = \frac{S(\mathbf{U}, L)}{\log_2 N(L)}$$

$$\mathcal{U}_{s^L} = -\frac{\log_2 Pr(s^L)}{\log_2 N(L)}$$

The above extensive and transtensive quantities cane be calculated directly from cylinder histograms and the intensive quantities can be approximated from cylinder statistics at large L. These techniques are analogous to the standard methods used to calculate the $f(\alpha)$ spectrum for a given process.

One of the great benefits of using the techniques of computational mechanics is that we can obtain good estimates of the asymptotic quantities directly from ϵ-machines. To the degree that the ϵ-machine provides a good model of the process, asymptotic properties as well as cylinder statistics and fluctuations for cylinders of any length L can be calculated. Low probability events, i.e., large deviations, can be inferred, whereas if we were not trying to model the process and relied only on cylinder histogram statistics this would not be possible.

The means by which we obtain good estimates of the asymptotic quantities directly from ϵ-machines are twofold. In the first place we can calculate the energy of a given cylinder by using the set of edge probabilities associated with the graph. In particular $\mathcal{U}_{s^L}$ is completely determined by the frequency with which each edge of the ϵ-machine is visited for the given cylinder

$$\mu(s^L) = \prod_{i=0}^{L-1} p(s_i|v_i) = p_1^{c_1^L} p_2^{c_2^L} \cdots p_E^{c_E^L}$$

where $\left\{ p_i \equiv p_{v \to v'} : i = 0, \ldots, E-1, v, v' \in \mathbf{V}, s \in \mathbf{A} \right\}$ is the set of conditional transition probabilities associated with the edges $\mathbf{E}$ and c_i^L is the number of times that the given sequence visits the i^{th} edge. Then

$$\mathcal{U}_{s^L} = -\frac{\log_2 \mu(s^L)}{\log_2 N(L)}$$

We can calculate $S(\mathcal{U}, L)$, and hence the fluctuation spectrum, by considering the combinatorics of edge visitations for paths in the graph.

Secondly we can relate the energies and entropies of the process directly to the eigenvalue spectrum of the associated ϵ-machine. Specifically we derive the relation

$$\mathcal{U}(\beta) = \beta^{-1}(S(\mathcal{U}(\beta)) - \log_2 \lambda_\beta)$$

in terms of the parameter β. β is in fact the Rényi α discussed above where we have changed notation here to stress that the parameter is to be thought of as an inverse temperature. λ_β is the largest eigenvalue of T_β and $S(\mathcal{U}(\beta)) = h_\mu(S_\beta)$ can be calculated as the metric entropy of the "stochastisized machine" S_β given by

$$(S_\beta)_{ij} = \lambda_\beta^{-1} r_i^{-1} (T_\beta)_{ij} r_j$$

where $\vec{r}$ is the right eigenvector of T_β associated with λ_β.

Section 3 Summary of Parts II, III, IV, and V

Part II

In part II we briefly discuss the motivation for a definition of complexity, introduce the formalism laid out in the previous section and use it to obtain a "superuniversal" relationship between the entropies and complexities for various parameter values of the logistic map. This reflects our assertion about the general, i.e.. superuniversal, shape that such a plot should have. In this particular case superuniversality reflects the fact that the position of a system in the complexity-entropy plot is less a function of parameter than the characteristics of the system. For instance, all periodic systems are grouped together in the complexity-entropy plot regardless of where they are in the bifurcation diagram, i.e., the state space-parameter plot. In general systems with similar complexity-entropy values occur at widely separated parameter values.

Part III

In part III we discuss a particular set of values in the complexity-entropy plot discussed in part II, as a means of describing the structure of the generic period doubling route to chaos.

We treat the onset of chaos as a phase transition signaled by the divergence of the finitary complexity as $L \to \infty$.

In particular we show that the lower envelope of values in the complexity-entropy plot, in the positive entropy regime, corresponds to the period two band merging sequence of parameter values in the logistic map. To verify this we derive an analytic expression for this envelope and show that the generic band merging machines give values which lie on this curve.

Given the divergence of the finitary complexity at the critical point we conclude that the language is not regular there, i.e., the language is not representable in terms of a finite digraph. We use the structure of the language at the critical point to show that it can be represented by a type of language known as a restricted indexed contest free language and derive the structure of the machine which recognizes this language.

This is a useful development in that it shows that even for the "simplest" route to a divergent finitary complexity in the logistic map the language at the phase transition is non-regular. This suggests that for more complicated sequences of parameter values such as general sequences of Misiurewicz parameter values or those leading to the Fibonacci map the languages may be arbitrarily complicated at the critical point.

Part IV

In part IV we catalogue the semigroup structure as well as the entropies and complexities of some archetypal processes as a means of demonstrating the rich structure of the set of finitary processes. In particular the systems we shall refer to as strictly sofic systems introduce subtleties into the computational mechanics of finitary processes which involve the phenomenon we call stochastic phase locking. This phenomenon introduces a weak form of infinite memory into the modeling of data sets produced by strictly sofic processes which may have implications for describing the physics of such processes.

Part V

In part V we develop the thermodynamic formalism of ϵ-machines described briefly in the previous section. We calculate the fluctuation spectrum for some archetypal examples from cylinder histograms and directly from the ϵ-machines. The efficacy of the ϵ-machine representation is clearly demonstrated at least for these examples. We argue further however that the ϵ-machine representation itself is more fundamental than the derivative statistics we obtain from it. While such statistics are useful, interesting, and currently very popular, a coherent model of the process under study, as provided by computational mechanics, seems to provide a closer approximation to what we think of as understanding.

Section 4 Remaining Issues

The emphasis in this thesis has been to develop an inductive formalism for the analysis and modeling of complex nonlinear systems from the data produced by the observation of such systems. Much time was spent on developing the techniques and determining the consistency of the formalism. In parts II, III, IV, and V "toy" systems, like the logistic map whose properties have been explored by other means, were examined in keeping with the tradition of testing a formalism on systems whose properties are at least to some extent understood. As a result the main goal of analyzing data sets thought to have been produced by complex nonlinear systems remains. Apart from issues of computational efficiency much of the apparatus necessary for analyzing complex time series is in place and such analyses are planned for the near future on available astrophysical and geophysical data sets as well as more laboratory oriented data sets such as velocity time series from turbulent fluids.

There are remaining theoretical issues as well. One of the primary goals is to extend the formalism described in this thesis to spatially distributed systems. This is currently underway and described in [41]. As well as extending the formalism important issues are understanding

the more general sets of behaviors of the reconstructed machines expected in these extended cases, and the determination of whether Chomsky's hierarchy will be useful in this domain.

Another fundamental issue in establishing whether or not the types of hierarchical description of dynamical systems advocated in this thesis are really useful would be to find systems whose description demands the resources of higher levels in the Chomsky hierarchy. As discussed above there are indications that even the logistic map at particular parameter values will yield such systems, but at this point this is merely conjecture. There are sequences of parameter values in the logistic map as well as other systems, hopefully leading to a "phase transition", which can be studied systematically.

As alluded to above a fundamental issue in dynamics in general as well as in computational mechanics in particular is the determination of effective generating partitions for dynamical systems. If in fact the determination of a generating partition from data were straightforward for any system the techniques of computational mechanics would be, dare it be said, purely mechanical. This not being the case there is still some art involved in picking a "good" phase space partition. This is currently an active area of research.

We have focused to some extent on measures of complexity associated with nonlinear dynamical systems. As previously mentioned, there is currently a plethora of definitions for such measures in the literature. From a historical perspective many of these definitions will die and others will coalesce into the essential definition or definitions. In the mean time, it is felt that any attempts to scale this tower of Babel would expedite the historical process and would therefore be useful. In this spirit a review of the present situation is currently underway.

In parts IV and V we really only scratch the surface of the determination of structure for ϵ-machines. There are many properties of the semigroups which might be useful in certain contexts. Examples are cosets and normal sub-semigroups which might help to determine the architecture of the set of periodic orbits of the inferred symbolic dynamical system. As is muched discussed in the dynamics literature the structure of this set determines much about the dynamics

in a large class of systems. Further structure might be exploited by working out the transfer matrix formalism for the topological and stochastic Markov matrices associated with ϵ-machines. Certainly the trace of the iterated topological machine is directly related to the structure of the set of periodic orbits. Another reason for studying such structures is that if in fact they provide convenient ways of describing the set of periodic orbits in classical systems they will probably be useful in the context of quantum chaos because it is the set of periodic orbits that provides the sturdiest bridge between classically chaotic systems and their quantum counterparts.

Similarly a detailed study of the structure of the entropy vs. energy plots for ϵ-machines, as defined in part V, for a larger set of systems will perhaps help to determine general properties of phase transitions in nonlinear systems (either as a function of parameter or the more general transitions discussed in parts II and III) as well as nontrivial fluctuation spectra in complex nonlinear systems.

Bibliography

[1] R. L. Adler and B. Weiss. Entropy, a complete metric invariant for automorphisms of the torus. *Proc. Nat. Acad. Sci. USA*, 57:1573, 1967.

[2] V. M. Alekseyev and M. V. Jacobson. Symbolic dynamics. *Phys. Rep.*, 25:287, 1981.

[3] V. I. Arnold and A. Avez. *Ergodic problems of classical mechanics.* Benjamin, New York, 1968.

[4] C. P. Bachas and B.A. Huberman. Complexity and relaxation of hierarchical structures. *Phys. Rev. Let.*, 57:1965, 1986.

[5] H. Bai-Lin. *Elementary Symbolic Dynamics and Chaos in Dissipative Systems.* World Scientific, 1989.

[6] R. Baierlein. *Atoms and Information Theory; an Introduction to Statistical Mechanics.* W. H. Freeman, San Francisco, 1971.

[7] C. H. Bennett. Dissipation, information, computational complexity, and the definition of organization. In D. Pines, editor, *Emerging Syntheses in the Sciences.* Addison-Wesley, Redwood City, 1988.

[8] A. Bialas and R. Peschanski. Moments of rapidity distributions as a measure of short-range fluctuations in high-energy collisions. *Nucl. Phys. B*, B273:703, 1986.

[9] R. E. Blahut. *Principles and Practice of Information Theory.* Addision-Wesley, 1987.

[10] L. Brillouin. *Science and Information Theory.* Academic Press, 1962.

[11] J. A. Bucklew. *Large Deviation Techniques in Decision, Simulation, and Estimation.* Wiley-Interscience, New York, 1990.

[12] H.B. Callen. *Thermodynamics and an Introduction to Thermostatics.* John Wiley And Sons, New York, 1985.

[13] S. Chern and P. A. Griffiths. *Pfaffian Systems in Involution*. Mathematical Sciences Research Institute, Berkeley, 1983.

[14] N. Chomsky. Three models for the description of language. *IRE Trans. Info. Th.*, 2:113, 1956.

[15] P. Collet and J.-P. Eckmann. *Maps of the Unit Interval as Dynamical Systems*. Birkhauser, Berlin, 1980.

[16] I. P. Cornfeld, S. V. Fomin, and Ya. G. Sinai. *Ergodic Theory*. Springer-Verlag, Berlin, 1982.

[17] J. P. Crutchfield. *Noisy Chaos*. PhD thesis, University of California, Santa Cruz, 1983. published by University Microfilms Intl, Minnesota.

[18] J. P. Crutchfield. Knowledge and meaning ... chaos and complexity. In L. Lam and H. C. Morris, editors, *Modeling Complex Systems*, Berlin, 1991. Springer-Verlag.

[19] J. P. Crutchfield and B. S. McNamara. Equations of motion from a data series. *Complex Systems*, 1:417, 1987.

[20] J. P. Crutchfield and N. H. Packard. Symbolic dynamics of one-dimensional maps: Entropies, finite precision, and noise. *Intl. J. Theo. Phys.*, 21:433, 1982.

[21] J. P. Crutchfield and N. H. Packard. Symbolic dynamics of noisy chaos. *Physica*, 7D:201, 1983.

[22] J. P. Crutchfield, N. H. Packard, J. D. Farmer, and R. S. Shaw. Chaos. *Sci. Am.*, 255:46, 1986.

[23] J. P. Crutchfield and K. Young. Finitary ϵ-machines: A sofic primer. in preparation, 1989.

[24] J. P. Crutchfield and K. Young. Inferring statistical complexity. *Phys. Rev. Let.*, 63:105, 1989.

[25] 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.

[26] J. P. Crutchfield and K. Young. ϵ-machine spectroscopy. preprint, 1991.

[27] J.-P. Eckmann and D. Ruelle. Ergodic theory of chaos and strange attractors. *Rev. Mod. Phys.*, 57:617, 1985.

[28] R. S. Ellis. *Entropy, Large Deviations, and Statistical Mechanics*, volume 271 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, New York, 1985.

[29] I. E. Farquhar. *Ergodic Theory in Statistical Mechanics*. Wiley-Interscience, New York, 1964.

[30] J. Ford. How random is a coin toss? *Physics Today*, April 1983.

[31] A. Fraser. Using hidden markov models to predict chaos. preprint, 1990.

[32] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, New York, 1979.

[33] J. W. Gibbs. *Elementary Principles of Statistical Mechanics*. Dover, New York, 1960.

[34] J. P. Gollub and S. V. Benson. Many routes to turbulent convection. *J. Fluid Mech.*, 100:449, 1980.

[35] P. Grassberger. Toward a quantitative theory of self-generated complexity. *Intl. J. Theo. Phys.*, 25:907, 1986.

[36] P. Grassberger and H. Kantz. Generating partitions for the dissipative henon map. *Phys. Let. A*, 113:235, 1985.

[37] P. Grassberger and I. Procaccia. Characterization of strange attractors. *Phys. Rev. Lett.*, 50:346, 1983.

[38] J. Guckenheimer and P. Holmes. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*. Springer-Verlag, New York, 1983.

[39] J. Hadamard. Les surfaces a coubures oposees et leurs lignes geodesiques. *J. de. Math. Pures. Appl.*, 4:27, 1898.

[40] T.C. Halsey, M.H. Jensen, L.P. Kadanoff, I. Procaccia, and B.I. Shraiman. Fractal measures and their singularities: the characterization of strange sets. *Phys. Rev. A*, 33:1141, 1986.

[41] J. E. Hanson and J. P. Crutchfield. The attractor-basin portrait of a cellular automaton. submitted, 1991.

[42] R. V. Hartley. Transmission of information. *Bell Sys. Tech. J.*, 7:535, 1928.

[43] J. E. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation.* Addison-Wesley, Reading, MA, 1979.

[44] J. J. Hopfield. Neural networks and physical systems with emergent collective behavior. *Proc. Natl. Acad. Sci.*, 79:2554, 1982.

[45] C. Howson and P. Urbach. *Scientific Reasoning: The Bayesian Approach.* Open Court, La Salle, Illinois, 1989.

[46] S. Kirkpatrick, C. D. Gelatt, and H. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671, 1983.

[47] A. N. Kolmogorov. Local structure of turbulence in an incompressible liquid for very large reynolds numbers. *Proc. Acad. Sci. USSR, Geochem. Sec.*, 30:299, 1949.

[48] A. N. Kolmogorov. Three approaches to the concept of the amount of information. *Prob. Info. Trans.*, 1:1, 1965.

[49] B. Kosko. *Neural Networks and Fuzzy Systems.* Prentice-Hall, New York, 1990.

[50] N. S. Krylov. *Works on the Foundations of Statistical Physics.* Princeton University Press, Princeton, 1979.

[51] T. S. Kuhn. *The Structure of Scientific Revolutions.* University of Chicago Press, Chicago, 1962.

[52] L. D. Landau. *C. R. Dokl. Acad. Sci. USSR*, 44:311, 1944.

[53] J. L. Lebowitz and O. Penrose. Modern ergodic theory. *Phys. Today*, 26:2, 1973.

[54] E. Levin, N. Tishby, and S. A. Solla. A statistical approach to learning and generalization in layered neural networks. In R. Rivest, D. Haussler, and M . K. Warmuth, editors, *Computational Learning Theory*, page 245. Morgan Kaufmann, 1989.

[55] 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.

[56] A. Lindenmayer and P. Prusinkiewicz. Developmental models of multicellular organisms: A computer graphics perspective. In C. G. Langton, editor, *Artificial Life*, page 221. Addison-Wesley, 1989.

[57] B. B. Mandelbrot. Intermittent turbulence in self-similar cascades: divergence of high moments and dimension of the carrier. *J. Fluid. Mech.*, 62:331, 1974.

[58] B. Marcus, R. Adler, and D. Lind. An introduction to symbolic dynamics and its applications. Unpublished Textbook, 1991.

[59] R. M. May. Simple mathematical models with very complicated dynamics. *Nature*, 261:459, 1976.

[60] G. Mayer-Kress, editor. *Dimensions and Entropies in Chaotic Systems*. Springer, 1986.

[61] D. S. Ornstein. *Ergodic Theory, Randomness, and Dynamical Systems*. Yale University Press, New Haven, 1974.

[62] D. S. Ornstein. Ergodic theory, randomness, and chaos. *Science*, 243:182, 1989.

[63] N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw. Geometry from a time series. *Phys. Rev. Let.*, 45:712, 1980.

[64] H. Pagels and S. Lloyd. Complexity as thermodynamic depth. *Ann. Phys.*, 188:186, 1988.

[65] W. Parry. Intrinsic markov chains. *Trans. Amer. Math. Soc.*, 112:55, 1964.

[66] O. Penrose. *Foundations of Statistical Mechanics*. Pergamon Press, Oxford, 1970.

[67] Y. B. Pesin. *Math. Surveys*, 32:55, 1977.

[68] K. Petersen. *Ergodic Theory*. Cambridge University Press, Cambridge, 1983.

[69] P. Pfeuty and G. Toulouse. *Introduction to the Renormalization Group and to Critical Phenomena*. Wiley, 1977.

[70] F. J. Pineda. Generalization of backpropagation to recurrent neural networks. *Phys. Rev. Lett.*, 18:2229, 1987.

[71] H. Poincare. *Les Methodes Nouvelles de la Mecanique Celeste*. Gauthier-Villars, Paris, 1892.

[72] K. R. Popper. *The Logic of Scientific Inquiry.* Basic Books, New York, 1959.

[73] L.E. Reichl. *A Modern Course in Statistical Physics.* University of Texas Press, Austin, 1988.

[74] A. Renyi. On the dimension and entropy of probability distributions. *Acta Math. Hung.*, 10:193, 1959.

[75] A. Renyi. Some fundamental questions of information theory. In *Selected Papers of Alfred Renyi, Vol. 2*, page 526. Akademii Kiado, Budapest, 1976.

[76] J. Rissanen. *Stochastic Complexity in Statistical Inquiry.* World Scientific, Singapore, 1989.

[77] D. Ruelle. *Thermodynamic Formalism.* Addision Wesley, Reading, MA, 1978.

[78] D. Ruelle and F. Takens. On the nature of turbulence. *Comm. Math. Phys.*, 20:167, 1974.

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

[80] J. Scargle. Studies in astronomical time series analysis. i. modeling random processes in the time domain. *Ap. J. Supp. Ser.*, 45:1, 1981.

[81] D. Schertzer and S. Lovejoy. Physical modelling and analysis of rain and clouds by anisotropic scaling multiplicative processes. *J. Geop. Res.*, 92:9693, 1987.

[82] H. G. Schuster. *Deterministic Chaos: An Introduction.* VCH Publishing, New York, 1988.

[83] C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication.* University of Illinois Press, Champaign-Urbana, 1962.

[84] R. Shaw. Strange attractors, chaotic behavior, and information flow. *Z. Naturforsh.*, 36a:80, 1981.

[85] R. Shaw. *The Dripping Faucet as a Model Chaotic System.* Aerial Press, Santa Cruz, California, 1984.

[86] P. Walters. *An Introduction to Ergodic Theory.* Springer-Verlag, New York, 1982.

[87] B. Weiss. Subshifts of finite type and sofic systems. *Monastsh. Math.*, 77:462, 1973.

[88] J. A. Wheeler and W. H. Zurek. *Quantum Theory and Measurement*. Princeton University Press, 1983.

[89] K. Wiesenfeld, J. Theiler, and B. Mcnamara. Self-organized criticality in a deterministic automaton. *Phys. Rev. Lett.*, 65:949, 1990.

PART II
Inferring Statistical Complexity

Abstract[29]

Statistical mechanics is used to describe the observed information processing complexity of nonlinear dynamical systems. We introduce a measure of complexity distinct from and dual to the information theoretic entropies and dimensions. A technique is presented that directly reconstructs minimal equations of motion from the recursive structure of measurement sequences. Application to the period doubling cascade demonstrates a form of superuniversality that refers only to the entropy and complexity of a data stream.

[29] This Part first appeared as: J.P. Crutchfield and K. Young. Inferring Statistical Complexity. *Phy. Rev. Lett.* **63** (1989):105

Inferring Statistical Complexity

Prior to the discovery of chaos, physical processes were broadly described in terms of two extreme models of behavior: periodic and random. both are simple, but in distinct senses: the former since the behavior is temporally repetitive; the latter since it affords a compact statistical description. the existence of chaos demonstrates that there is a rich spectrum of unpredictability spanning these two extremes. Information-theoretic descriptions of this spectrum, (say) using the dynamical entropies, measure the raw diversity of temporal patterns: Periodic behavior has low information content; random, high content. What this misses, however, is the statistical simplicity of random behavior.[7] For example, periodic processes are readily forecast by using a stored finite pattern; stochastic processes by using a source of random numbers. Behavior between these extremes, while of intermediate information content, is more complex in that the most concise description is an amalgam of regular and stochastic processes.

To address the deficiency of conventional entropies this Letter introduces a measure of complexity that quantifies the intrinsic computation of a physical process. The basis of our approach is an abstract notion of complexity: the information contained in the minimum number of equivalence classes obtained by reducing a data set modulo a symmetry. In the following this is given concrete form for the case of data sets produced by nonlinear dynamical systems and reduced by time translation invariance, the symmetry appropriate to forecasting. Central to this is a procedure for reconstructing computationally equivalent machines, whose properties lead to quantitative estimates of complexity and entropy. this is developed using the statistical mechanics of orbit ensembles, rather than focusing on the computational complexity of individual orbits.[4],[15],[12],[3] As an application example, we find that the period doubling route to chaos exhibits a λ-like phase transition with a second order component and that at the accumulation point the behavior has infinite complexity, but is not computation universal.

The first step is obtaining a data stream. The (unknowable) exact states of an observed system are translated into a sequence of symbols via a measurement channel.[6] This process is described by a parametrized partition M_ε of the state space, consisting of cells of size ε that are sampled every τ time units. a measurement sequence consists of the successive elements of M_ε visited over time by the system's state. Using the instrument $\{M_\varepsilon, \tau\}$, a sequence of states $\{x_{\vec{n}}\}$ is mapped into a sequence of symbols $\{s_n : s_n \in A\}$ where $A = \{0, ..., k-1\}$ is the alphabet of labels for the k ($\approx \varepsilon^{-m_{bed}}$) partition elements and m_{bed} is the data set's embedding dimension.[7] a common example, to which we shall return, is the logistic map of the interval, $x_{n+1} = rx_n(1 - x_n)$, observed with the binary generating partition $M_{\frac{1}{2}} = \{[0, 0.5), [0.5, 1)\}$ whose elements are labeled with $A = \{0, 1\}$.[8] We shall refer to the computational models reconstructed from such data as *ε-machines*, in order to emphasize their dependence on the measuring instrument.

Given the data stream in the form of a long measuring sequence, the first step in machine reconstruction is the construction of a tree. A *tree* $T = \{\mathbf{n}, \mathbf{l}\}$ consists of *nodes* $\mathbf{n} = \{n_i\}$ and directed, labeled *links* $\mathbf{l} = \{l_i\}$ connecting them in a hierarchical structure with no closed paths. An L-level *subtree* T_n^L at node n is a tree that starts at node n and contains all nodes that can be reached within L links. To construct a tree from a measurement sequence we simply parse the latter for all length-L sequences (L-cylinders) and from this construct the tree with links up to level L that are labeled with individual symbols up to that time. We add probabilistic structure to the tree by recording for each node n_i the number $N_i(L)$ of occurrences of the associated L-cylinder relative to the total number $N(L)$ observed, $p_{n_i}^T(L) = \frac{n_i(L)}{N(L)}$. This gives a hierarchical approximation of the measure in orbit space. Tree representation of data streams are closely related to the hierarchical algorithm used for measuring dynamical entropies.[6],[8]

The ε-machines are represented by a class of labeled, directed multigraphs, or *l-digraphs*.[10] They are related to the Shannon graphs of information theory,[18] to Weiss's sofic systems,[11] in symbolic dynamics, to discrete finite automata (DFA) [14] in computation theory, and to

regular languages in Chomsky's hierarchy.[5] In the most general formulation, we are concerned with probabilistic versions of these. Their topological structure is described by an l-digraph $G = \{\mathbf{V}, \mathbf{E}\}$ that consists of *vertices* $\mathbf{V}=\{v_i\}$ and directed *edges* $\mathbf{E}=\{e_i\}$ connecting them, each of which is labeled by a symbol $s \in A$.

To reconstruct a topological ε-machine we define an equivalence relation, *subtree similarity*, denoted $\approx$, on the nodes of the tree by the condition that the L-subtrees are identical: $n_i \approx n_j$ if and only if (iff) $T^L_{n_i} = T^L_{n_j}$. Subtree equivalence means that the link structure is identical. This equivalence relation induces a set of equivalence classes $\{\mathbf{C}^L_l: l= 1, ..., K\}$ given by $\mathbf{C}^L_l = \{n \in \mathbf{n}: n_i \in \mathbf{C}^L_l \text{ and } n_i \in \mathbf{C}^L_l \text{ iff } n_i \approx n_i\}$. We refer to the archetypal subtree link structure for each class as a *morph*. A graph G_L is then constructed by associating a vertex to each tree node L-level equivalence class. Two vertices v_k and v_l are connected by a directed edge if the transition exists in the tree between nodes in the equivalence classes, $n_i \rightarrow n_j : n_i \in \mathbf{C}^L_k, n_j \in \mathbf{C}^L_k$. The corresponding edge is labeled by the symbol(s) associated with the tree nodes in the two equivalence classes.

In this way ε-machine reconstruction deduce from the diversity of individual temporal patterns "generalized states", associated with graph vertices, that are optimal for forecasting. the topological ε-machines so reconstructed capture the essential computational aspects of the data stream by virtue of the following instantiation of Occam's razor.

Theorem—. topological reconstruction of G_L produces the minimal machine recognizing the language and the generalized states specified up to L-cylinders by the measurement sequence.

The generalization to reconstructing metric ε-machines that contain the probabilistic structure of the data stream follows by a straightforward extension of subtree similarity. Two L-subtrees are δ similar if they are topologically similar and their corresponding links individually are equally probable within some $\delta \geq 0$. There is also a motivating theorem: Metric reconstruction yields minimal metric ε-machines.

In order to reconstruct an ε-machine it is necessary to have a measure of the "goodness of

fit" for determining ε, τ, δ, and the level L of subtree approximation. this is given by the *graph indeterminacy*, which measures the degree of ambiguity in transitions between graph vertices. The indeterminacy [6] I_G of a labeled digraph G is defined as the weighted conditional entropy

$$I_G = \sum_{v \in \mathbf{V}} p_v \sum_{s \in A} p(s|v) \sum_{v' \in \mathbf{V}} p(v'|v;s) \log p(v'|v;s)$$

where $p(v'|v;s)$ is the transition probability from vertex v to v' along an edge labeled with symbol s, $p(s|v)$ is the probability that s is emitted on leaving v, and p_v is the probability of vertex v; logarithms are to base 2. An ε-machine is reconstructable from L-level equivalence classes if I_{G_L} vanishes. Finite indeterminacy. at some given L, ε, τ, and δ, indicates a residual amount of extrinsic noise at the level approximation.

We now turn to the statistical mechanical description of ε-machines, the central result of which is a natural definition of *complexity* that is dual to the dynamical entropies and dimensions. We define the partition function for n-cylinders $\{s^n\}$ as

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

where α is a formal parameter related to the inverse temperature of statistical mechanics. The α-order total Renyi information, [6] or *free information*, in the n-cylinders is given by $H_\alpha(n) = (1-\alpha)^{-1} \log Z_\alpha(n)$. The dynamical α-order entropies are given by the thermodynamic (large orbit space volume) limit

$$h_\alpha = \lim_{n \to \infty} (1-\alpha)^{-1} n^{-1} \log Z_\alpha(n)$$

An ε-machine is described by a set $\left\{T^{(s)} : s \in A\right\}$ of transition matrices, one for each symbol $s \in A$, where $T^{(s)} = \{p_{s,ij}\}$ with $p_{s,ij} = p(v_j|v_i;s)$. We define the probabilistic connection matrix as $T_\alpha = \{t_{ij}\} = \sum_{s \in A} T_\alpha^{(s)}$, where the parametrized matrix $T_\alpha^{(s)} = \left\{p_{s,ij}^\alpha\right\}$. T defines a Green's function on the tree of measurement sequences. The eigenvalue spectrum $S[T_\alpha]$ will be denoted $\{\lambda_i(\alpha)\}$. The largest eigenvalue λ_α of T_α is real for ε-machines. the

associated eigenvector $\hat{p}_\alpha = \{p_v^\alpha : v \in \mathbf{V}\}$ has non-negative elements that give the asymptotic vertex probabilities.

As a measure of the information processing capacity of an ε-machine, we define the α-order graph complexity as the Renyi entropy of $\hat{p}_\alpha$,

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

This is an intensive thermodynamic quantity that measures the average amount of α information contained in the morphs. It also quantifies the informational fluctuations in the data stream. Fluctuations in the free information are measured via the total excess (α) entropy for the L-cylinders [9],[16],[13]

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

It is useful in its own right since if a machine is finite, i.e., $F_1(L)$ is finite, the process is weak Bernoulli.[16] $F_\alpha(L)$ is a measure of fluctuations of finite cylinder set statistics from asymptotic. In the thermodynamic limit, $L \to \infty$, the α-order complexity is simply proportional to F_α. Heuristically speaking, the complexity measures the average amount of mathematical work required to produce a fluctuation.

Developing the partition function in terms of eigenvalues of T_α, the α-order total entropies can also be expressed in terms of the ε-machine eigenvalue λ_α : $H_\alpha(n) \approx (1-\alpha)^{-1} \log \lambda_\alpha^n$, $n \to \infty$. The Renyi entropy spectrum is then $h_\alpha(n) = (1-\alpha)^{-1} \log \lambda_\alpha$. For completeness, we note that the spectrum d_α of Renyi dimensions can be similarly related to the reconstructed ε-machine via the ε scaling of H_α.

There are two important cases for the α-order graph complexity that we now explicitly consider. the first is the topological case, when $\alpha = 0$. T_0 is the digraphs connection matrix, the Renyi entropy $h_0 = \log \lambda_0$ is the topological entropy h, and the graph complexity is the probabilistic algorithmic complexity, $C_0(G) = \log |\mathbf{V}|$. In the case of cellular automata

and assuming one has the equations of motion, this quantity has been referred to as the algorithmic complexity.[19] It is important, however, to distinguish C_0 from the Chaitin-Kolmogorov complexity, which goes under the same name and describes the complexity of individual measurement sequences.[4],[15],[12],[3] The quantities that we have defined here are *probabilistic* and referred to Turing machines with a random register.

The second (metric) case of interest is the "high temperature limit", when $\alpha = 1$. Here h_α is the metric entropy: $h_\mu = \lim_{\alpha \to 1} h_\alpha = -\frac{d\lambda_\alpha}{d\alpha}$. C_α is the graph complexity $C_1 = -\sum_{v \in \mathbf{V}} p_v \log p_v$, that is, the Shannon entropy of $\hat{p}_1$.[13],[1]

We demonstrate the practical implementation of ε-machine reconstruction and the foregoing formalism on the well-studied period-doubling cascade, as exhibited by the logistic map. Figure 26 displays topological ε-machines for several notable parameter values. Figure 27 shows the graph complexity $C_\alpha(L)$ versus the specific entropy $L^{-1}H_\alpha(L)$ for the metric case, $\alpha = 1$, at parameter values associated with period-doubling cascades of various periodicities. Of particular interest is the appearance of a phase transition as a function of specific entropy. The observed divergence in C_1 indicates a form of "superuniversality" for the transition, since we are only considering the complexity's dependence on the information-theoretic characterization, viz, H_α, of the measurement sequences, and not the dependence on parameter value. the lower bound on C_1, attained for periodic behavior and at band mergings, indicates a λ-like phase transition with a second-order component. The remaining chaotic data are always significantly above the lower bound.

This example shows that C_α is a measure of complexity distinct from and dual to standard information measures of dimension d_α and entropy h_α. The latter, given by the ε-machine's eigenvalue, simply measures the diversity of observed patterns in a data stream: the more random the source, the more patterns, and so the higher the information content. The graph complexity, however, given by the ε-machine's eigenvector, measures the computational resources required to reproduce a data stream.[17],[2] It vanishes for trivially periodic and

for purely random data sets. A reconstructed ε-machine reflects a balanced utilization of deterministic and random information processing. We claim this model basis is the proper one for describing the computational complexity of physical processes, since the latter always have some residual extrinsic fluctuations.

We have argued for a chaotic measurement theory that exploits the intimate relationship between information, computation and forecasting.[6],[8],[7] This Letter establishes the connections constructively and within the framework of statistical mechanics. With a direct measure of an ε-machine's complexity, the theory gives a computation-theoretic foundation to the notion of model optimality and, most importantly, a measure of the computational complexity of estimated models.[7] The generality of ε-machine reconstruction and its ability to infer generalized states from a data stream suggest that it captures essential aspects of learning. That it can be put into a statistical mechanical framework, therefore, suggests the existence of a complete theoretical basis for "artificial science": the fully automated deduction of optimal models of physical processes.[7]

This work was funded in part by ONR Contract No. N00014-86-K-0154 and by the 1988 Cray-Sponsored University Research and Development Grant Program. The authors thank Professor Carson Jeffries for his continuing support.

Figures

Figure Captions

Figure 26 Topological ϵ-machine l-digraphs for the logistic map at (a) the first period doubling accumulation r_c=3.569945671..., (b) the band merging $r_{2b\rightarrow 1b}$=3.67859..., (c) the "typical" chaotic value r=3.7, and (d) the most chaotic value r=4.0. (a) and (c) show approximations of infinite ϵ-machines. The start vertex is indicated by a double circle; all states are accepting; otherwise see ref. [14].

Figure 27 Graph complexity C_l vs. specific entropy $H_l(16)/16$, using the binary, generating partition {[0,0.5),[0.5,1)}, for the logistic map at 193 parameter values r in [3,4] associated with various period doubling cascades. For most, the underlying tree was constructed from 32-cylinders and machines from 16-cylinders. From high-entropy data sets smaller cylinders were used as determined by storage. Note the phase transition (divergence) at H^* approximately equal to 0.28. Below H^* behavior is periodic and C_α=H_α=log(period). Above H^*, the data are chaotic. The lower bound C_α=log(B) is attained at $B\rightarrow B/2$ band mergings.

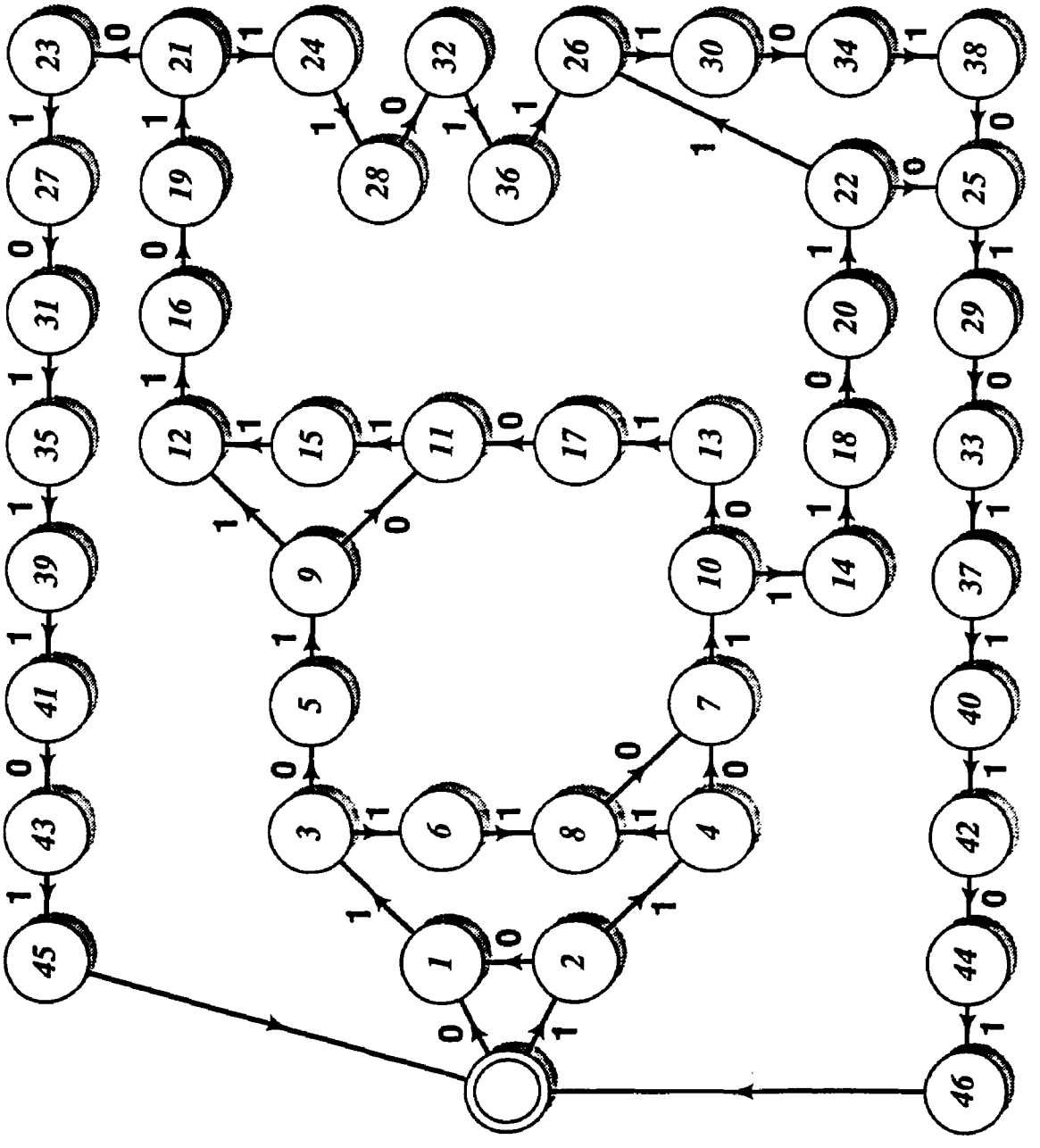

23
21
24
32
26
30
34
38
27
19
28
36
22
25
31
16
20
29
35
12
15
11
17
13
18
33
39
9
10
14
37
41
5
7
40
43
3
6
8
4
42
45
1
2
44
46

Figure 26(a)

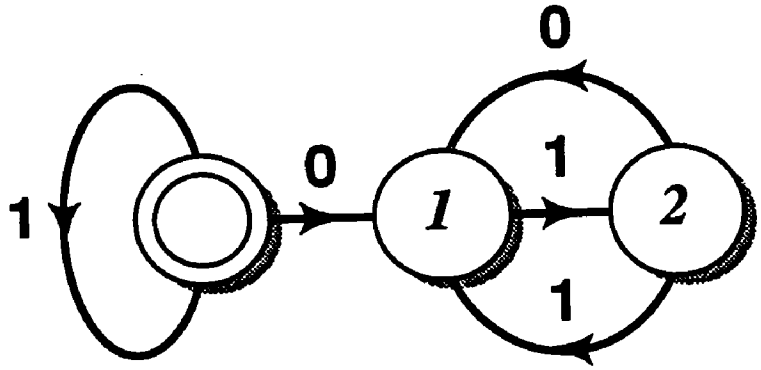


Figure 26(b)

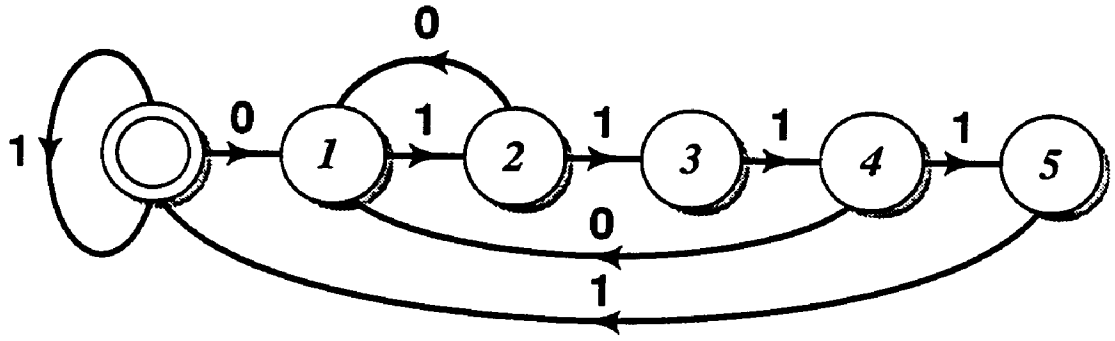


Figure 26(c)

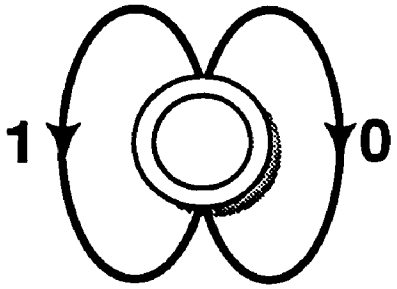


Figure 26(d)

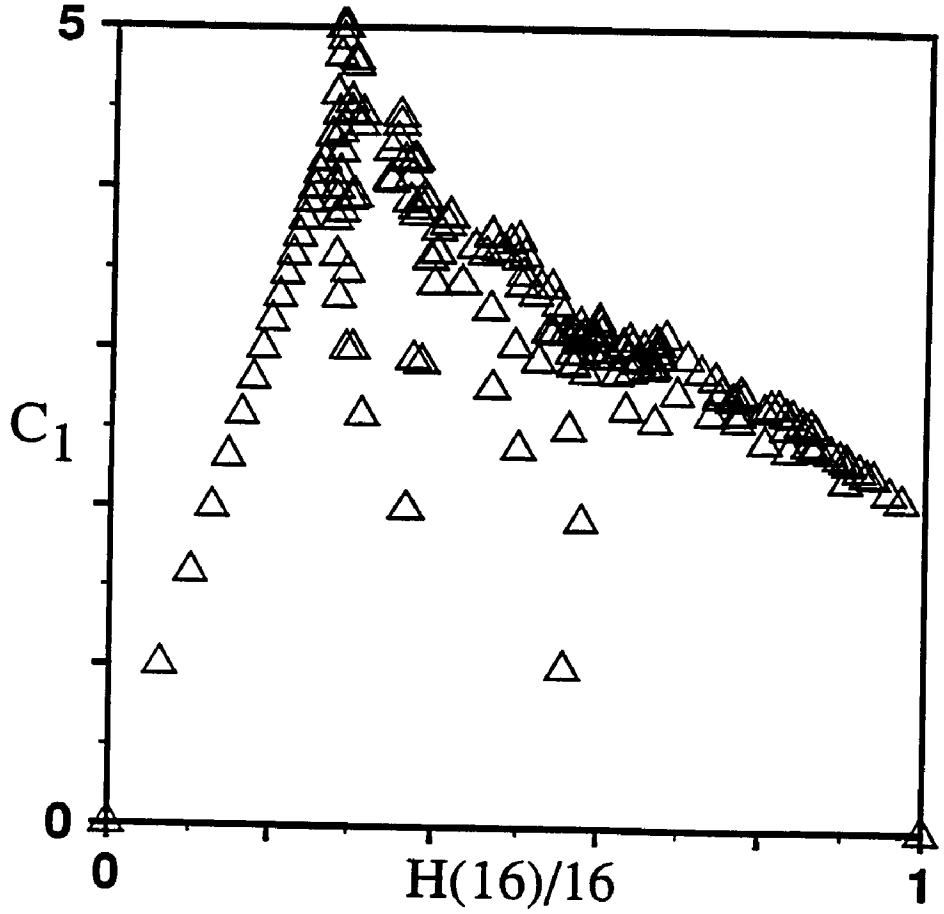


Figure 27

Bibliography

[1] C. P. Bachas and B.A. Huberman. Complexity and relaxation of hierarchical structures. *Phys. Rev. Let.*, 57:1965, 1986.

[2] C. H. Bennett. Dissipation, information, computational complexity, and the definition of organization. In D. Pines, editor, *Emerging Syntheses in the Sciences.* Addison-Wesley, Redwood City, 1988.

[3] A. A. Brudno. Entropy and the complexity of the trajectories of a dynamical system. *Trans. Moscow Math. Soc.*, 44:127, 1983.

[4] G. Chaitin. On the length of programs for computing finite binary sequences. *J. ACM*, 13:145, 1966.

[5] N. Chomsky. Three models for the description of language. *IRE Trans. Info. Th.*, 2:113, 1956.

[6] J. P. Crutchfield. *Noisy Chaos.* PhD thesis, University of California, Santa Cruz, 1983. published by University Microfilms Intl, Minnesota.

[7] J. P. Crutchfield and B. S. McNamara. Equations of motion from a data series. *Complex Systems*, 1:417, 1987.

[8] J. P. Crutchfield and N. H. Packard. Symbolic dynamics of one-dimensional maps: Entropies, finite precision, and noise. *Intl. J. Theo. Phys.*, 21:433, 1982.

[9] J. P. Crutchfield and N. H. Packard. Symbolic dynamics of noisy chaos. *Physica*, 7D:201, 1983.

[10] D. M. Cvetkovic, M. Doob, and H. Sachs. *Spectra of Graphs.* Academic Press, New York, 1980.

[11] R. Fischer. Sofic systems and graphs. *Monastsh. Math*, 80:179, 1975.

[12] J. Ford. Chaos: Solving the unsolvable, predicting the unpredictable. In M. F. Barnsley and S. G. Demko, editors, *Chaotic Dynamics and Fractals*, page 1. Academic Press, Orlando, 1986.

[13] P. Grassberger. Toward a quantitative theory of self-generated complexity. *Intl. J. Theo. Phys.*, 25:907, 1986.

[14] J. E. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation.* Addison-Wesley, Reading, MA, 1979.

[15] A. N. Kolmogorov. Three approaches to the concept of the amount of information. *Prob. Info. Trans.*, 1:1, 1965.

[16] N. H. Packard. *Measurements of Chaos in the Presence of Noise.* PhD thesis, University of California, Santa Cruz, 1982.

[17] H. Pagels and S. Lloyd. Complexity as thermodynamic depth. *Ann. Phys.*, 188:186, 1988.

[18] C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication.* University of Illinois Press, Champaign-Urbana, 1962.

[19] S. Wolfram. Computation theory of cellular automata. *Comm. Math. Phys.*, 96:15, 1984.

PART III
Computation at The Onset of Chaos

Abstract[30]

Computation at levels beyond storage and transmission of information appears in physical systems at phase transitions. We investigate this phenomenon using minimal computational models of dynamical systems that undergo a transition to chaos as a function of a nonlinearity parameter. For period-doubling and band-merging cascades, we derive expressions for the entropy, the interdependence of ϵ-machine complexity and entropy, and the latent complexity of the transition to chaos. At the transition deterministic finite automaton models diverge in size. Although there is no regular or context-free Chomsky grammar in this case, we give finite descriptions at the higher computational level of context-free Lindenmayer systems. We construct a restricted indexed context-free grammar and its associated one-way nondeterministic nested stack automaton for the cascade limit language.

This analysis of a family of dynamical systems suggests a complexity theoretic description of phase transitions based on the informational diversity and computational complexity of observed data that is independent of particular system control parameters. The approach gives a much more refined picture of the architecture of critical states than is available via correlation functions, mutual information, and statistical mechanics generally. The analytic methods establish quantitatively the longstanding observation that significant computation is associated with the critical states found at the border between order and chaos.

[30] This Part first appeared as:J.P. Crutchfield and K. Young. Computation at the Onset of Chaos. In *Complexity, Entropy, and the Physics of Information*, edited by W.H. Zurek. Santa Fe Institutes Studies in the Sciences of Complexity, vol. VIII. Redwood City. CA: Addison-Wesley, 1990, 223.

Beyond a Clock and a Coin Flip

The clock and the coin flip signify the two best understood behaviors that a physical system can exhibit. Utter regularity and utter randomness are the dynamical legacy of two millenia of physical thought. Only within this century, however, has their fundamental place been established. Today, realistic models of time-dependent behavior necessarily incorporate elements of both.

The regularity and Laplacian determinism of a clock are fundamental to much of physical theory. Einstein's careful philosophical consideration of the role of time is a noteworthy example. The use of a mechanical device to mark regular events is the cornerstone of relativity theory.[31] A completely predictable system, which we shall denote by $\mathbf{P}_t$, is essentially a clock; the hands indicate the current state and the mechanism advances them to the next state without choice. For a predictable system some fixed pattern is repeated every (say) t seconds.

Diametrically opposed, the coin flip, a picaresque example of ideal randomness, is the basic model underlying probability and ergodic theories. The next state in such a system is statistically independent of the preceding and is reached by exercising maximum choice. In ergodic theory the formal model of the coin flip is the Bernoulli flow $\mathbf{B}_t$, a coin flip every t seconds.

We take $\mathbf{B}_t$ and $\mathbf{P}_t$ as the basic processes with which to model the complexity of nonlinear dynamical systems. In attempting to describe a particular set of observations, if we find that they repeat then we can describe them as having been produced by some variant of $\mathbf{P}_t$. Whereas, if they are completely unpredictable then their generating process is essentially the same as $\mathbf{B}_t$. Any real system S, of course, will contain elements of both and so naturally we ask whether it is always the case that some observed behavior can be decomposed into these separate components. Is $\mathbf{S} = \mathbf{B}_t \otimes \mathbf{P}_t$? Both ergodic and probability theories say that this cannot be done so simply in general. Ornstein showed that there are ergodic systems that cannot be separated into

[31] It is not an idle speculation to wonder what happen's to Einstein's universe if his clock contains an irreducible element of randomness, or more realistically, if it is chaotic.

completely predictable and completely random processes.[56] The Wold-Kolmogorov spectral decomposition states that although the frequency spectrum of a stationary process consists of a singular spectral component associated with periodic and almost periodic behavior and a broadband continuous component associated with an absolutely continuous measure, there remain other statistical elements beyond these.[46, 47, 69]

What is this other behavior, captured neither by clocks nor by coin flips? A partial answer comes from computation theory and is the subject of the following.

The most general model of deterministic computation is the universal Turing machine (UTM).[32] Any computational aspect of a regular process like $\mathbf{P}_t$ can be programmed and so modeled with this machine. In order that the Turing machine *readily* model processes like $\mathbf{B}_t$ we augment it with a random register whose state it samples with a special instruction.[33] The result is the Bernoulli-Turing machine (BTM). It captures both the completely predictable via its subset of deterministic operations and the completely unpredictable by accessing its stochastic register. If the data is completely random, a BTM models it most efficiently by guessing. A BTM reads and prints the contents of its "Bernoulli" register, rather than implementing some large deterministic computation to generate pseudo-random numbers. What are the implications for physical theory? A variant of the Church-Turing thesis is appropriate: the Bernoulli-Turing machine is powerful enough to describe even the "other stuff" of ergodic and probability theories.

Let us delve a little further into these considerations by drawing parallels. One goal here is to infer how much of a data stream can be ascribed to a certain set of models $\{\mathbf{B}_t, \mathbf{P}_t\}$. This model basis induces a set of equivalences in the space of stationary signals. Thus, starting with the abstract notions of strict determinism and randomness, we obtain a decomposition of

---

[32] This statement is something of an article of faith that is formulated by the Church-Turing Thesis: any reasonably specifiable computation can be articulated as a program for a UTM.[38]

[33] This register can also be modeled with a second tape containing random bits. In this case, the resulting machine is referred to as an "Random Oracle" Turing Machine.[34] What we have in mind, although formally equivalent, is that the machine in question is physically coupled to an information source whose bits are random with respect the computation at hand. Thus, we do not require ideal random bits.

that space. A quantity that is constant in each equivalence class is an invariant of the modeling decomposition. Of course, we are also interested in those cases where the model basis is inadequate; where more of the computational power of the BTM must be invoked. When this occurs, it hints that the model basis should be expanded. This will then refine the decomposition and lead to new invariants.

An analogous, but restricted type of decomposition is also pursued formally in ergodic and computation theories by showing how particular examples can be mapped onto one another. The motivations being that the structure of the decomposition is a representation of the defining equivalence concept and, furthermore, the latter can be quantified by an invariant. A classic problem in ergodic theory has been to identify those systems that are isomorphic to B_t. The associated invariant used for this is the metric entropy, introduced into dynamical systems theory by Kolmogorov[43, 44] and Sinai[66] from Shannon's information theory.[63] Two Bernoulli processes are equivalent if they have the same entropy.[56] Similarly, in computation theory there has been a continuing effort to establish an equivalence between various hard-to-solve, but easily-verified, problems. This is the class of nondeterministic polynomial (NP) problems. If one can guess the correct answer, it can be verified as such in polynomial time. The equivalence between NP problems, called NP-completeness, requires that within a polynomial number of TM steps a problem can be reduced to one hardest problem.[34] The invariant of this polynomial-time reduction equivalence is the growth rate, as a function of problem size, of the computation required to solve the problem. This growth rate is called the algorithmic complexity.[34]

The complementarity between these two endeavors can be made more explicit when both are focused on the single problem of modeling chaotic dynamical systems. Ergodic theory is seen to classify complicated behavior in terms of information production properties, e.g. via the metric entropy. Computation theory describes the same behavior via the intrinsic amount of computation that is performed by the dynamical system. This is quantified in terms of

[34] In fact, the invariant actually used is a much coarsened version of the algorithmic complexity: a polynomial time reduction is required only to preserve the exponential character of solving a hard problem.

machine size (memory) and the number of machine steps to reproduce behavior.[35] It turns out, as explained in more detail below, that this type of algorithmic measure of complexity is equivalent to entropy. As a remedy to this we introduce a complexity measure based on BTMs that is actually complementary to the entropy.

The emphasis in the following is that the tools of each field are complementary and both approaches are necessary to completely describe physical complexity. The basic result is that if one is careful to restrict the class of computational models assumed to be the least powerful necessary to capture behavior, then much of the abstract theory of computation and complexity can be constructively implemented.[36] From this viewpoint, phase transitions in physical systems are seen to support high levels of computation. And conversely, computers are seen to be physical systems designed with a subset of "critical" degrees of freedom that support computational fluctuations.

The discussion has a top-down organization with three major parts. The first, consisting of this section and the next, introduces the motivations and general formalism of applying computational ideas to modeling dynamical systems. The second part develops the basic tools of ϵ-machine reconstruction and a statistical mechanical description of the machines themselves. The third part applies the tools to the particular class of complex behavior seen in cascade transitions to chaos. A few words on further applications conclude the presentation.

Conditional Complexity

The basic concept of complexity that allows for dynamical systems and computation theories to be profitably linked relies on a generalized notion of structure that we will refer to generically as "symmetry". In addition to repetitive structure, we also consider statistical

---

[35] We note that computation theory also allows one to formalize how much effort is required to infer a dynamical system from observed data. Although related, this is not our present concern.[18]

[36] At the highest computation level of universal Turing machines, descriptions of physical complexity are simply not constructive since finding the minimal TM program for a given problem is undecidable in general.[38]

regularity to be one example of symmetry. The idea is that a data set is complex if it is the composite of many symmetries.

To connect back to the preceding discussion, we take as two basic dynamical symmetries those represented by the model basis $\{\mathbf{B}_t, \mathbf{P}_t\}$. A complex process will have, at the very least, some nontrivial combination of these components. Simply predictable behavior and purely random behavior will not be complex. The corresponding complexity spectrum is schematically illustrated in figure 28.

More formally, we define the conditional complexity $C(D|S)$ to be the amount of information in equivalence classes induced by the symmetry S in the data D plus the amount of data that is "unexplained" by S. If we had some way of enumerating all symmetries, then the absolute complexity $C(D)$ would be

$$C(D) = \inf_{\{S\}} C(D|S)$$

And we would say that an object is complex if, after reduction, it is its own symmetry. In that case, there are no symmetries in the object, other than itself.[37] If D is the best model of itself, then there is no unexplained data, but the model is large: $C(D|D) \propto \text{length}(D)$. Conversely, if there is no model, then all of the data is unexplained: $C(D|\emptyset) \propto \text{length}(D)$. The infimum formalizes the notion of considering all possible model "bases" and choosing those that yield the most compact description.[38]

This definition of conditional modeling complexity mirrors that for algorithmic randomness[9, 10, 45, 48, 50] and is closely related to computational approaches to inductive inference.[67]

[37] Or, said another way, the complex object is only described by a large number of equivalence classes induced by inappropriate symmetries. The latter can be illustrated by considering an inappropriate description of a simple object. A square wave signal is infinitely complex with respect to a Fourier basis. But this is not an intrinsic property of square waves, only of the choice of model basis. There is a model basis that gives a very simple description of a square wave.

[38] This computational framework for modeling also applies, in principle, to estimating symbolic equations of motion from noisy continuous data.[25] Generally, minimization is an application of Occam's Razor in which the description is considered to be a "theory" explaining the data.[42] Rissanen's minimum description length principle, the coding theoretic version of this philosophical axiom, yields asymptotically optimal representations.[61, 62]

A string s is random if it is its own shortest UTM description. The latter is a complexity measure called the Chaitin-Kolmogorov complexity $K(\mathrm{s})$ of the string s. In the above notation $K(\mathrm{s}) = C(\mathrm{s}|UTM)$. The class of "symmetries" referred to here are those computable by a deterministic UTM. After factoring these out, any residual "unidentified" or "unexplained" data is taken as input to the UTM program. With respect to the inferred symmetries, this data is "noise". It is included in measuring the size of the minimal UTM representation. $K(\mathrm{s})$ measures the size of two components: an emulation program and input data to that emulation. To reconstruct s the UTM first reads in the program portion in order to emulate the computational part of the description. This computes the inferred symmetries. The (emulated) machine then queries the input tape as necessary to disambiguate indeterminant branchings in the computation of s. $K(\mathrm{s})$ should not be confused with the proposed measure of physical complexity based on BTMs, $C(\mathrm{s}|BTM)$, which include statistical symmetries. There is, in fact, a degeneracy of terminology here that is easily described and avoided.

Consider the data in question to be an orbit $x_t(x_0)$ of duration t starting at state x_0 of a dynamical system admitting an absolutely continuous invariant measure.[39] The algorithmic complexity[27] $\mathcal{A}(x_t(x_0))$ is the growth rate of the Chaitin-Kolmogorov complexity with longer orbits

$$\mathcal{A}(x_t(x_0)) = \lim_{t\to\infty} \frac{K(x_t(x_0))}{t}$$

Note that this artifice removes constant terms in the Chaitin-Kolmogorov complexity, such as those due to the particular implementation of the UTM, and gives a quantity that is machine independent. Then, the algorithmic complexity is the dynamical system's metric entropy, except for orbits starting at a measure zero set of initial conditions. These statements connect the notion of complexity of single strings with that of the ensemble of typical orbits. The Chaitin-Kolmogorov complexity is the same as informational measures of randomness, but

[39] In information theoretic terms we are requiring stationarity and ergodicity of the source.

is distinct from the BTM complexity.[40] To avoid this terminological ambiguity we shall minimize references to algorithmic and Chaitin-Kolmogorov complexities since in most physical situations they measure the same dynamical property captured by the information theoretic phrase "entropy". "Complexity" shall refer to conditional complexity with respect to BTM computational models. We could qualify it further by using "physical complexity", but this is somewhat misleading since it applies equally well outside of physics.[41]

We are not aware of any means of enumerating the space of symmetries and so the above definition of absolute complexity, while of theoretical interest, is of little immediate application. Nonetheless, we can posit that symmetries S be effectively computable in order to be relevant to scientific investigation. According to the physical variant of the Church-Turing thesis, then, S can be implemented on a BTM. Which is to say that as far as realizability is concerned, the unifying class of symmetries we have in mind is represented by operations of a BTM. Although the mathematical specification for a BTM is small; its range of computation is vast; at least as large as the underlying UTM. It is, in fact, unnecessarily powerful so that many questions, such as finding a minimal program for given data, are undecidable and many quantities, such as the conditional complexity $C(D|BTM)$, are noncomputable. More to the point, adopting too general a computational model results in there being little to say about a wide range of physical processes.

Practical measures of complexity are based on lower levels of Chomsky's computational hierarchy.[42] Indeed, Turing machines appear only at the pinnacle of this graded hierarchy. The following concentrates on deterministic finite automata (DFA) and stack automata (SA) complexity, the lowest two levels in the hierarchy. DFAs represent strictly clock and coin flip modeling. SAs are DFAs augmented by an infinite memory with restricted pushdown stack access. We will demonstrate how DFA models break down at a chaotic phase transition and

[40] We are necessarily skipping over a number of details, such as how the state x_t is discretized into a string over a finite alphabet. The basic point made here has been emphasized some time ago.[27, 8]

[41] This definition of complexity and its basic properties as represented in figure 28 were presented by the first author at the International Workshop on "Dimensions and Entropies in Chaotic Systems", Pecos, New Mexico, 11-16 September 1985.

[42] Further development of this topic is given elsewhere.[16, 19]

how higher levels of computational model arise naturally. Estimating complexity types beyond SAs, such as linear bounded automata (LBA), is fraught with certain intriguing difficulties and will not be attempted here. Nonetheless, setting the problem context as broadly as we have just done is useful to indicate the eventual goals we have in mind and to contrast the present approach to other longstanding proposals that UTMs are the appropriate framework with which to describe the complexity of natural processes.[43] Even with the restriction to Chomsky's lower levels a good deal of progress can be made since, as will become clear, contemporary statistical mechanics is largely associated with DFA modeling.

Reconstructing ϵ-Machines

To effectively measure intrinsic computational properties of a physical system we infer an ϵ-machine from a data stream obtained via a measuring instrument.[29] An ϵ-machine is a stochastic automaton of the minimal computational power yielding a finite description of the data stream. Minimality is essential. It restricts the scope of properties detected in the ϵ-machine to be no larger than those possessed by the underlying physical system. We will assume that the data stream is governed by a stationary measure. That is, the probabilities of fixed length blocks of measurements exist and are time-translation invariant.

The goal, then, is to reconstruct from a given physical process a computationally equivalent machine. The reconstruction technique, discussed in the following, is quite general and applies directly to the modeling task for forecasting temporal or spatio-temporal data series. The resulting minimal machine's structure indicates the inherent information processing, i.e. transmission and computation, of the original physical process. The associated complexity measure quantifies the ϵ-machine's informational size; in one limit, it is the logarithm of the number of machine states.

[43] We have in mind Kolmogorov's work[48] over many years that often emphasizes dynamical and physical aspects of this problem. Also, Bennett's notion of "logical depth" and his analysis of physical processes typically employ UTM models.[5] Wolfram's suggestion[71] that the computational properties of intractability and undecidability will play an important role in future theoretical physics assumes UTMs as the model basis. More recently, Zurek[73] has taken up UTM descriptions of thermodynamic processes. The information metric used there was also developed from a conditional complexity.[17]

The machine's states are associated with historical contexts, called morphs, that are optimal for forecasting. Although the simplest (topological) representation of an ϵ-machine at the lowest computational level (DFAs) is in the form of labeled directed graphs, the full development captures the probabilistic (metric) properties of the data stream. Our complexity measure unifies a number of disparate attempts to describe the information processing of nonlinear physical systems.[25, 27, 28, 70, 65, 6, 4, 35, 59] The following two sections develop the reconstruction method for the machines and their statistical mechanics.

The initial task of inferring automata from observed data falls under the purview of grammatical inference within formal learning theory.[18] The inference technique uses a particular choice S of symmetry that is appropriate to forecasting the data stream in order to estimate the conditional complexity $C(D|S)$. The aim is to infer generalized "states" in the data stream that are optimal for forecasting. We will identify these states with measurement sequences giving rise to the same set of possible future sequences.[44] Using the temporal translation invariance guaranteed by stationarity, we identify these states using a sliding window that advances one measurement at a time through the sequence. This leads to the second step in the inference technique, the construction of a parse tree for the measurement sequence probability distribution. This is a coarse-grained representation of the underlying process's measure in orbit space. The state identification requirement then leads to an equivalence relation on the parse tree. The machine states correspond to the induced equivalence classes; the state transitions, to the observed transitions in the tree between the classes. We now give a more formal development of the inference method.

The first step is to obtain a data stream. The main modeling *ansatz* is that the underlying process is governed by a noisy discrete-time dynamical system

$$\vec{x}_{n+1} = \vec{F}(\vec{x}_n) + \vec{\xi}_n, \quad \vec{x}_0 \in \mathrm{M}$$

[44] We note that the same construction can be done for past possibilities. We shall discuss this alternative elsewhere.

where M is the m-dimensional space of states, $\vec{x}_0 = (x_0^0, x_0^1, \ldots, x_0^{m-1})$ is the system's initial state, $\vec{F}$ is the dynamic, the governing deterministic equations of motion, and $\vec{\xi}_n$ represents external time-dependent fluctuations. We shall concentrate on the deterministic case in the following. The (unknowable) exact states of the observed system are translated[45] into a sequence of symbols via a measurement channel.[14] This process is described by a parametrized partition

$$P_\epsilon = \left\{ c_i : \bigcup_{i=0}^{k-1} c_i = \mathrm{M},\ c_i \cap c_j = \emptyset, i \neq j;\ i,j = 0,\ldots,k-1 \right\}$$

of the state space M, consisting of cells c_i of volume ϵ^m that are sampled every τ time units. A measurement sequence consists of the labels from the successive elements of P_ϵ visited over time by the system's state. Using the instrument $\mathbf{I} = \{P_\epsilon, \tau\}$, a sequence of states $\{\vec{x}_n\}$ is mapped into a sequence of symbols $\{s_n : s_n \in \mathbf{A}\}$, where $\mathbf{A} = \{0,\ldots,k-1\}$ is the alphabet of labels for the k ($\approx \epsilon^{-m}$) partition elements.[25] A common example, to which we shall return near the end, is the logistic map of the interval, $x_{n+1} = rx_n(1-x_n)$, observed with the binary generating partition $P_{\frac{1}{2}} = \{[0.,.5),[.5,1.]\}$ whose elements are labeled with $\mathbf{A} = \{0,1\}$.[27] The computational models reconstructed from such data are referred to as ϵ-machines in order to emphasize their dependence on the measuring instrument I.

Given the data stream in the form of a long measurement sequence $\mathbf{s} = \{s_0 s_1 s_2 \cdots : s_i \in \mathbf{A}\}$, the second step in machine inference is the construction of a parse tree. A tree $T = \{\mathbf{n}, \mathbf{l}\}$ consists of nodes $\mathbf{n} = \{n_i\}$ and directed, labeled links $\mathbf{l} = \{l_i\}$ connecting them in a hierarchical structure with no closed paths. The links are labeled by the measurement symbols $s \in \mathbf{A}$. An L-level subtree T_n^L is a tree that starts at node n and contains all nodes below n that can be reached within L links. To construct a tree from a measurement sequence we simply parse the latter for all length L sequences and from this construct the tree with links up to level L that are labeled with individual symbols up to that time. We refer to length L subsequences

---

[45] We ignore for brevity's sake the question of extracting from a single component $\{x_n^i\}$ an adequate reconstructed state space.[58]

$s^L = \left\{ s_i \cdots s_j \cdots s_{i+L-1} : s_j = (\mathbf{s})_j \right\}$ as L-cylinders.[46] Hence an L level tree has a length L path corresponding to each distinct observed L-cylinder. Probabilistic structure is added to the tree by recording for each node n_i the number $N_i(L)$ of occurrences of the associated L-cylinder relative to the total number *N(L)* observed,

$$p^T_{n_i}(L) = \frac{N_i(L)}{N(L)}$$

This gives a hierarchical approximation of the measure in orbit space $\mathbf{M} \otimes \mathbf{time}$. Tree representations of data streams are closely related to the hierarchical algorithm used for estimating dynamical entropies.[27, 14]

At the lowest computational level ϵ-machines are represented by a class of labeled, directed multigraph, or l-digraphs.[30] They are related to the Shannon graphs of information theory,[63] to Weiss's sofic systems in symbolic dynamics,[33] to discrete finite automata in computation theory,[38] and to regular languages in Chomsky's hierarchy.[11] Here we are concerned with probabilistic versions of these. Their topological structure is described by an l-digraph $G = \{\mathbf{V}, \mathbf{E}\}$ that consists of vertices $\mathbf{V} = \{v_i\}$ and directed edges $\mathbf{E} = \{e_i\}$ connecting them, each of the latter is labeled by a symbol $s \in \mathbf{A}$.

To reconstruct a topological ϵ-machine we define an equivalence relation, subtree similarity, denoted $\sim$, on the nodes of the tree T by the condition that the L-subtrees are identical:

$$n \sim n' \text{ if and only if } T^L_n = T^L_{n'}$$

Subtree equivalence means that the link structure is identical. This equivalence relation induces on T, and so on the measurement sequence **s**, a set of equivalence classes $\left\{ \mathbf{C}^L_m : m = 1, \ldots, K \right\}$ given by

$$\mathbf{C}^L_l = \left\{ n \in \mathbf{n} : n \in \mathbf{C}^L_l \text{ and } n' \in \mathbf{C}^L_l \text{ iff } n \sim n' \right\}$$

[46] The picture here is that a particular L-cylinder is a name for that bundle of orbits $\{\vec{x}_n\}$ each of which visited the sequence of partition elements indexed by the L-cylinder.

We refer to the archetypal subtree link structure for each class as a "morph". An l-digraph G_L is then constructed by associating a vertex to each tree node L-level equivalence class; that is, $\mathbf{V} = \left\{\mathbf{C}_m^L\right\}$. Two vertices v_k and v_l are connected by a directed edge $e = (v_k \to v_l)$ if the transition exists in T between nodes in the equivalence classes,

$$n \to n' : n \in \mathbf{C}_k^L, n' \in \mathbf{C}_l^L$$

The corresponding edge is labeled by the symbol(s) $s \in \mathbf{A}$ associated with the tree links connecting the tree nodes in the two equivalence classes

$$\mathbf{E} = \left\{ e = (v_k, v_l; s) : \; v_k \underset{s}{\to} v_l \text{ iff } n \underset{s}{\to} n';\; n \in \mathbf{C}_k^L, n' \in \mathbf{C}_l^L, s \in \mathbf{A} \right\}$$

In this way, ϵ-machine reconstruction deduces from the diversity of individual patterns in the data stream "generalized states", the morphs, associated with the graph vertices, that are optimal for forecasting. The topological ϵ-machines so reconstructed capture the essential computational aspects of the data stream by virtue of the following instantiation of Occam's Razor.

Theorem: Topological reconstruction of G_L produces the minimal and unique machine recognizing the language and the generalized states specified up to L-cylinders by the measurement sequence.

The generalization to reconstructing metric ϵ-machines that contain the probabilistic structure of the data stream follows by a straightforward extension of subtree similarity. Two L-subtrees are δ-similar if they are topologically similar and their corresponding links individually are equally probable within some $\delta \geq 0$. There is also a motivating theorem: metric reconstruction yields minimal metric ϵ-machines.

In order to reconstruct an ϵ-machine it is necessary to have a measure of the "goodness of fit" for determining ϵ, τ, δ, and the level L of subtree approximation. This is given by the graph indeterminacy, which measures the degree of ambiguity in transitions between graph vertices. The indeterminacy[14] I_G of a labeled digraph G is defined as the weighted conditional entropy

$$I_G = \sum_{v \in \mathbf{V}} p_v \sum_{s \in \mathbf{A}} p(s|v) \sum_{v' \in \mathbf{V}} p(v'|v;s) \log p(v'|v;s)$$

where $p(v'|v;s)$ is the transition probability from vertex v to v' along an edge labeled with symbol s, $p(s|v)$ is the probability that s is emitted on leaving v, and p_v is the probability of vertex v. A deterministically-accepting ϵ-machine is reconstructible from L-level equivalence classes if I_{G_L} vanishes. Finite indeterminacy, at some given $\{L, \epsilon, \tau, \delta\}$ indicates a residual amount of extrinsic noise at that level of approximation. In this case, the optimal machine in a set of machines consistent with the data is the smallest that minimizes the indeterminacy.[18]

Statistical Mechanics of ϵ-Machines

Many of the important properties of these stochastic automata models are given concisely using a statistical mechanical formalism that describes the coarse-grained scaling structure of orbit space. We recall some definitions and results necessary for our calculations.[29] The statistical structure of an ϵ-machine is given by a parametrized stochastic connection matrix

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

that is the sum over each symbol $s \in \mathbf{A}$ in the alphabet $\mathbf{A} = \{i : \ i = 0, \ldots, k-1; \ k = \mathcal{O}(\epsilon^{-m})\}$ of the state transition matrices

$$T_\alpha^{(s)} = \left\{ e^{\alpha \log p(v_i | v_j; s)} \right\}$$

for the vertices $v_i \in \mathbf{V}$. We will distinguish two subsets of vertices. The first $\mathbf{V_t}$ consists of those associated with transient states; the second $\mathbf{V_r}$, consists of recurrent states.

The α-order total Renyi entropy,[60] or "free information", of the measurement sequence up to n-cylinders is given by

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

where the partition function is

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

with the probabilities $p(s^n)$ defined on the n-cylinders $\{s^n\}$. The Renyi specific entropy, i.e. entropy per measurement, is approximated[27] from the n-cylinder distribution by

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

$$\text{or } h'_\alpha(n) = H_\alpha(n) - H_\alpha(n-1)$$

and is given asymptotically by

$$h_\alpha = \lim_{n \to \infty} h_\alpha(n)$$

The parameter α has several interpretations, all of interest in the present context. From the physical point of view, α $(= 1 - \beta)$ plays the role of the inverse temperature β in the statistical mechanics of spin systems.[39] The spin states correspond to measurements; a configuration of spins on a spatial lattice to a temporal sequence of measurements. Just as the temperature increases the probability of different spin configurations by increasing the number of available states, α accentuates different subsets of measurement sequences in the asymptotic distribution. From the point of view of Bayesian inference α is a Lagrange multiplier specifying a maximum entropy distribution consistent with the maximum likelihood distribution of observed cylinder probabilities.[41] Following symbolic dynamics terminology, $\alpha = 0$ will be referred to as the topological or counting case; $\alpha = 1$, as the metric or probabilistic case or high temperature limit. Varying α moves continuously from topological to metric machines. Originally in his studies of generalized information measures, Renyi introduced α as just this type of interpolation parameter and noted that the α-entropy has the character of a Laplace transform of a distribution.[60] Here there is the somewhat pragmatic, and possibly more important, requirement for α: it gives the proper algebra of trajectories in orbit space. That is, α is necessary for computing measurement sequence probabilities from the stochastic connection matrix T_α. Without it, products of T_α fail to distinguish distinct sequences.

An ϵ-machine's structure determines several key quantities. The first is the stochastic DFA measure of complexity. The α-order graph complexity is defined as

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

where the probabilities p_v are defined on the vertices $v \in V$ of the ϵ-machine's l-digraph. The graph complexity is a measure of an ϵ-machine's information processing capacity in terms of the amount of information stored in the morphs. As mentioned briefly later, the complexity is related to the mutual information of the past and future semi-infinite sequences and to the convergence[26, 28] of the entropy estimates $h_\alpha(n)$. It can be interpreted, then, as a measure of the amount of mathematical work necessary to produce a fluctuation from asymptotic statistics.

The entropies and complexities are dual in the sense that the former is determined by the principal eigenvalue λ_α of T_α,

$$h_\alpha = (1-\alpha)^{-1} \log_2 \lambda_\alpha$$

and the latter by the associated left eigenvector of T_α

$$\vec{p}_\alpha = \{p_v^\alpha : v \in \mathbf{V}\}$$

that gives the asymptotic vertex probabilities.

The specific entropy is also given directly in terms of the stochastic connection matrix transition probabilities

$$h_\alpha = \sum_{v \in \mathbf{V}} \frac{p_v^\alpha}{1-\alpha} \log \sum_{\substack{v' \in \mathbf{V} \\ s \in \mathbf{A}}} p^\alpha(v|v';s)$$

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

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

$\vec{p}_e$ is given by the left eigenvector of the ϵ-machine's edge graph. The transition complexity C_α^e is simply related to the entropy and graph complexity by

$$C_\alpha^e = C_\alpha + h_\alpha$$

There are, thus, only two independent quantities for a finite DFA ϵ-machine.[18]

The two limits for α mentioned above warrant explicit discussion. For the first, topological case ($\alpha = 0$), T_0 is the 1-digraph's connection matrix. The Renyi entropy $h_0 = \log \lambda_0$ is the topological entropy h. And the graph complexity is

$$C_0(G) = \log |\mathbf{V}|$$

This is $C(\mathrm{s}|\mathrm{DFA})$: the size of the minimal DFA description, or "program", required to *produce* sequences in the observed measurement language of which s is a member. This topological complexity counts all of the reconstructed states. It is similar to the regular language complexity developed for cellular automaton generated spatial patterns.[70] The DFAs in that case were constructed from known equations of motion and an assumed neighborhood template. Another related topological complexity counts just the recurrent states $\mathbf{V}_\mathbf{r}$. The distinction between this and C_0 should be clear from the context in which they are used in later sections.

In the second, metric case ($\alpha = 1$), h_α becomes the metric entropy

$$h_\mu = \lim_{\alpha \to 1} h_\alpha = -\frac{d\lambda_\alpha}{d\alpha}$$

The metric complexity

$$C_\mu = \lim_{\alpha \to 1} C_\alpha = -\sum_{v \in \mathbf{V}} p_v \log p_v$$

is the Shannon information contained in the morphs.[47] Following the preceding remarks, the metric entropy is also given directly in terms of the stochastic connection matrix

$$h_\mu = \sum_{v \in \mathbf{V}} p_v \sum_{\substack{v' \in \mathbf{V} \\ s \in \mathbf{A}}} p(v|v';s) \log p(v|v';s)$$

A central requirement in identifying models from observed data is that a particular inference methodology produces a sequence of hypotheses that converge to the correct one describing the underlying process. The complexity can be used as a diagnostic for this since it is a direct

[47] Cf. "set complexity" version of the regular language complexity[35] and "diversity" of undirected, unlabeled trees.[4]

measure of the size of the hypothesized stochastic DFA at a given reconstruction cylinder length. The identification method outlined in the preceding section converges with increasing cylinder length if the rate of change of the complexity vanishes. If, for example,

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

vanishes, then the noisy dynamical system has been identified. If it does not vanish, then c_α is a measure of the rate of divergence of the model size and so quantifies a higher level of computational complexity. In this case, the model basis must be augmented in an attempt to find a finite description at some higher level. The following sections will demonstrate how this can happen. A more complete discussion of reconstructing various hierarchies of models is found elsewhere.[19]

Period-Doubling Cascades

To give this general framework substance and to indicate the importance of quantifying computation in physical processes, the following sections address a concrete problem: the complexity of cascade transitions to chaos. The onset of chaos often occurs as a transition from an ordered (solid) phase of periodic behavior to a disordered (gas) phase of chaotic behavior. A cascade transition to chaos consists of a convergent sequence of individual "bifurcations", either pitchfork (period-doubling) in the periodic regimes or band-merging in the chaotic regimes.[48]

The canonical model class of these transitions is parametrized two-lap maps of the unit interval, $x_{n+1} = f(x_n)$, $x_n \in [0,1]$, with negative Schwartzian derivative; that is, those maps with two monotone pieces and admitting only a single attractor. We assign to the domain of each piece the letters of the binary alphabet $\Sigma = \{0,1\}$. The sequence space Σ^* consists of all 0-1 sequences. Some of these maps, such as the piecewise-linear tent map described in a later

[48] The latter are not, strictly speaking, bifurcations in which an eigenvalue of the linearized problem crosses the unit circle. The more general sense of bifurcation is nonetheless a useful shorthand for qualitative changes in behavior as a function of a control parameter.

section, need not have the period-doubling portion of the cascade. Iterated maps are canonical models of cascade transitions in the sense that the same bifurcation sequence occurring in a set of nonlinear ordinary differential equations (say) is topologically equivalent to that found in some parametrized map.[13, 37, 32]

Although ϵ-machines were developed in the context of reconstructing computational models from data series, the underlying theory provides an analytic approach to calculating entropies and complexities for a number of dynamical systems. This allows us to derive in the following explicit bounds on the complexity and entropy for cascade routes to chaos.

We focus on the periodic behavior near pitchfork bifurcations and chaotic behavior at band-mergings with arbitrary basic periodicity.[21, 23] In distinction to the description of universality of the period-doubling route to chaos in terms of parameter variation,[31] we have found a phase transition in complexity that is not explicitly dependent on control parameters.[29] The relationship between the entropy and complexity of cascades can be said to be super-universal in this sense. This is similar to the topological equivalence of unimodal maps of the interval,[55, 51, 52, 36, 12] except that it accounts for statistical and computational structures associated with the behavior classes.

In this and the next sections we derive the total entropy and complexity as a function of cylinder length n for the set of ϵ-machines describing the behavior at the different parameter values for the period-doubling and band-merging cascades. The sections following this then develop several consequences, viz. the order and the latent complexity of the cascade transition. With these statistical mechanical results established, the discussion turns to a detailed analysis of the higher level computation at the transition itself.

In the periodic regime below the periodicity $q = 1$ cascade transition we find the ϵ-machines for m-order period-doubling $2^m \to 2^{m+1}$ $(m = 0, 1, 2, 3)$ shown in figures 29 - 32.

For periodic behavior the measure on the n-cylinders $\{s^n\}$ is uniform; as is the measure on the recurrent ϵ-machine states $\mathbf{V}_r$. Consider behavior with period $P = q \times 2^m$ at a given

m-order period-doubling with basic cascade periodicity q. The uniformity allows us to directly estimate the total entropy in terms of the number $N(n,m)$ of n-cylinders with $n > P$

$$\begin{aligned} H_\alpha(n,m) &= (1-\alpha)^{-1} \log \sum_{s^n \in \{s^n\}} p^\alpha(s^n) \\ &= (1-\alpha)^{-1} \log \frac{N(n,m)}{N^\alpha(n,m)} \\ &= \log N(n,m) \end{aligned}$$

For periodic behavior and assuming$n > P$ the number of n-cylinders is given by the period $N(n,m) = P$. The total entropy is then $H_\alpha(n,m) = \log P$. Note that, in this case, h_α vanishes.

Similarly, the complexity is given in terms of the number $V_r = |\mathbf{V}_\mathbf{r}|$ of recurrent states

$$\begin{aligned} C_\alpha &= (1-\alpha)^{-1} \log \sum_{v \in \mathbf{V}} p_v^\alpha \\ &= (1-\alpha)^{-1} \log |\mathbf{V}_\mathbf{r}|^{1-\alpha} \\ &= \log V_r \end{aligned}$$

The number V_r of vertices is also given by the period for periodic behavior and so we find $C_\alpha = \log P$. Thus, for periodic behavior the relationship between the total and specific entropies and complexity is simple

$$C_\alpha = H_\alpha$$

$$\text{or } C_\alpha = n h_\alpha(n)$$

This relationship is generally true for periodic behavior and is not restricted to the situation where dynamical systems have produced the data. Where noted in the following we will also use $C_0 = \log |\mathbf{V}|$ to measure the total number of machine states.

Chaotic Cascades

In the chaotic regime the situation is much more interesting. The ϵ-machines at periodicity $q = 1$ and m-order band-mergings$2^m \to 2^{m-1}$, $m = 0, 1, 2, 3$, are shown in figures 33 - 36.

The graph complexity is still given by the number V_r of recurrent states as above. The main analytic task comes in estimating the total entropy. In contrast to the periodic regime

the number of distinct subsequences grows with n-cylinder length for all n. Asymptotically, the growth rate of this count is given by the specific topological entropy. In order to estimate the total topological entropy at finite n, however, more careful counting is required than in the periodic case. This section develops an exact counting technique for all cylinder lengths that applies at chaotic parameter values where the orbit $f^n(x^*)$ of the critical point x^*, where $f'(x^*) = 0$, is asymptotically periodic. These orbits are unstable and embedded in the chaotic attractor. The set of such values is countable. At these (Misiurewicz) parameters there is an absolutely continuous invariant measure.[54]

There is an additional problem with the arguments used in the periodic case. The uniform distribution of cylinders no longer holds. The main consequence is that we cannot simply translate counting $N(n,m)$ directly into an estimate of $H_{\alpha\neq 0}(n,m)$. One measure of the degree to which this is the case is given by the difference in the topological entropy h and the metric entropy h_μ.[27]

Approximations for the total Renyi entropy can be developed using the exact cylinder counting methods outlined below and the machine state and transition probabilities from $\left\{T_\alpha^{(s)}\right\}$. The central idea for this is that the states represent a Markov partition of the symbol sequence space Σ^*. There are invariant subsets of Σ^*, each of which converges at its own rate to "equilibrium". Each subset obeys the Shannon-McMillan theorem[7] individually. At each cylinder length each subset is associated with a machine state. And so the growth in the total entropy in each subset is governed by the machine's probabilistic properties. Since the cylinder counting technique captures a sufficient amount of the structure, however, we will not develop the total Renyi entropy approximations here and instead focus on the total topological entropy.

We now turn to an explicit estimate of $N(n,m)$ for various cases. Although the techniques apply to all Misiurewicz parameters, we shall work through the periodicity $q = 1\ 2 \rightarrow 1, 4 \rightarrow 2$, and $1 \rightarrow 0$ band-merging transitions (figure 33 - 36) in detail, and then quote the general formula for arbitrary order of band-merging.

The tree for $2 \rightarrow 1$ band merging n-cylinders is shown in figure 37.

An exact expression for $N(n,1)$ derives from splitting the enumeration of unique n-cylinders as represented on the tree into recurrent and transient parts. For two bands, Figure 37 illustrates the transient spine, the set of tree nodes associated with transient graph states, while schematically collapsing that portion of the tree associated with asymptotic graph vertices. The latter is shown in Figure 38. As will become clear the structure of the transient spine in the tree determines the organization of the counting method.

The sum for the n^{th} level, i.e. for the number of n-cylinders, is

$$N(n,1) = 1 + \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} 2^i + \sum_{i=0}^{\lfloor \frac{n-1}{2} \rfloor} 2^i$$

where $\lfloor k \rfloor$ is the largest non-negative integer less than k. The second term on the right counts the number of tree nodes that branch at even numbered levels, the third term is the number that branch at odd levels, and the first term counts the transient spine that adds a single cylinder. For $n \geq 2$ and even, this can be developed into a renormalized expression that yields a closed form as follows

$$\begin{aligned} N(n,1) &= 1 + 2\sum_{i=0}^{\frac{n-2}{2}} 2^i \\ &= 1 + 2\left(1 + 2\sum_{i=0}^{\frac{n-2}{2}} 2^i - 2 \cdot 2^{\frac{n-2}{2}}\right) \\ &= 1 + 2\left(N(n,1) - 2^{\frac{n}{2}}\right) \\ \text{or } \; N(n,1) &= 2\left(2^{\frac{n}{2}} - 2^{-1}\right) \end{aligned}$$

For $n \geq 2$ and odd, we find $N(n,1) = 3 \cdot 2^{\frac{n-1}{2}} - 1$. This gives an upper bound on the growth envelope as a function of n. The former, a lower bound.

The analogous expression for the $4 \rightarrow 2$ band cylinder count can be explicitly developed. Figure 39 shows the transient spine on the tree that determines the counting structure. In this

case, the sum is

$$N(n,2) = 2 + 2^{\lfloor \frac{n-4}{4} \rfloor} + 2^{\lfloor \frac{n-4}{4} \rfloor} + \sum_{i=0}^{\lfloor \frac{n-3}{4} \rfloor} 2^i + \sum_{i=0}^{\lfloor \frac{n-5}{8} \rfloor} 2^i + \sum_{i=0}^{\lfloor \frac{n-4}{4} \rfloor} 2^i + \sum_{i=0}^{\lfloor \frac{n-6}{8} \rfloor} 2^i$$

There are seven terms on the right hand side. In order they account for

1. The two transient cycles, begun on 0 and 1, each of which contributes 1 node per level;
2. Cycles on the attractor that are fed into the attractor via non-periodic transients (second and third terms);
3. Sum over tree nodes that branch by a factor of 2 at level $k + 4i, k = 3,4,5,6$, respectively.

The sum greatly simplifies upon rescaling the indices to obtain a self-similar form. For $n \geq P = 4$ and $n = 4i$, we find

$$\begin{aligned} N(n,2) &= 2\left(1 + 2^{\frac{n-4}{4}} + \sum_{i=0}^{\frac{n-4}{4}} 2^i + \sum_{i=0}^{\frac{n-8}{8}} 2^i\right) \\ &= 2 + 4\left(1 + \sum_{i=0}^{\frac{n-4}{4}} 2^i\right) \\ &= 2 + 4\left(1 + 2\sum_{i=0}^{\frac{n-4}{4}} 2^i - 2^{\frac{n}{4}}\right) \\ &= 2 + 2\left(N(n,2) - 2^{\frac{n+4}{4}}\right) \end{aligned}$$

$$\text{or } N(n,2) = 2^2 2^{\frac{n}{4}} - 2$$

There are three other phases for the upper bound as a function of n.

For completeness we note that this approach also works for the single band ($m = 0$) case

$$\begin{aligned} N(n,0) &= 1 + \sum_{i=0}^{n-1} 2^i \\ &= 1 + \left(2 + 2\sum_{i=0}^{n-1} 2^i - 2^n - 1\right) \\ &= 2N(n,0) - 2^n \end{aligned}$$

$$\text{or } N(n,0) = 2^n$$

The preceding calculations were restricted by the choice of a particular phase of the asymptotic cycle at which to count the cylinders. With a little more effort a general expression for all phases is found. Noting the similarity of the l-digraph structures between different order band-mergings and generalizing the preceding recursive technique yields an expression for arbitrary order band-merging. This takes into account the fact that the generation of new n-cylinders via branching occurs at different phases on the various limbs of the transient spine. The number of n-cylinders from the exact enumeration for the $q = 1\ 2^m \rightarrow 2^{m-1}$ band-merging is

$$N(n,m) = \begin{cases} 2^n & m = 0 \\ 2^m\left(b_{n,m}2^{n2^{-m}} - 2^{-1}\right) & m \neq 0 \end{cases}$$

where $n > P = 2^m$ and $b_{n,m} = (1+\hat{n})2^{-\hat{n}}$ and $\hat{n} = 2^{-m}(n \bmod 2^m)$ account for the effect of relative branching phases in the spine. This coefficient is bounded

$$b_{min} = \inf_{\{n,m\neq 0\}} b_{n,m} = 1$$
$$b_{max} = \sup_{\{n,m\neq 0\}} b_{n,m} = 3\cdot 2^{-\frac{3}{2}} \approx 1.0606602$$

The second bound follows from noting that the maximum occurs when, for example, $n = 2^m + 2^{m-1}$. Note that the maximum and minimum values of the prefactor are independent of the phase and of n and m. We will ignore the detailed phase dependence and simply write b instead of $b_{n,m}$ and consider the lower bound case of $b = 1$.

Recalling that $C_0 = \log|V_r| = m$, we have

$$N(n) = 2^{C_0}\left(b2^{n2^{-C_0}} - 2^{-1}\right)$$

and the total (topological) entropy is given by

$$H_0(n) = \log_2 N(n)$$
$$H_0(n) = C_0 + \log_2\left(2^{n2^{-C_0}} - 2^{-1}\right)$$

where we have set $b = 1$. The first term recovers the linear interdependence that derives from the asymptotic periodicity; cf. the period-doubling case. The second term is due to the additional

feature of chaotic behavior that, in the band-merging case, is reflected in the branching and transients in the l-digraph structure. In terms of the modeling decomposition introduced at the beginning, the first term corresponds to the periodic process $\mathbf{P}_t$ and the branching portion of the second term, to components isomorphic to the Bernoulli process $\mathbf{B}_t$.

From the development of the argument, we see that the factor 2^{-m} in the exponent controls the branching rate in the asymptotic cycle and so should be related to the rate of increase of the number of cylinders. The topological entropy is the growth rate of H_0 and so can now be determined directly

$$h_0(m) = \lim_{n \to \infty} \frac{H_0(n)}{n} = 2^{-m}$$

Rewriting the general expression for the lower bound in a chaotic cascade makes it clear how h_0 controls the total entropy

$$N(n, m) = V_r \left(2^{nh} - 2^{-1} \right)$$

where $h = \frac{f}{V_r}$ is the branching ratio of the number of vertices f that branch to the total number V_r of recurrent states.

The above derivation used periodicity $q = 1$. For general periodicity band-merging, we have $V_r = q \cdot 2^m$ and $f = 1$. It is clear that the expression works for a much wider range of ϵ-machines with isolated branching within a cycle that do not derive from cascade systems. Indeed, the results concern the relationship between eigenvalues and asymptotic state probabilities in the family of labeled Markov chains with isolated branching among cyclic recurrent states.

As a subset of all Misiurewicz parameter values, band-merging behavior has the simplest computational structure. In closing this section, we should point out that there are other cascade-related families of Misiurewicz parameters whose machines are substantially more complicated in the sense that the stochastic element is more than an isolated branching. Each family is described by starting with a general labeled Markov chain as the lowest order machine.

The other family members are obtained by applications of a period-doubling operator.[13] Each is a product of a periodic process and the basic stochastic machine. As a result of this simple decomposition, the complexity-entropy analysis can be carried out. This will be reported elsewhere. It explains many of the complexity-entropy properties above the lower bound case of band-merging. The numerical experiments later give examples of all these types of behavior.

Cascade Phase Transition

The preceding results are used in this section to demonstrate that the cascade route to chaos has a complexity-entropy phase transition. It was established some time ago that this route to chaos is a phase transition as a function of a nonlinearity parameter,[31] with an external (dis)ordering field[23] and a natural (dis)order parameter.[20] Here we focus on the information processing properties of this transition. First, we estimate for finite cylinder lengths the complexity and specific entropy at the transition. Second, we define and compute the transition's latent complexity that gives the computational difference between ϵ-machines above and below the transition. Finally, we discuss the transition's order.

Given the lower bound expressions for the entropy and complexity above and below the transition to chaos as a function of cylinder length n, we can easily estimate the complexities $C'(n)$ and $C''(n)$ and the critical entropy $H_c(n)$. Figure 40 gives a schematic representation of the transition and shows the definitions of the various quantities. The transition is defined as the divergence in the slope of the chaotic branch of the complexity-entropy curve. That is, the critical entropy H_c and complexity C' are defined by the condition

$$\frac{\partial H}{\partial C} = 0$$

From this, we find

$$C' = \log_2 n - \log_2 \log_2 y$$

$$nH_c = C' + \log_2 \left(by - 2^{-1} \right)$$

where $y = 2^{n2-C'}$ is the solution of

$$y \log_e y - y + \frac{1}{2} = 0$$

that is, $y \approx 2.155535035$. Numerical solution for $n = 16$ gives

$$C'(16) \approx 3.851982$$

$$C''(16) \approx 4.579279$$

$$H_c(16) \approx 0.286205$$

at $b = 1$.

The latent complexity ΔC of the transition we define as the difference at the critical entropy H_c of the complexities on the periodic and chaotic branches

$$\Delta C = C'' - C'$$

Along the periodic branch the entropy and complexity are equal and so from the previous development we see that

$$nH_c = C'' = C' + \log_2 \left(by - \frac{1}{2} \right)$$

$$\text{or } \Delta C = \log_2 \left(by - \frac{1}{2} \right)$$

For $b = 1$ this gives by numerical solution

$$\Delta C \approx 0.7272976887 \text{ bit}$$

which, we note, is independent of cylinder length.

In classifying this transition thermodynamically, the complexity plays the role of a heat capacity. It is by our definition a computational "capacity". Just as the thermodynamic temperature controls the multiplicity of available states, H appears as an "informational" temperature and H_c as a critical amount of information (energy) per symbol (spin) at which long range fluctuations occur. The overall shape is then similar to a lambda phase transition in

that there is a gradual increase in the capacity from both sides and a jump discontinuity in it at the transition. The properties supporting this follow from the bounds developed earlier. And so, there is at least one component of the cascade transition that is a second order transition, i.e. that associated with periodicity $q = 1$. There is also a certain degeneracy due to the phase dependence of the coefficient $b_{n,m}$. This is a small effect, but it does indicate a range of different limiting values as $n \to \infty$ for the chaotic critical complexity C'. It does not change the order of the transition. To completely characterize the transition, though, an upper bound on complexity at fixed n is also needed. This requires accounting for the typical chaotic parameters, by which we mean those associated with aperiodic behavior of the critical point. An approach to this problem will be reported elsewhere.

It should also be emphasized that the above properties were derived for finite cylinder lengths; that is, far away from the thermodynamic limit of infinite cylinders. The overall shape and qualitative properties hold not only in the thermodynamic limit but also at each finite size. In the thermodynamic limit the entropy estimates $n^{-1}H(n)$ go over to the entropy growth rates h_α. As a result, all of the periodic behavior lies on the $h_\alpha = 0$ line in the (h_α, C_α)-plane. This limiting behavior is consistent with a zero temperature phase transition of a one-spatial-dimension spin system with finite interaction range.

This analysis of the cascade phase transition should be contrasted with the conventional descriptions based on correlation function and mutual information decay. The correlation length of a statistical mechanical system is defined most generally as the minimum size L at which there is no qualitative statistical difference between the system of size L and the infinite (thermodynamic limit) system. This is equivalent in the present context to defining a correlation length L_α at which L-cylinder α-order statistics are close to asymptotic.[49] If we consider the total entropy $H_\alpha(L)$ as the (dis)order parameter of interest, then for finite ϵ-machines,[50] away from the

[49] Cf. the entropy "convergence knee" n^*.[28]

[50] The statistical mechanical argument, from which the following is taken, equivalently assumes exponential decay of the correlation function.

transition on the chaotic side, we expect its convergence to asymptotic statistics to behave like

$$2^{H_\alpha(L)} \propto 2^{\frac{L}{L_\alpha}}$$

But for L sufficiently large

$$2^{H_\alpha(L)} \propto 2^{h_\alpha L}$$

where $h_\alpha = \log_2 \lambda_\alpha$. By this argument, the correlation length is simply related to the inverse of the specific entropy: $L_\alpha \propto h_\alpha^{-1}$. We would conclude, then, that the correlation function description of the phase transition is equivalent in many respects to that based on specific entropy.

Unfortunately, this argument, which is often used in statistical mechanics, confuses the rate of decay of correlation with the correlation length. These quantities are proportional only assuming exponential decay or, in the present case, assuming finite ϵ-machines. The argument does indicate that as the transition is approached the correlation length diverges since the specific entropy vanishes. For all behavior with zero metric entropy, periodic or exactly at the transition, the correlation length is infinite. As typically defined, it is of little use in distinguishing the various types of zero entropy behavior.

The correlation length in statistical mechanics is determined by the decay of the two-point autocorrelation function

$$C(L) = \langle s_i s_{i+L} \rangle = \frac{1}{N} \sum_{i=0}^{N-1} \left(s_i s_{i+L} - s_i^2 \right)$$

Its information theoretic analog is the two–point 1-cylinder mutual information

$$I_\alpha(\mathbf{s}_i, \mathbf{s}_{i+L}) = H_\alpha(\mathbf{s}_i) - H_\alpha(\mathbf{s}_{i+L}|\mathbf{s}_i)$$

where $\mathbf{s}_i$ is the i^{th} symbol in the sequence $\mathbf{s}$ and $H_\alpha(\cdot)$ is the Renyi entropy.[51] Using this to describe phase transitions is an improvement over the correlation function in that, for periodic

[51] The correlation length is most closely related to I_2.

data, it depends on the period P : $I_\alpha \propto \log P$. In contrast, the correlation function in this case does not decay and gives an infinite correlation length.

The convergence of cylinder statistics to their asymptotic (thermodynamic limit) values is most directly studied via the total excess entropy[29, 26, 57]

$$F_\alpha(L) = H_\alpha(L) - h_\alpha L$$

It measures the total deviation from asymptotic statistics, up to L-cylinders.[52] As $L \to \infty$, it measures the average mutual information between semi-infinite past and future sequences. It follows from standard information theoretic inequalities that the two-point 1-cylinder mutual information is an overestimate of the excess entropy and so of the convergence properties. In particular,

$$L^{-1}F_\alpha(L) \leq I_\alpha(\mathrm{s}_i, \mathrm{s}_{i+L})$$

since I_α ignores statistical dependence on the symbols between s_i and s_{i+L}. The DFA ϵ-machine complexity is directly related to the total excess entropy[29]

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

As a tool to investigate computational properties, the two-point mutual information is too coarse, since it gives at most an upper bound on the DFA complexity.

At the transition correlation extends over arbitrarily long temporal and spatial scales and fluctuations dominate. It is the latter that support computation at higher levels in Chomsky's hierarchy. The computational properties at the phase transition are captured by the diverging ϵ-machines' structure. To the extent that their computational structure can be analyzed, a more refined understanding of the phase transition can be obtained.

---

[52] A scaling theory for entropy convergence to the thermodynamic limit that includes the effect of extrinsic noise has been is described previously.[28]

Cascade Limit Language

The preceding section dealt with the statistical character of the cascade transition, but we actually have much more information available from the ϵ-machines. Although the DFA model diverges in size, its detailed computational properties at the phase transition reveal a finite description at a higher level in Chomsky's hierarchy. With this we obtain a much finer classification than is typical in phase transition theory.

The structure of the limiting machine can be inferred from the sequence of machines reconstructed at $2^m \rightarrow 2^{m+1}$ period-doubling bifurcation on the periodic side and from those reconstructed at $2^m \rightarrow 2^{m-1}$ band-merging on the chaotic side. (Compare figures 29 and 33, 30 and 34, 31 and 35, 32 and 36.) All graphs have transient states of pair-wise similar structure, except that the chaotic machines have a period 2^{m-1} unstable cycle. All graphs have recurrent states of period 2^m. In the periodic machines this cycle is deterministic. In the chaotic machines, although the states are visited deterministically, the edges have a single nondeterministic branching.

The order of the phase transition depends on the structural differences between the ϵ-machines above and below the transition to chaos. In general, if this structural difference alters the complexity at constant entropy, then the transition will be second order. At the transition to chaos via period doubling there is a difference in the complexities due to

1. The single vertex in the asymptotic cycle that branches; and
2. The transient 2^{m-1} cycle in the machines on the chaotic side.

At constant complexity the uncertainty developed by the chaotic branching and the nature of the transient spine determine the amount of dynamic information production required to make the change from predictable to chaotic ϵ-machines.

The following two subsections summarize results discussed in detail elsewhere.

Critical Machine

The machine M that accepts the sequences produced at the transition, although minimal, has an infinite number of states. The growth of machine size $|\mathbf{V}(L)|$ versus reconstruction cylinder size L at the transition is demonstrated in figure 41. The maximum growth is linear with slope $c_0 = 3$. Consequently, the complexity diverges logarithmically.[53] The growth curve itself is composed of pieces with alternating slope 2 and slope 4

$$|\mathbf{V}(L)| = \left\{ \begin{array}{ll} 2L & 2^i \leq L < 3 \cdot 2^{i-1} \\ 4L & 3 \cdot 2^{i-1} \leq L < 2^{i+1} \end{array} \right\}$$

The slope 2 learning regions correspond to inferring more of the states that link the upper and lower branches of the machine. (The basic structure will be made clearer in the discussion of figure 42 below.) The slope 4 regions are associated with picking up groups of states along the long deterministic chains that are the upper and lower branches. Recalling the definition of c_α in a previous section, we note that finite c_0 indicates a constant level of complexity using a more powerful computational model than $\{\mathbf{P}_t, \mathbf{B}_t\}$.

Self-similarity of machine structure at the limit is evident if the machine is displayed in its "dedecorated" form. A portion of the infinite l-digraph at the transition is shown in figure 42 in this form. A decoration of an l-digraph is the insertion of a deterministic chain of states between two states.[68] In a dedecorated l-digraph chains of states are replaced with a single edge labeled with the equivalent symbol sequence. In the figure structures with a chain followed by a single branching have been replaced with a single branching each of whose edges are labeled with the original symbol sequence between the states. The dedecoration makes the self-similarity in the infinite machine structure readily apparent.

The strict regularity in the limit machine structure indicates a uniformity in the underlying computation modeled at a higher level. Indeed, the latter can be inferred from the infinite machine by applying the equivalence class morph reconstruction algorithm to the machine itself.[54] The

[53] The total entropy also depends logarithmically on cylinder length.

[54] The general framework for reconstructing machines at different levels in a computational hierarchy is presented elsewhere.[19]

result is the non-DFA machine M_c shown in figure 43, where the states in dedecorated M (figure 42) are coalesced into new equivalence classes based on the subtree similarity applied to the sequence of state transitions. The additional feature that must be inferred, once this higher level machine is reconstructed, is the production rule for the edge labels. These describe strings that double in length according to the production $B \rightarrow BB'$, where B is a register variable and B' is the contents of B with the last symbol complemented. The production appends to the register's contents the string B'.

On a state transition the contents of the register are output either directly or as the string B'. The 1-digraph edges are labeled accordingly in the figure. On a transition from states signified by squares, the register production is performed first and then the transition is made. The machine begins in the start state with a "1" in the register.

M_c accepts the full language L_c produced at the transition including the transient strings with various prefixes. At its core, though, is the simple recursive production ($B \rightarrow BB'$) for the itinerary ω_c of the critical point x^*. We will now explore the structure of this sequence in more detail in order to see just what computational capabilities it requires. We shall demonstrate how and where it fits in the Chomsky hierarchy.

Critical Language and Grammar

Before detailing the formal language properties of the symbol sequences generated at the cascade transition, several definitions of restricted languages are in order. First, of course, is the critical language itself L_c which we take to be the set of all subsequences produced asymptotically by the dynamical system at the cascade transition. M_c is a deterministic acceptor of L_c. Second, the most restricted language, denoted L_1, is the sequence of the itinerary of the map's maximum x^*. That is,

$$L_1 = \{\omega_c : \omega = s_1 s_2 s_3 \cdots \text{ and } f^i(x^*) < x^* \Rightarrow s_i = 0, \text{ otherwise } s_i = 1\}$$

a single sequence. Third, a slight generalization of this, L_2, consists of all length 2^n subwords of ω_c that start at the first symbol

$$L_2 = \left\{ \omega : \omega = s_1 s_2 \cdots s_{2^i} \text{ , i} = 0, 1, 2, \ldots \text{ and } s_j = [w_c]_j \right\}$$

where $[\omega]_k = s_k$ if $\omega = s_1 s_2 s_3 \cdots s_k \cdots$. Finally, we define L_3 to be the set of subsequences of any length that start at the first symbol of ω_c

$$L_3 = \left\{ \omega : \omega = s_1 s_2 \cdots s_i \text{ , i} = 1, 2, 3, \ldots \text{ and } s_j = [w_c]_j \right\}$$

Note that L_c is the further generalization including subsequences that start at any symbol in ω_c

$$L_c = \left\{ \omega_k : \omega_k = s_1 s_2 \cdots s_i \text{ , i} = 1, 2, 3, \ldots \text{ and } s_j = [w_c]_{j+k},\ k \geq 0 \right\}$$

With these various languages, we can begin to delineate the formal properties of the transition behavior. First, we note that an infinite number of words occur in L_c even though the metric entropy is zero. Additionally, there are an infinite number of inadmissible sequences and so an infinite number of words in the complement language $\bar{L}_c$, i.e. words not in L_c. One consequence is that the transition is not described by a subshift of finite type since there is no finite list of words whose concatenation generates L_c.[3]

Second, in formal language theory "pumping lemmas" are used to prove that certain languages are not in some language class.[38] Typically this is tantamount to demonstrating that particular recurrence or cyclic properties of the class are not obeyed by sufficiently long words in the language in question. Regular languages (RL) are those accepted by DFAs. Using the pumping lemma for regular languages it is easy to demonstrate that $L \in \{L_1, L_2, L_3, L_c\}$ is not regular. This follows from noting that there is no length n such that each word $z \in L$ with $|z| \geq n$ can be broken into three subwords, $z = uvw$ with $|uv| \leq n$, where the middle (nonempty) subword can be repeated arbitrarily many times. That is, sufficiently long strings cannot be decomposed such that $z \in L \Rightarrow uv^i w \in L\ \forall i \geq 0$. In fact, no substrings can be arbitrarily pumped. The lack of such a cyclic property also follows from noting that in M

all the states are transient and there are no transient cycles. The observation of this structural property also leads to the conclusion that L_c is also not finitely-described at the next level of the complexity hierarchy: context-free languages (CFL), i.e. those accepted by pushdown automata. This can be established directly using the pumping lemma for context-free languages.

Third, in the structural analysis of M we found states at which the following production is applied: $A \rightarrow AA'$, where $A' = s_0 \cdots \bar{s}_k$ if $A = s_0 \cdots s_k$ and $\bar{s}$ is the complement of s. This production generates L_1 and L_2. It is most concisely expressed as a context-free Lindenmayer system.[49] The general class is called 0L grammars: $G = \{\Sigma, P, \alpha\}$ consisting of the symbol alphabet, production rules, and start string, respectively. This computational model is a class of parallel rewrite automata in which all symbols in a word have the production rules simultaneously applied, with the neighboring symbols playing no role in the selection of which production. The symbol alphabet is $\Sigma = \{0, 1\}$. The production rules P are quite simple$P = \{0 \rightarrow 11, 1 \rightarrow 10\}$ and start with the string $\alpha = \{1\}$. This system generates the infinite sequence L_1 and allowing the choice of when to stop the productions, it generates $L_2 = \{1, 10, 1011, 10111010, \ldots\}$.

Although the L-system model of the transition behavior is quite simple, as a class of models its parallel nature is somewhat inappropriate. L-systems produce both "early" and "late" symbols in a string at every production step; whereas the dynamical system in question produces symbols sequentially. This point is even more obvious when these symbol sequences are considered as sequential measurements. The associated L-system model would imply that the generating process had an infinite memory of past measurements and accessed them arbitrarily quickly. The model class is too powerful.

This can be remedied by converting the 0L-system to its equivalent in the Chomsky hierarchy of sequential computation.[38] The Chomsky equivalent is a restricted indexed context-free grammar $G_c = \{\mathrm{N}, \mathrm{I}, \mathrm{T}, \mathrm{F}, \mathrm{P}, S\}$.[1] A central feature of the indexed grammars is that they are a natural extension of the context-free languages that allow for a limited type of context-sensitivity via indexed productions, while maintaining properties of context-free languages,

such as closure and decidability, that are important for compilation. For the limit language the components are defined as follows. $\mathbf{N} = \{S, T\}$ is the set of nonterminal variables with S the start symbol; $\mathbf{I} = \{A, B, C, D, E, F\}$ is the set of intermediate variables; $\mathbf{T} = \{0, 1\}$ is the set of terminal symbols; $\mathbf{P} = \{S \rightarrow Tg, T \rightarrow Tf, T \rightarrow BA, C \rightarrow BB, D \rightarrow BA, E \rightarrow 0, F \rightarrow 1\}$ is the set of productions; and $\mathbf{F} = \{f, g\}$ with $f = [A \rightarrow C, B \rightarrow D]$ and $g = [A \rightarrow E, B \rightarrow F]$ are indexed productions. The grammar just given is in its "normal" form since the variables in the indexed productions $\mathbf{F}$ do not have productions in $\mathbf{P}$. The indexed grammar is restricted in that there are no intermediate variables with productions that produce new indices. The latter occurs only via the $S \rightarrow Tg$ and $T \rightarrow Tf$ productions. Note that once this is no longer used, via the application of $T \rightarrow BA$, no new indices appear.

The above indexed grammar sequentially produces symbols in words from L_2. Two example "left-most" derivations are

$$
\begin{aligned}
(1) \quad & S \rightarrow Tg \rightarrow BgAg \\
& \rightarrow FAg \rightarrow 1Ag \rightarrow 1E \rightarrow 10 \\
(2) \quad & S \rightarrow Tg \rightarrow Tfg \rightarrow BfgAfg \\
& \rightarrow DgAfg \rightarrow BgAgAfg \rightarrow FAgAfg \\
& \rightarrow 1AgAfg \rightarrow 1EAfg \rightarrow 10Afg \\
& \rightarrow 10Cg \rightarrow 10BgBg \rightarrow 10FBg \\
& \rightarrow 101Bg \rightarrow 101F \rightarrow 1011
\end{aligned}
$$

Productions are applied to the leftmost nonterminal in each step. Consequently, the terminal symbols $\{0, 1\}$ are produced sequentially left to right in "temporal" order. In the first line, notice how the indices distribute over the variables produced by the production $T \rightarrow BA$. When an indexed production is used an index is consumed: as in $Bg \rightarrow F$ in going from the first to the second line above.

All of the languages in the Chomsky hierarchy have dual representations as grammars and as automata. The machine corresponding to an indexed context-free language is the nested stack

automaton (NSA).[2] This is a generalization of the pushdown automaton: a finite state control augmented with a last-in first-out memory or stack. An NSA has the additional ability to move into the stack in a read-only mode and to insert a new (nested) stack at the current stack symbol being read. It cannot move higher in the stack until it has finished with the nested stack and removed it. The restricted indexed context-free grammar for L_2 is recognized by the one-way nondeterministic NSA (1NNSA) shown in figure 44. The start state is q. The various actions label the state transition edges. $ denotes the top of the current stack and the cent sign, the current stack bottom. The actions are one of three forms

1. $\alpha \rightarrow \beta$, where α and β are patterns of symbols on the top of the current stack;
2. $\alpha \rightarrow$ {1, -1}, where the latter indicates moving the head up and down the stack, respectively, upon seeing the pattern α at current stack top.
3. (t,$t) $\rightarrow$ (1,$), where t is a symbol read off of the input tape and compared to the symbol at the top of the stack. The '1' indicates that the input head advances to the next symbol on the input tape. The symbol on the stack's top is removed: $t $\rightarrow$ $.

In all but one case the actions are in the form of a symbol pattern on the top of the stack leading to a replacement pattern and a stack head motion. The notation on the figure uses a component-wise shorthand. For example, the productions are implemented on the transition labeled ${S,T,T,C,D,E,F} $\rightarrow$ ${Tg,Tf,BA,BB,BA,0,1} which is shorthand for the individual transitions: $S $\rightarrow$ $Tg, $T $\rightarrow$ $Tf, $T $\rightarrow$ $BA, $C $\rightarrow$ $BA, $D $\rightarrow$ $BB, $E $\rightarrow$ $0, and $F $\rightarrow$ $1. The operation of the 1NNSA mimics the derivations in the indexed grammar. The nondeterminism here means that there exists some set of transitions that will accept words from L_2. L_3 is accepted by the same 1NNSA, but modified to accept when the end of the input string is reached and the previous input has been accepted.

There are three conclusions to draw from these formal language results. First, it should be emphasized that the particular details in the preceding analysis are not essential. Rather, the most important remark is that the description at this higher level is finite and, indeed, quite

small. Despite the infinite DFA complexity, a simple higher level description can be found once the computational model is augmented. Indeed, the deterministic Turing machine program to generate words in the limit language is simple: (i) copy the current string on the tape onto its end and (ii) invert the last bit. The limit language for the cascade transition uses little of the power of the indexed grammars. The latter can recognize, for example, context-sensitive languages. The limit machine is thus exceedingly weak in its implied computational structure. Also, the only nondeterminism in the 1NNSA comes from anticipating the length of the string to accept; a feature that can be replaced to give a deterministic and so less powerful automaton.

Second, it is relatively straightforward to build a continuous-state dynamical system with an embedded universal Turing machine.[55] With this in mind, and for its own sake, we note that by the above construction the cascade transition does not have universal computation embedded in it. Indeed, it barely aspires to be much more than a context-free grammar. With the formal language analysis we have bounded the complexity at the transition to be greater than regular and context-free languages and no more powerful than indexed context-free. Furthermore, the complexity at this level is measured by a linearly bounded DFA growth rate $c_0 = 3$. These properties leave open the possibility, though, that the language could be a one-way nondeterministic stack automaton (1NSA).[38]

Finally, we demonstrated by an explicit analysis that nontrivial computation, beyond information storage and transmission, arises at a phase transition. One is forced to go beyond DFA models to the higher stack automaton level since the former require an infinite representation. These properties are only hinted at by the infinite correlation length and the slow decay of two-point mutual information at the transition.

Logistic Map

The preceding analysis holds for a wide range of nonlinear systems since it rests only on the

[55] A two-dimensional map with an embedded 4 symbol, 7 state universal Turing machine[53] was constructed.[15]

symbolic dynamics and the associated probability structure. It is worthwhile, nonetheless, to test it quantitatively on particular examples. This is possible because it rests on a (re)constructive method that applies to any data stream. This section and the next report extensive numerical experiments on two one-dimensional maps. The first is the logistic map, defined shortly, and the second, the piecewise linear tent map.

The logistic map is a map of the unit interval given by

$$x_{n+1} = rx_n(1 - x_n),\ x_0 \in [0,1] \text{ and } r \in [0,4]$$

where the parameter r controls the degree of nonlinearity. $\frac{r}{4}$ is the map's height at its maximum $x^* = \frac{1}{2}$. This is one of the simplest, but nontrivial, nonlinear dynamical systems. It is an extremely rich system about which much is known.[13] It is fair to say, however, that even at the present time there are still a number of unsolved mathematical problems concerning the behavior at arbitrary chaotic parameter values. The (generating) measurement partition is $P_{\frac{1}{2}} = \{[0,.5),[.5,1.]\}$.

The machine complexity and information theoretic properties of this system have been reported previously.[29] Figure 45 shows the complexity versus specific entropy for 193 parameter values $r \in [3,4]$. One of the more interesting general features of the complexity-entropy plot is clearly demonstrated by this figure: all of the periodic behavior lies below the critical entropy H_c; and all of the chaotic, above. This is true even if the periodic behavior comes from cascade windows of periodicity $q > 1$ within the chaotic regime at high parameter values. The (H_α, C_α) plot, therefore, captures the essential information processing, i.e. computation and information production, in the period-doubling cascade independent of any explicit system control.

The lower bound derived in the previous sections applies exactly to the periodic data ($H < H_c$) and to the band-merging parameter values. The fit to the periodic data is extremely accurate, verifying the linear relationship except for high periods beyond that resolvable at the chosen reconstruction cylinder length. The fit in the chaotic regime is also quite good. (See

figure 46.) The data are systematically lower (~2%) in entropy due to the use of the topological entropy in the analysis. The measured critical entropy H_c and complexity C'' at the transition were 0.28 and 4.6, respectively.

Tent Map

The second numerical example, the tent map, is in some ways substantially simpler than the logistic map. It is given by

$$x_{n+1} = \begin{cases} ax_n & x_n \leq \frac{1}{2} \\ a(1-x_n) & x_n > \frac{1}{2} \end{cases}$$

where the parameter a controls the height $(= \frac{a}{2})$ of the map at the maximum $x^* = \frac{1}{2}$. The main simplicity is that there is no period-doubling cascade and, for that matter, there are no stable periodic orbits, except at the origin for $a < 1$. There is instead only a periodicity $q = 1$ chaotic band-merging cascade that springs from x^* at $a = 1$.

The piecewise linearity also lends itself to further analysis of the dynamics. Since the map has the same slope everywhere, the Lyapunov exponent λ, topological, metric, and Renyi specific entropies are all equal and given by the slope $\lambda = h_\alpha = \log_2 a$. We can simply refer to these as the specific entropy. From this, we deduce that, since $h_\alpha = 2^{-m}$ for $2^m \to 2^{m-1}$ band-mergings, the parameter values there are

$$a_{2^m \to 2^{m-1}} = 2^{2^{-m}}$$

For $L \geq 2^m$ the complexity is given by the band-merging period. And this, in turn, is given by the number of bands. Thus, we have $C_\alpha = -\log_2 h_\alpha$ or

$$C_\alpha = -\log_2 \log_2 a$$

as a lower bound for $a > 1$ and $L > 2^m$ at an m-order band merging.

Since there is no stable periodic behavior, other than period one, there is a forbidden region in the complexity-entropy plot below the critical entropy. The system cannot exist at finite "temperatures" below H_c, except at absolute zero $H_c = 0$.

Figure 47 gives the complexity-entropy plot for 200 parameter values $a \in [1, 2]$. There is a good deal of structure in this plot beyond the simple band-merging lower bounds we have concentrated on. Near each band-merging complexity-entropy point, there is a slanted cluster of points. These are associated with families of parameter values at which the iterates $f^n(x^*)$ are asymptotically periodic of various periods. We shall discuss this structure elsewhere, except to note here that it also appears in the logistic map, but is substantially clearer in this example.

Figure 48 shows band-merging data estimated from 16- and 20- cylinders along with the appropriate theoretical curves for those and in the thermodynamic limit ($L = 256$).

From the two numerical examples it is clear that the theory quite accurately predicts the complexity-entropy dependence. It can be easily extended in several directions. Most notably, the families of Misiurewicz parameters associated with unstable asymptotically periodic maxima can be completely analyzed. And this appears to give some insight into the general problem of the measure of parameter values where iterates of the maximum are asymptotically aperiodic. Additionally, the computational analysis is being applied to transitions to chaos via intermittency and via frequency-locking.

Computation at Phase Transitions

To obtain a detailed understanding of the computational structure of a phase transition, we have analyzed one example of a self-similar family of attractors. The period-doubling cascade is just one of many routes to chaos. The entire family is of nominal interest, providing a rather complete analysis of a phase transition and how statistical mechanics applies. More importantly, for general phase transitions the approach developed here indicates a type of superuniversality that is based only on the intrinsic information processing performed by a dynamical or, for that matter, physical system. This information processing consists of conventional communication theoretic quantities, that is the storage and transmission of information, and the computational

aspects, most clearly represented by the detailed structure and formal language properties of reconstructed ϵ-machines.

By analyzing in some detail a particular class of nonlinear systems, we have attempted to strengthen the conjecture that it is at phase transitions where high level computation occurs. Application to other examples, such as the phase transition in the 2D Ising spin system and cellular automata and lattice dynamical systems generally, will go even further toward establishing this general picture. These applications will be reported elsewhere. Nonetheless, it is clear that computational ideas provide a new set of tools for investigating the physics of phase transitions. The central conclusion is that via reconstruction they can be moved out of the realm of mathematics and theoretical computer science and applied to the scientific study and engineering of complex processes.

The association of high level computation and phase transitions is not made in isolation. Indeed, we should mention some early work addressing similar questions. Type IV cellular automata (CA) were conjectured by Wolfram to support nontrivial and perhaps universal computation.[72] These CA exhibit long-lived transients and propagating structures out of which elementary computations can be constructed. The first author and Norman Packard of the University of Illinois conjectured some years ago that type IV behavior was generated by CA on bifurcation sets in the discretized space of all CA rules. This was suggested by studies of bifurcations in a continuous-state lattice dynamical system as a function of a nonlinearity parameter.[24] The continuous local states were discretized to give CA with varying numbers of states. By comparing across a range of state-discretization and nonlinearity parameter, the CA bifurcation sets were found to be associated with bifurcations in the continuous-state lattice system. More recent work by Chris Langton of Los Alamos National Laboratory has confirmed this in substantially more detail via Monte Carlo sampling of CA rule space. This work uses mutual information, not machine complexity, measures of the behavior. As pointed out above, there is an inequality relating these measures. Mutual information versus entropy

density plots for hundreds of CA rules reveal a phase transition structure similar to that shown in the complexity-entropy diagram of figure 45. Seen in this light, the present paper augments these experimental results with an analytic demonstration of what appears to be a very general organization of the space of dynamical systems, whether discrete or continuous.

Complexity of Critical States

Recall that the Wold-Kolmogorov spectral decomposition says the spectrum of a stationary signal has three components.[46, 47, 69] The first is a singular measure consisting of δ-functions. This describes periodic behavior. The second component is associated with an absolutely continuous invariant measure and so broadband power spectra. The final component is unspecified and typically ignored. From the preceding investigation, though, we can make a comment and a conjecture. The comment is that finite stochastic DFA ϵ-machines capture the first two components of the decomposition: $\mathbf{P}_t$ and $\mathbf{B}_t$, respectively. The conjecture, and perhaps more interesting remark, is that the third component appears to be associated with higher levels in the computational hierarchy.

This conjecture can be compared to recent discussion of ergodic theory. Ornstein suggested that most "chaotic systems that arise naturally are abstractly the same as $\mathbf{B}_t$".[56] We can now see in what sense this can be true. If it is only at accumulation points of bifurcations that infinite DFA machines occur, then in the space of all dynamical systems the dimensionality of the set of such systems will be reduced and so the set's measure will be zero. From this viewpoint, high complexity (non-$\mathbf{B}_t$) systems would be rare. We can also see how the conjecture can be false. If there is some constraint restricting the space of systems in which we are interested, then with respect to that space infinite machines might have positive measure and so be likely. Systems, such as those found in biology, that survive by adaptively modifying themselves to better model and forecast their environment would tend to exhibit high levels of complexity. Within the space of successful adaptive systems, high complexity presumably is quite probable.

Another sense in which Ornstein's conjecture is too simplified is that the Bernoulli shift is computationally simple. It is equivalent, in one representation, to the piecewise linear Baker's transformation of the torus. In contrast, most "natural" physical systems are modeled with smooth (non-piecewise-linear) nonlinearities. The associated physical properties contribute substantially to a system's ability to generate complex behavior independent of the complexity of boundary and initial conditions.[22] Physical systems governed by a chaotic piecewise linear dynamic simply excavate microscopic fluctuations, amplifying them to determine macroscopic behavior.[64] This information transmission across scales is computationally trivial. It is not the central property of complex behavior and structure.

We have seen that away from the cascade phase transition it is only at band-merging parameters where chaotic behavior can be factored into periodic and Bernoulli components. A similar decomposition of ϵ-machines into periodic and finite stochastic components occurs at Misiurewicz parameters, where $f^n(x^*)$ is asymptotically periodic and unstable and the invariant measure is absolutely continuous. But these parameter values are countable. Furthermore, the measure of chaotic parameter values as one approaches a Misiurewicz value is positive and so is uncountable.[40] Typical parameters in this set appear to be characterized by aperiodic $f^n(x^*)$ that are in a Cantor set not containing x^*. Such a case is modeled by infinite DFA ϵ-machines. Taken together these indicate that a large fraction of "naturally arising" chaotic systems are isomorphic neither to $\mathbf{B}_t$ nor to $\mathbf{B}_t \otimes \mathbf{P}_t$.

Computational ergodic theory, the application of computational analysis in ergodic theory, would appear to be a useful direction in which to search for rigorous refined classifications of complex behavior. This suggests, for example, the use of DFA and SA complexity, and other higher forms, as invariants to distinguish further the behavior of dynamical systems, such as K-flows and zero-entropy flows. For example, although described by a minimal DFA with a single state, the inequivalence of the binary $\mathbf{B}_2(\frac{1}{2}, \frac{1}{2})$ and ternary $\mathbf{B}_3(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ Bernoulli shifts follows from the fact that they are not structurally equivalent. Not only do the entropies

differ, but the machine for the first has two edges; for the second, three. Restating this in entropy-complexity notation

$$\text{although} \quad C_\alpha(\mathbf{B_2}) = C_\alpha(\mathbf{B_3}) = 0,$$

$$h_\alpha(\mathbf{B_2}) \neq h_\alpha(\mathbf{B_3})$$

$$\text{and} \quad C^e_\alpha(\mathbf{B_2}) \neq C^e_\alpha(\mathbf{B_3})$$

The full entropy-complexity plane appears as a useful first step toward a more complete classification. Recall the (H_α, C_α) plots for the logistic and tent maps. (See figures 45 and 47.) Similar types of structural distinctions and new invariants will play a central role in computational ergodic theory.

But what does this have to say about physical systems? In what sense can a turbulent fluid, a noisy Josephson junction, or a quantum field, be said to perform a computation? The answer is that while computational aspects appear in almost any process, since we can always estimate some low level ϵ-machine, only nontrivial computation occurs in physical systems on the order-disorder border. Additionally these systems have very special phase-transition-like, or "critical", subspaces. ϵ-machine theory gives a much more refined description of such critical behavior than that currently provided by statistical mechanics. Indeed, the well-known phase transitions should be re-examined in this light. In addition to a more detailed structural theory of the dynamic mechanisms responsible for phase transitions, such studies will give an improved understanding of the macroscopic thermodynamic properties required for computation.

Computers are, in this view, physical systems designed to be in a critical state. They are constructed to support arbitrarily long time correlations within certain macroscopic "computational" degrees of freedom. This is achieved by decoupling these degrees of freedom from error-producing heat bath degrees of freedom. Computers are physical systems designed to be in continual phase transition within entropic-disordered environments. From the latter they derive the driving force that moves computations forward. But, at the same time, they must shield the computation from environmentally induced fluctuations.

As already emphasized, the general measure of complexity introduced at the beginning is not limited to stochastic DFAs and SAs, but applies in principle to any computational level or, indeed, to any modeling domain where a "symmetry" can be factored out of data. In particular, this can be done hierarchically as we have shown in the analysis of the computational properties of the cascade critical machine. In fact, a general hierarchical reconstruction is available.[19] The abstract definition of complexity applies to all levels of the Chomsky hierarchy where each computation level represents, in a concrete way, a class of symmetries with respect to which observed data is to be "expanded" or modeled. This notion of complexity is also not restricted to the Chomsky hierarchy. It can be applied, for example, to spatially-extended or network dynamical systems. Since these are computationally equivalent to parallel and distributed machines, respectively, ϵ-machine reconstruction suggests a constructive approach to parallel computation theory. We hope to return to these applications in the near future.

Acknowledgements

The authors have benefited from discussions with Charles Bennett, Eric Friedman, Chris Langton, Steve Omohundro, Norman Packard, Jim Propp, Jorma Rissanen, and Terry Speed. They are grateful to Professor Carson Jeffries for his continuing support. The authors thank the organizers of the Santa Fe Institute Workshop on "Complexity, Entropy, and Physics of Information" (May 1989) for the opportunity to present this work, which was supported by ONR contract N00014-86-K-0154.

Figures

Figure Captions

Figure 28 The complexity spectrum: complexity C as a function of the diversity of patterns. The latter is measured with the (normalized) Shannon entropy H. Regular data have low entropy; very random data have maximal entropy. However, their complexities are both low.

Figure 29 Topological 1-digraph for period 1 attractor.

Figure 30 Topological 1-digraph for period 2 attractor.

Figure 31 Topological 1-digraph for period 4 attractor.

Figure 32 Topological 1-digraph for period 8 attractor.

Figure 33 Topological 1-digraph for single band chaotic attractor.

Figure 34 Topological 1-digraph for $2 \longrightarrow 1$ band chaotic attractor.

Figure 35 Topological 1-digraph for $4 \longrightarrow 2$ band chaotic attractor.

Figure 36 Topological 1-digraph for $8 \longrightarrow 4$ band chaotic attractor.

Figure 37 Parse tree associated with two chaotic bands merging into one. Tree nodes are shown for the transient spine only. The subtrees associated with asymptotic behavior, and so also with the equivalence classes corresponding to recurrent graph vertex 1 in figure 34, are indicated schematically with triangles.

Figure 38 Subtree of nodes associated with asymptotic vertices in 1-digraph for two bands merging to one.

Figure 39 Transient spine for 4 $\rightarrow$ 2 band attractor. The asymptotic subtrees are labeled with the associated l-digraph vertex. (Compare figure 35.)

Figure 40 Complexity versus specific entropy estimate. Schematic representation of the cascade lambda transition at finite cylinder lengths. Below H_c the behavior is periodic; above, chaotic. The latent complexity is given by the difference of the complexities C'' and C' at the transition on the periodic and chaotic branches, respectively.

Figure 41 Growth of critical machine M. The number $|\mathbf{V}(L)|$ of reconstructed states versus cylinder length L for the logistic map at the periodicity $q = 1$ cascade transition. Reconstruction is from length 1 to length 64 cylinders on $2L$-cylinder trees.

Figure 42 Self-similarity of machine structure at cascade limit is shown in the dedecorated l-digraph of M.

Figure 43 Higher level production-rule machine, or stack automaton, M_c that accepts L_c.

Figure 44 One-way nondeterministic nested stack automaton for limit languages L_2 and L_3.

Figure 45 Observed complexity versus specific entropy estimate for the logistic map at 193 parameter values $r \in [3, 4]$ within both periodic and chaotic regimes. Estimates on 32-cylinder trees with 16-cylinder subtree machine reconstruction; where feasible.

Figure 46 Fit of logistic map periodic and chaotic data to corresponding functional forms. The data is from the periodicity 1 band-merging cascade and also includes all of the periodic data found in the preceding figure. The theoretical curves $C_0(H_0)$ are shown as solid lines.

Figure 47 Tent map complexity versus entropy at 200 parameter values $a \in [1, 2]$. The quantities were estimated with 20-cylinder reconstruction on 40-cylinder trees; where feasible.

Figure 48 Effect of cylinder length. Tent map data at 16- and 20-cylinders (triangle and square tokens, respectively) along with theoretical curves $C_0(H_0)$ for the same and in the thermodynamic limit ($L = 256$).

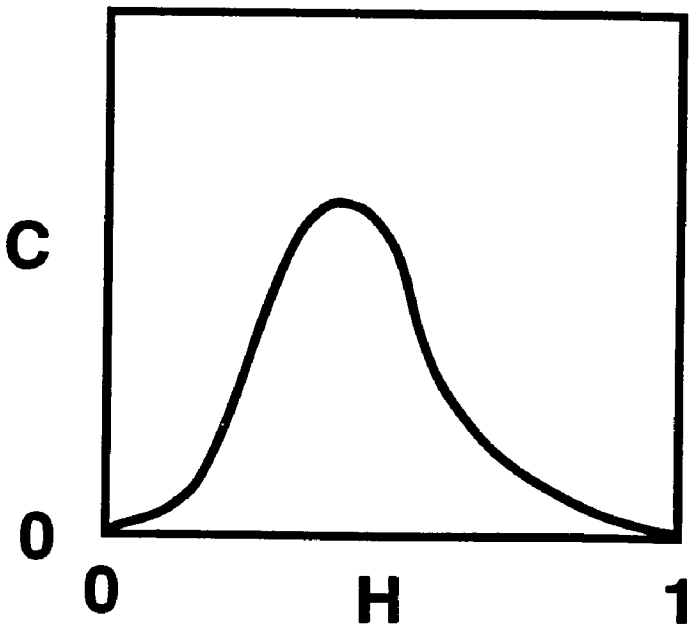


Figure 28

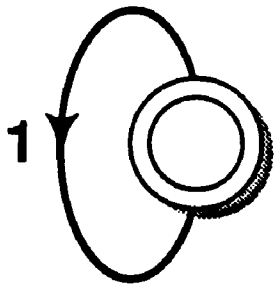


Figure 29

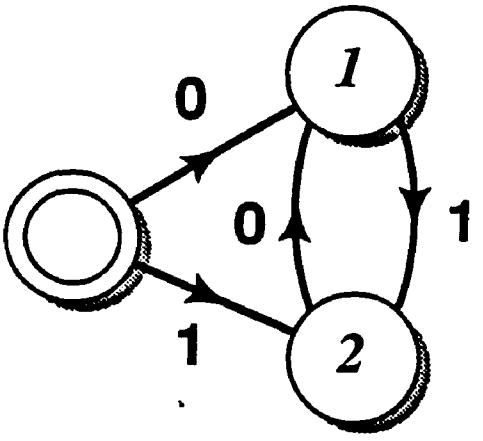


Figure 30

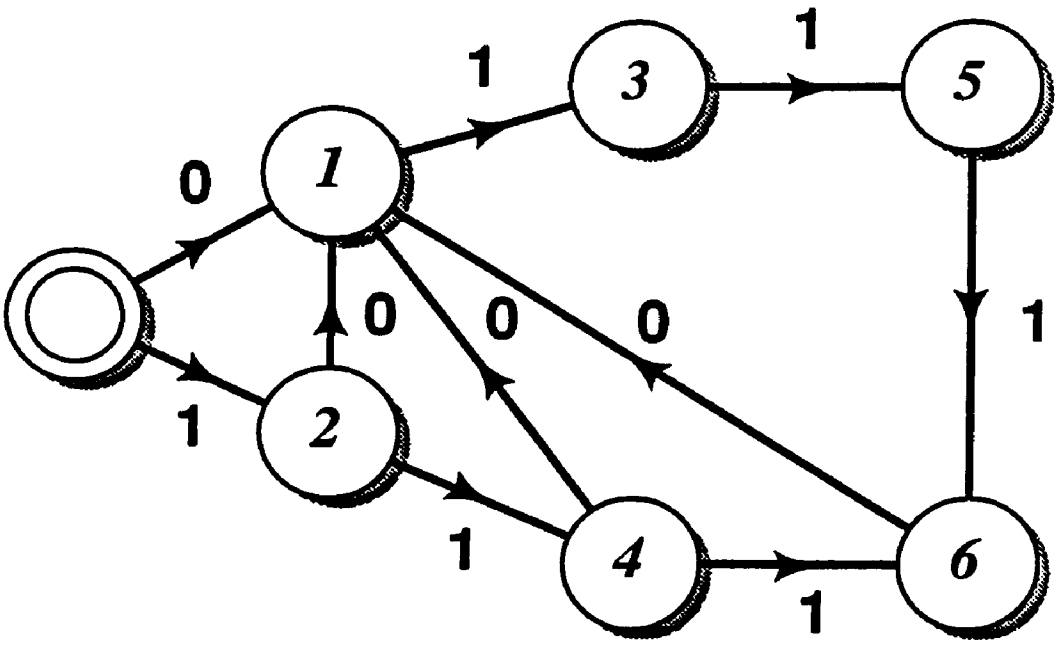


Figure 31

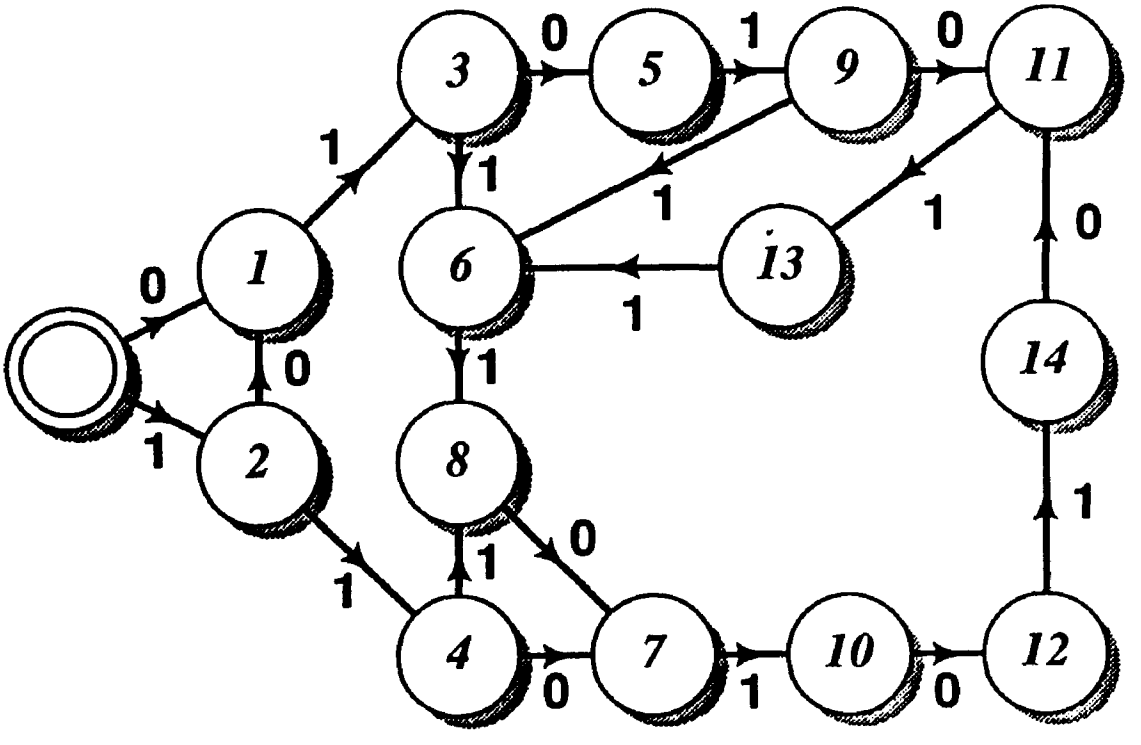


Figure 32

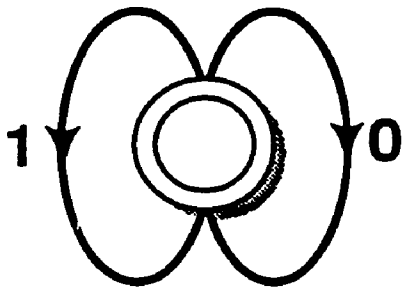


Figure 33

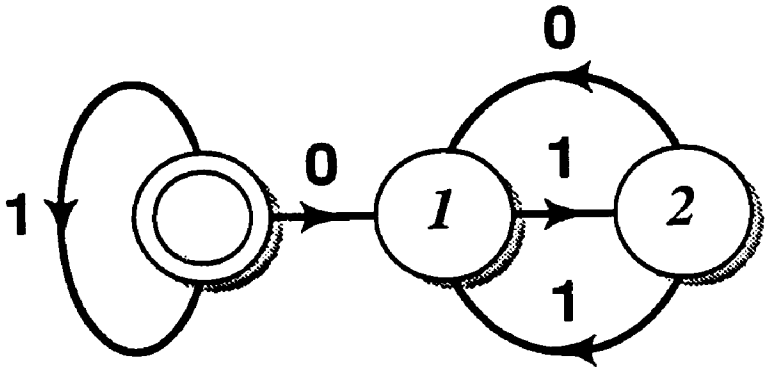


Figure 34

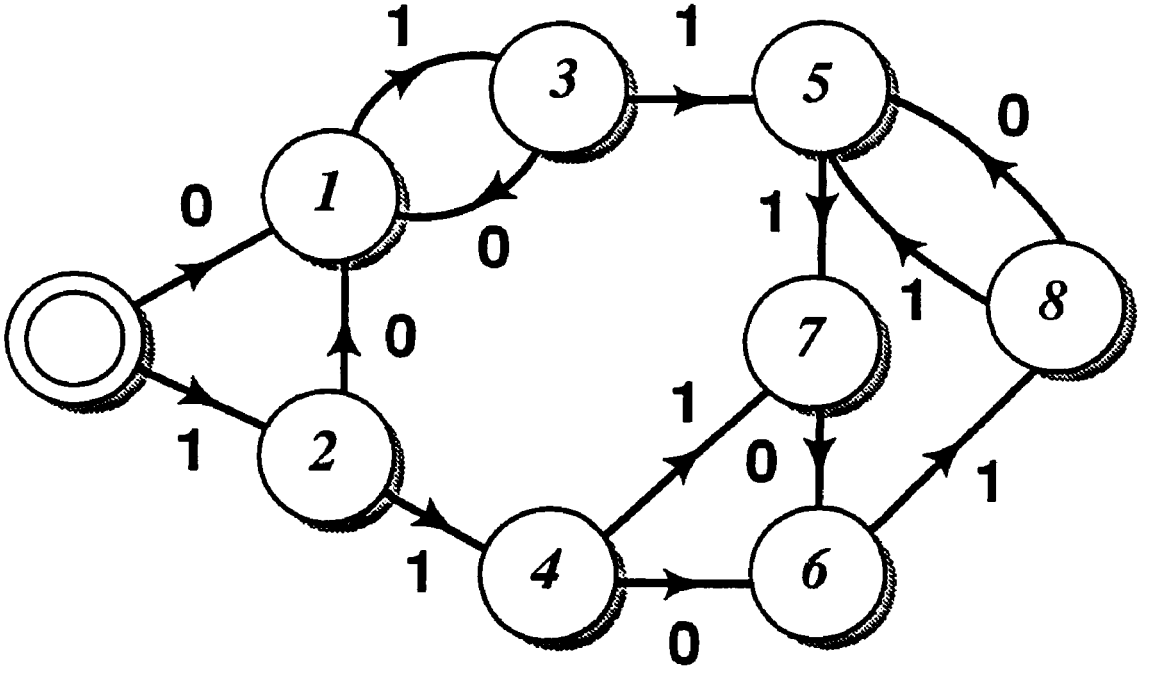


Figure 35

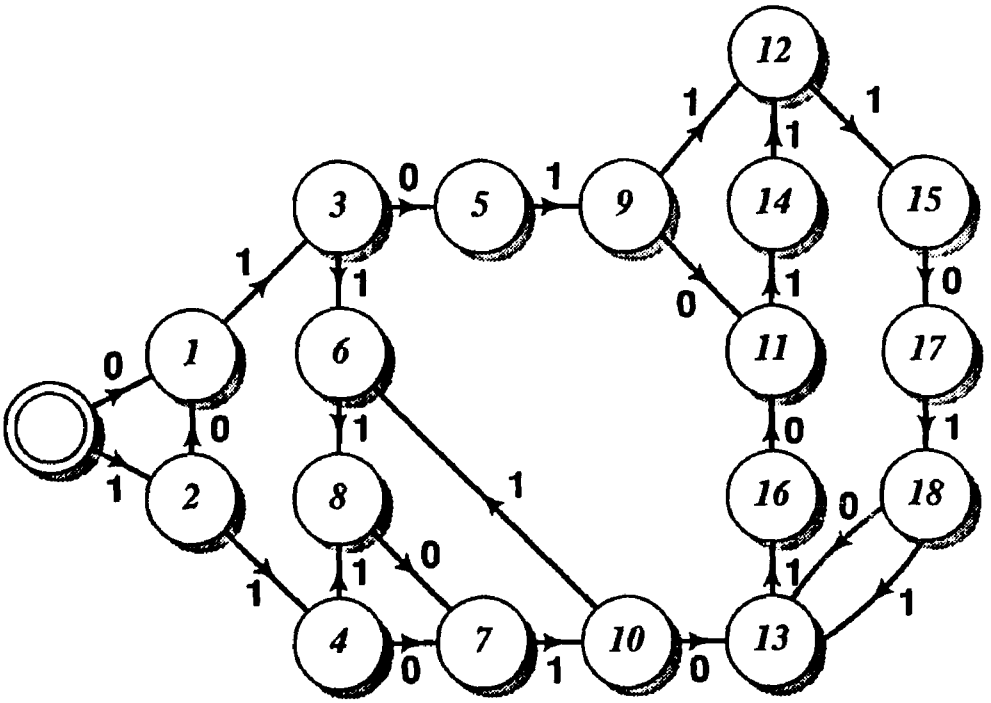

12
1
1
1
3
0
5
1
9
14
15
1
1
0
0
1
6
11
17
0
1
0
1
1
2
8
16
18
0
1
1
0
1
1
4
7
10
13
0
1
0

Figure 36

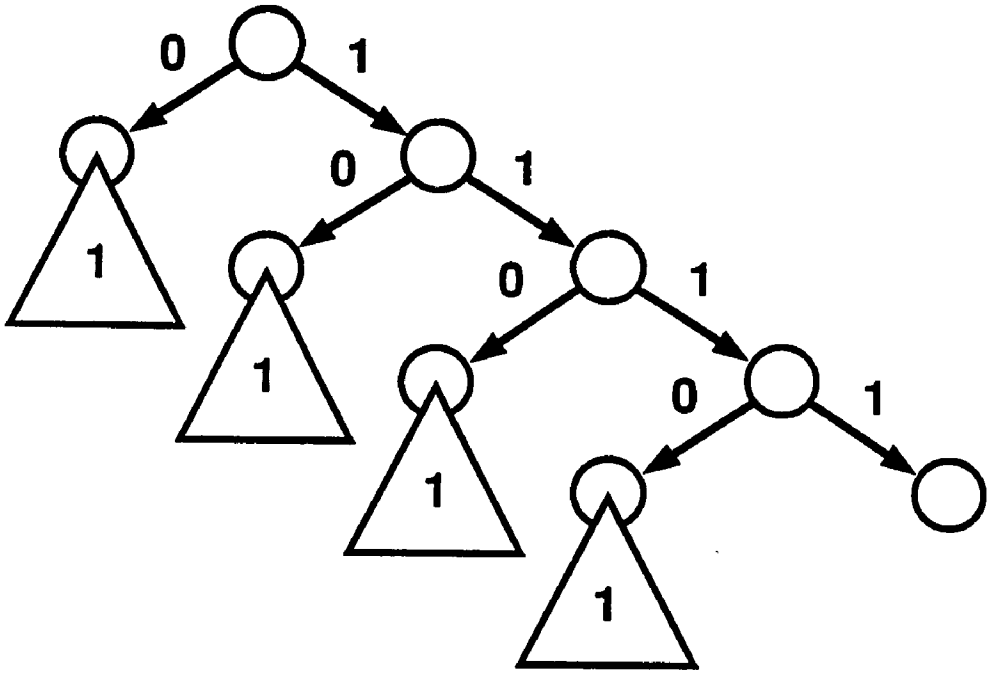


Figure 37

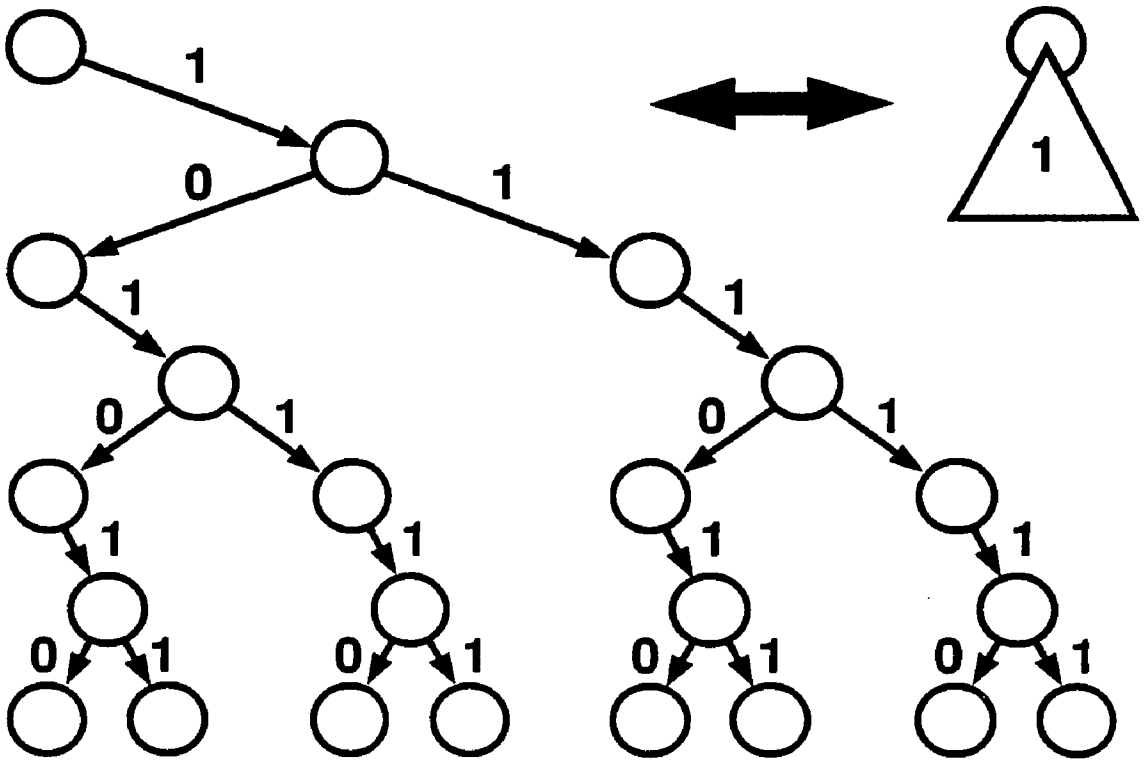


Figure 38

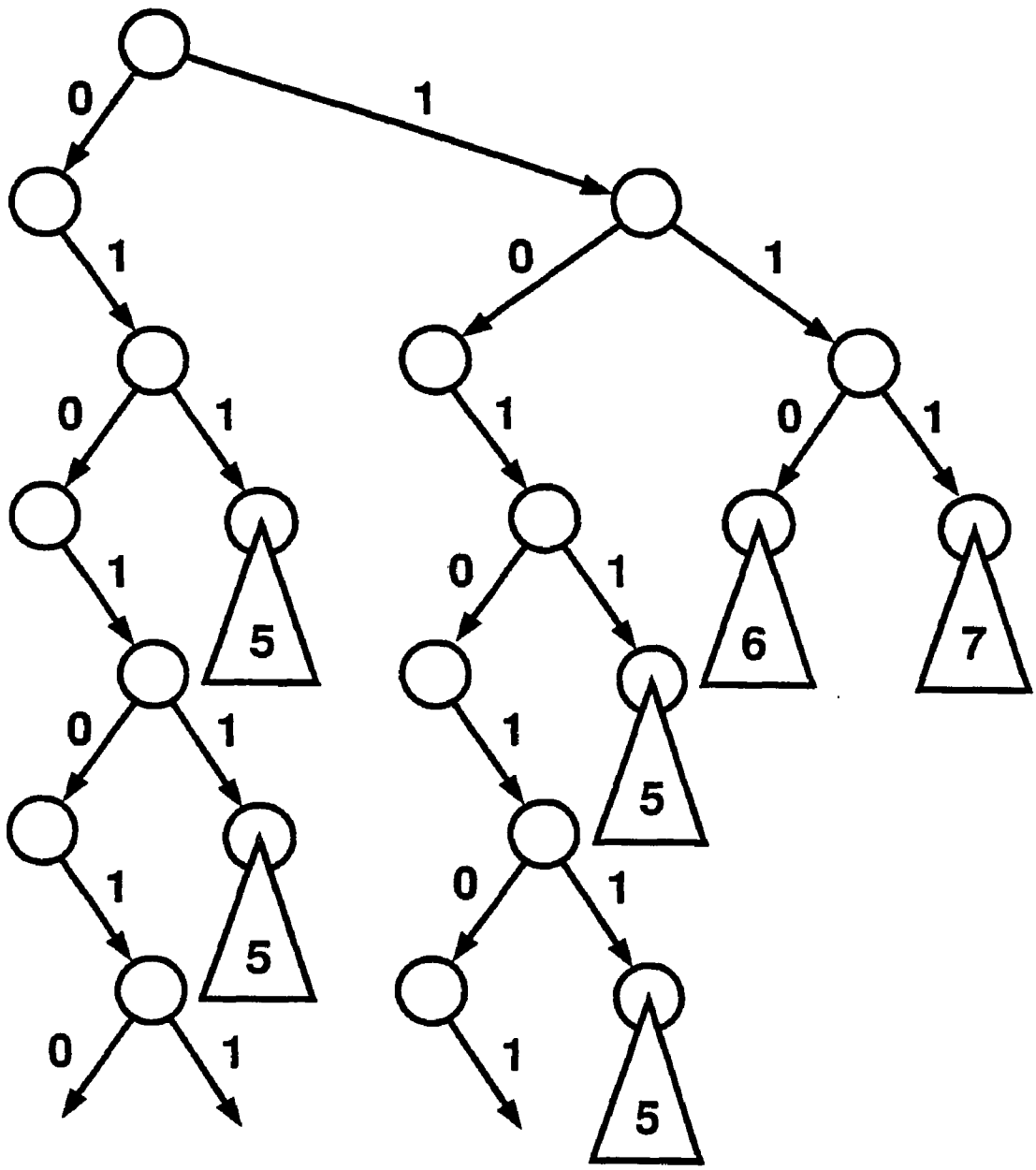


Figure 39

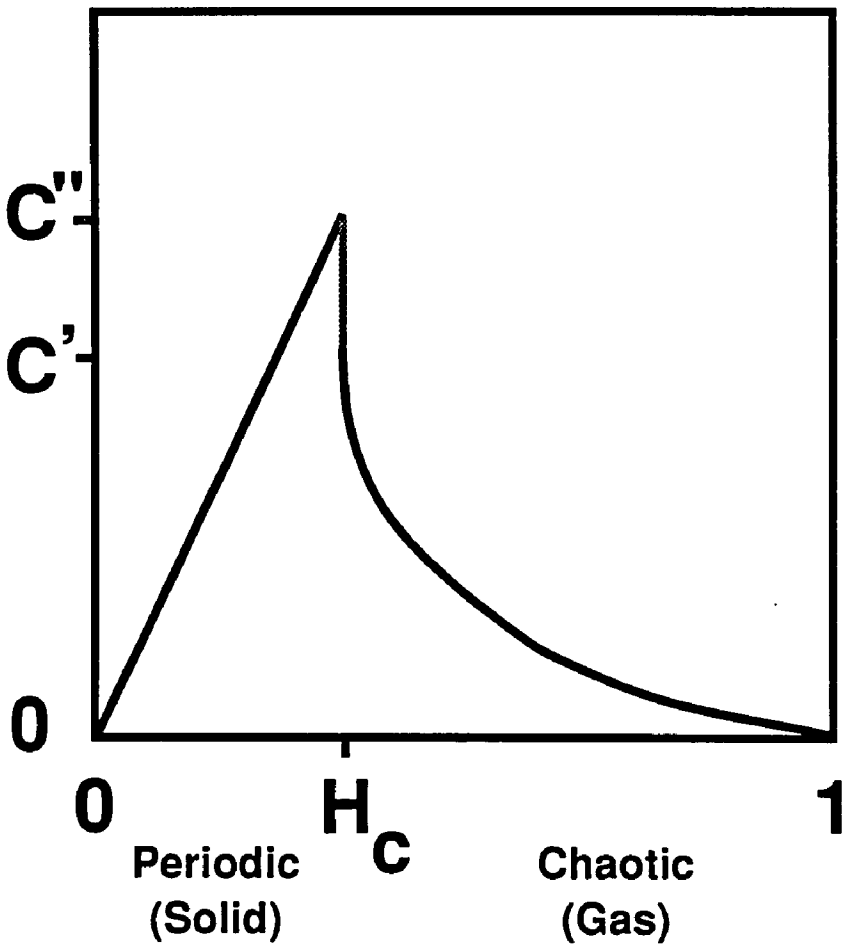


Figure 40

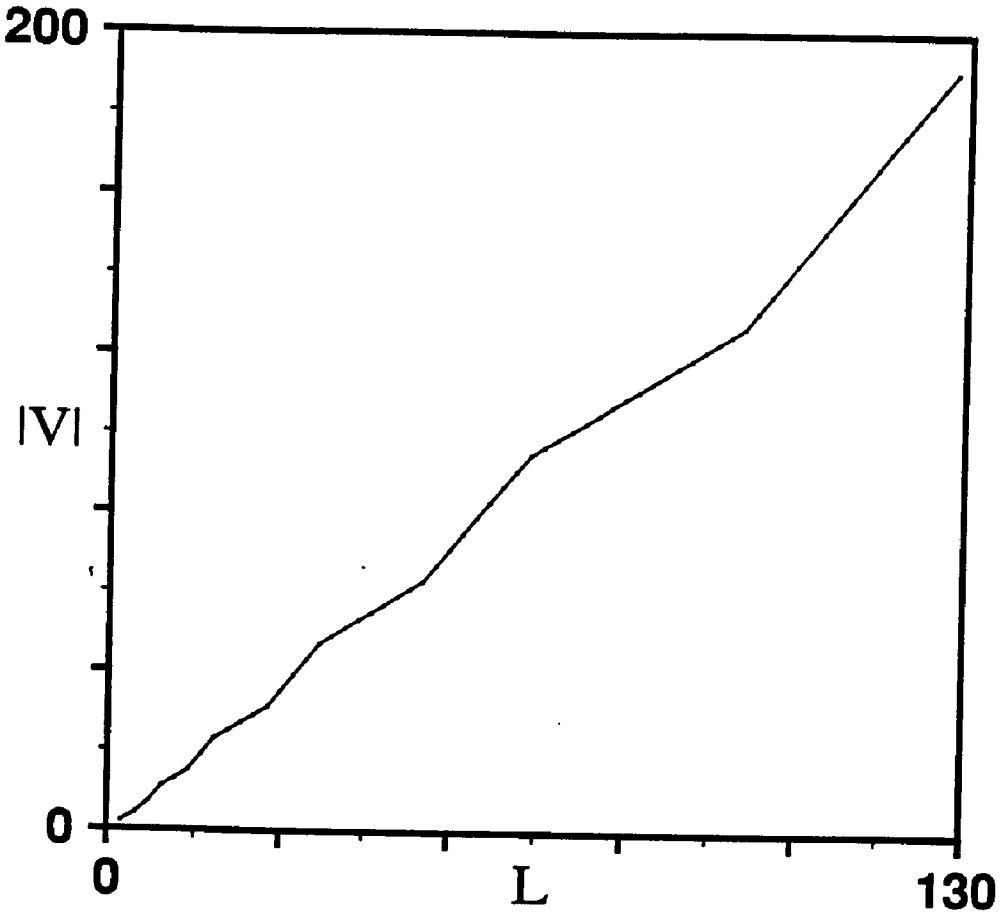


Figure 41

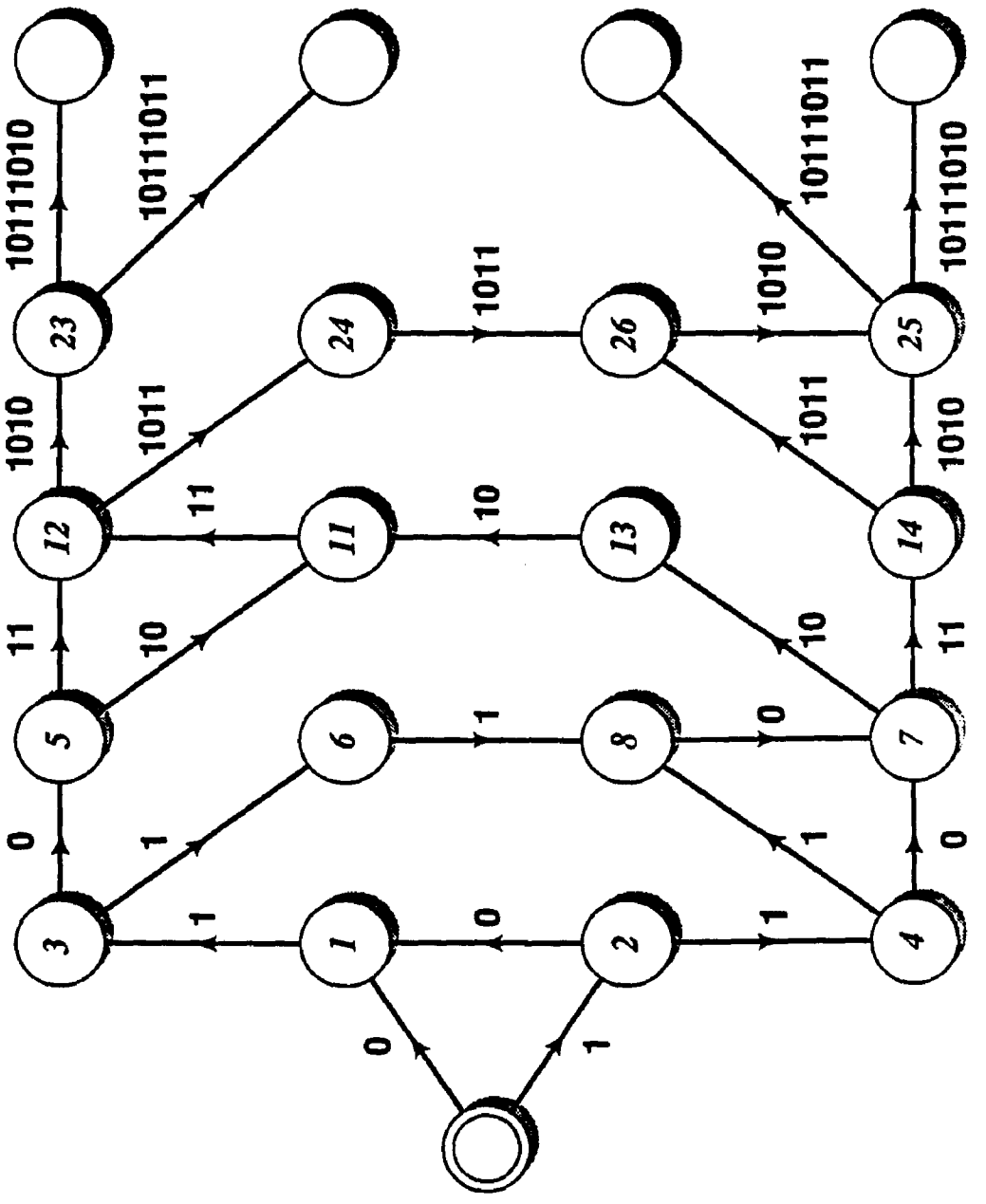

10111010
10111011
23
24
1011
26
1010
10111011
25
10111010
1010
1011
1011
1010
12
11
11
10
13
14
11
10
10
11
5
6
1
8
0
7
0
1
1
0
3
1
1
0
2
1
4
0
1
0
1

Figure 42

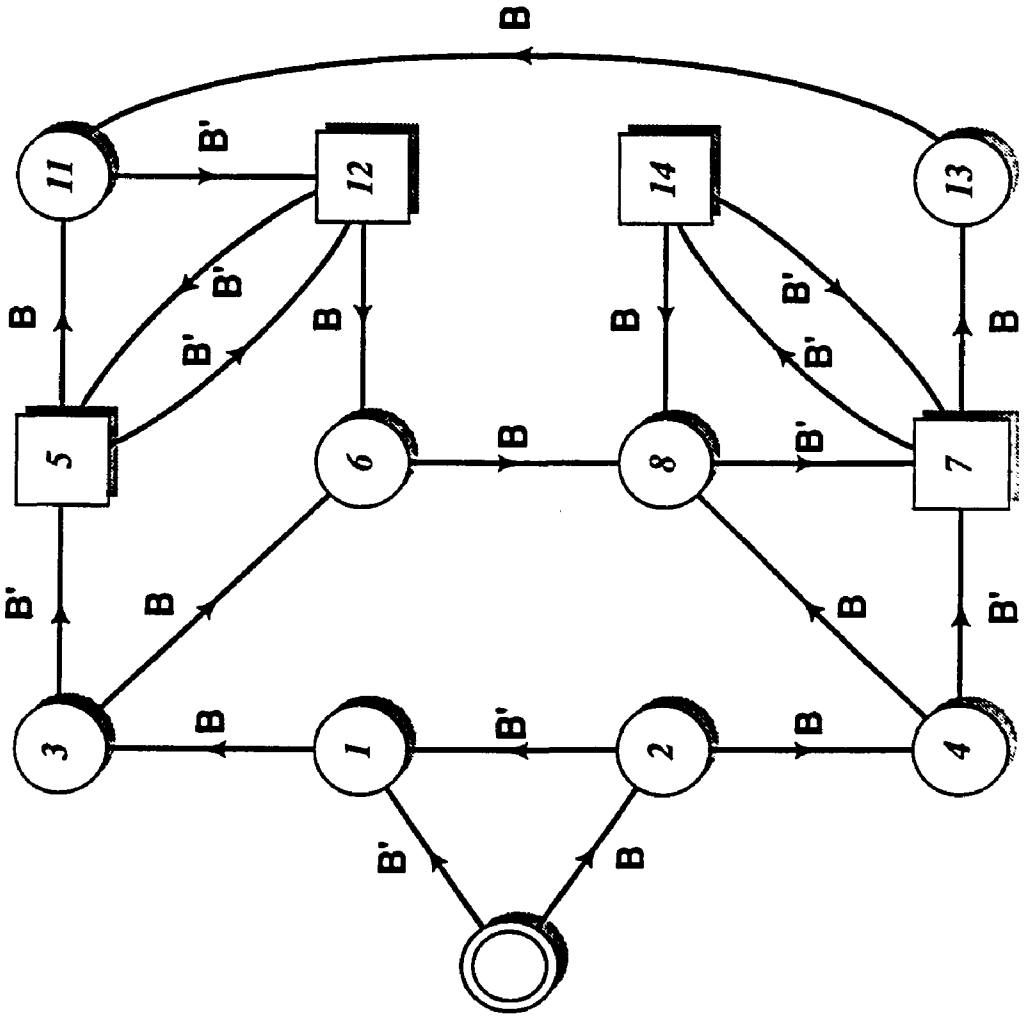


Figure 43

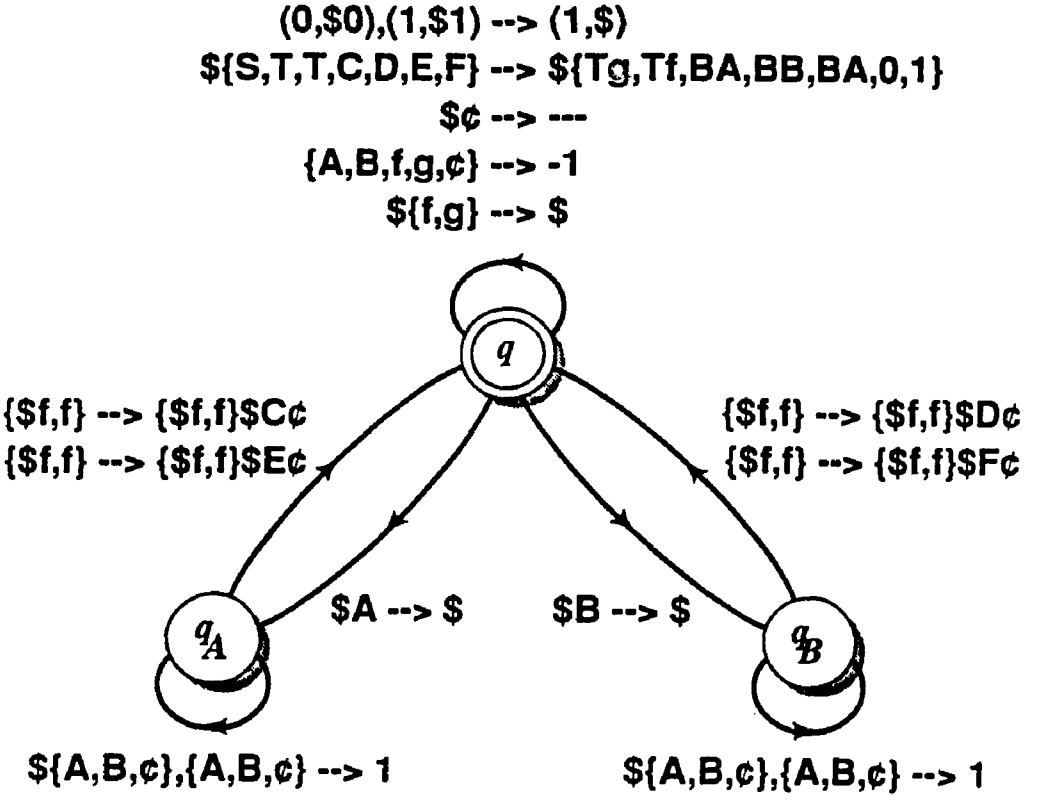

(0,$0),(1,$1) --> (1,$)
${S,T,T,C,D,E,F} --> ${Tg,Tf,BA,BB,BA,0,1}
$¢ --> ---
{A,B,f,g,¢} --> -1
${f,g} --> $
q
{$f,f} --> {$f,f}$C¢
{$f,f} --> {$f,f}$E¢
{$f,f} --> {$f,f}$D¢
{$f,f} --> {$f,f}$F¢
$A --> $
$B --> $
q_A
q_B
${A,B,¢},{A,B,¢} --> 1
${A,B,¢},{A,B,¢} --> 1

Figure 44

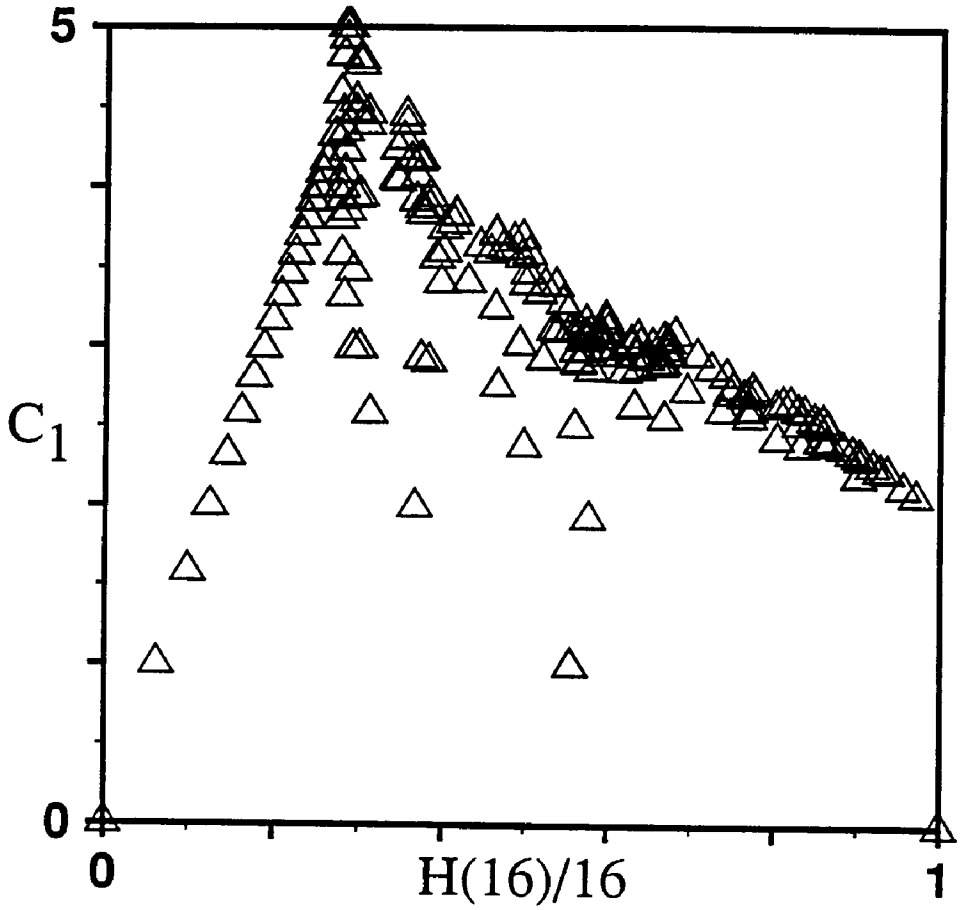


Figure 45

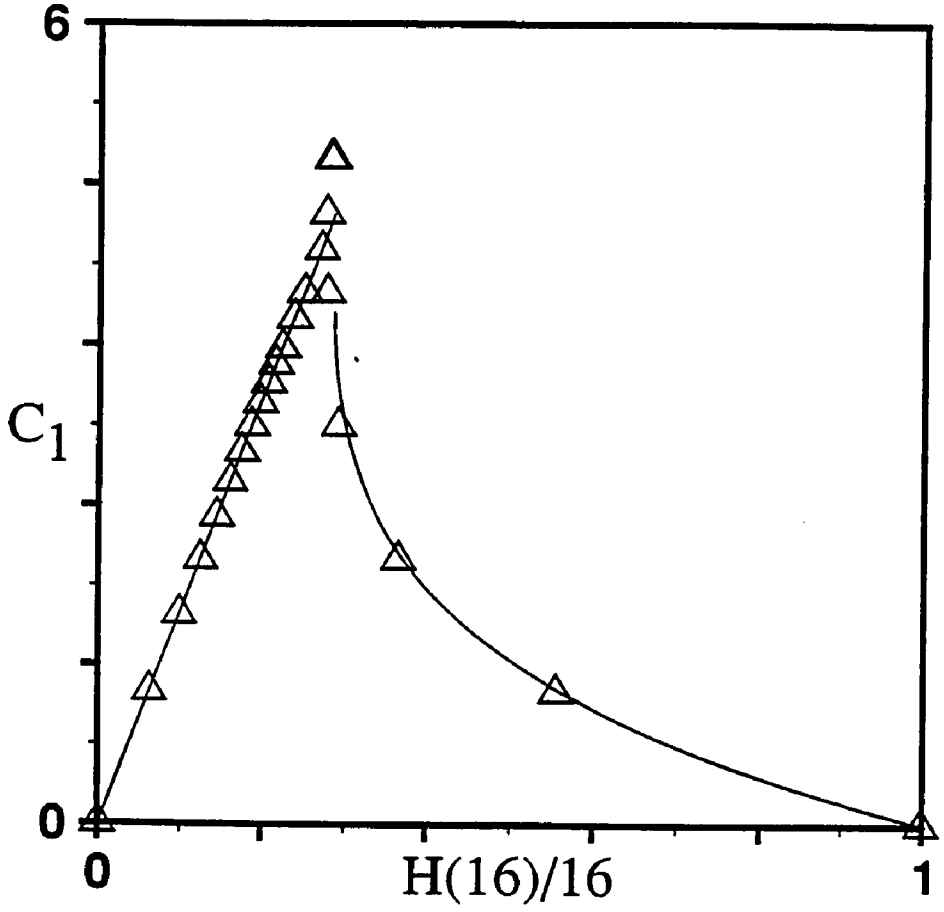


Figure 46

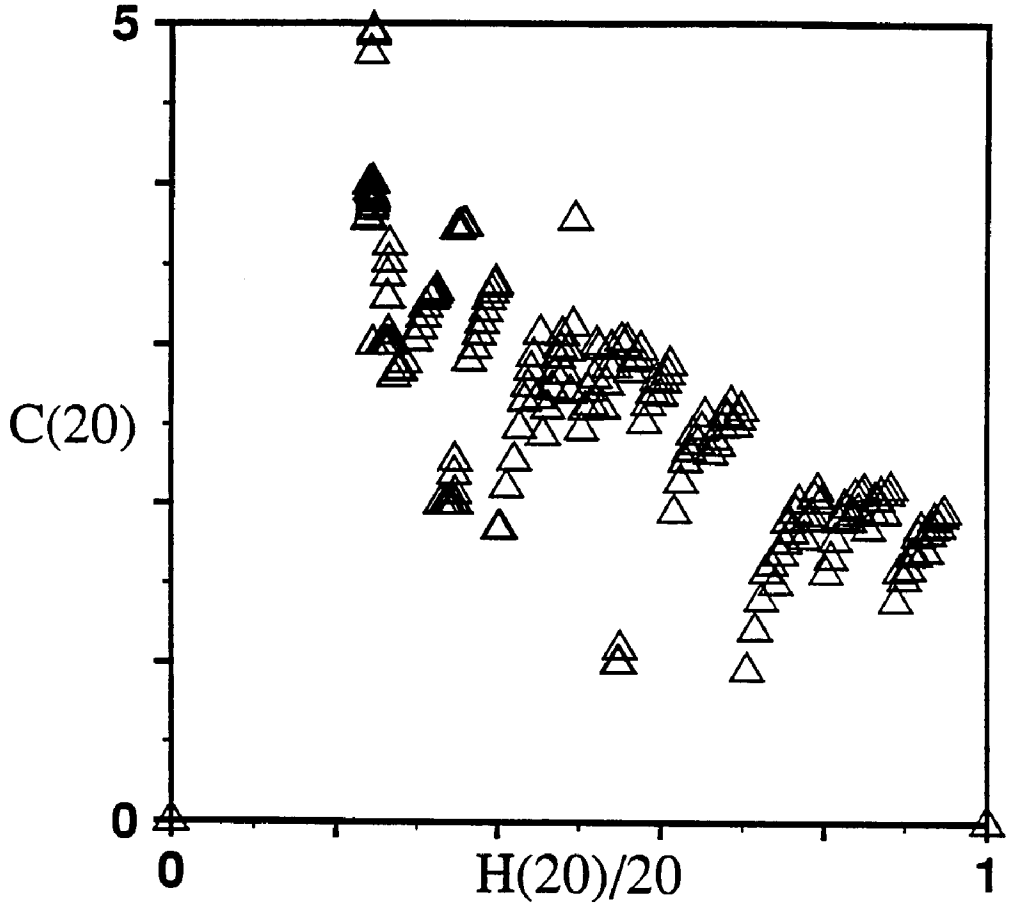


Figure 47

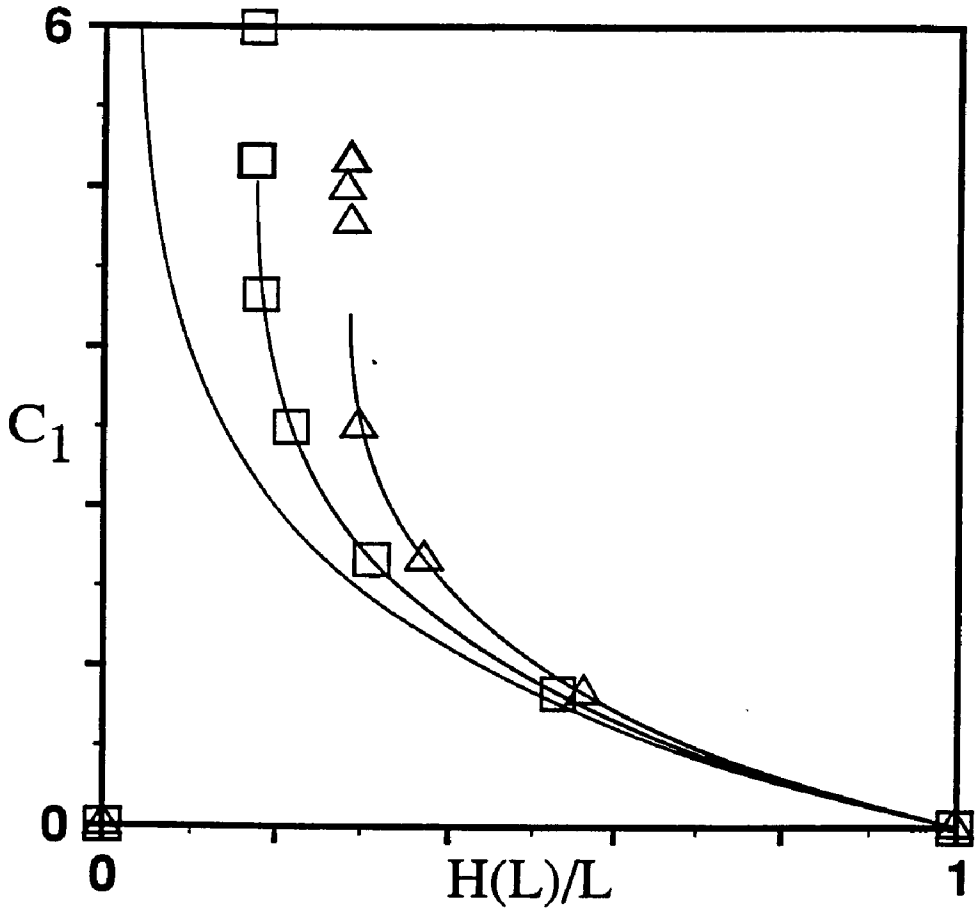


Figure 48

Bibliography

[1] A. V. Aho. Indexed grammars - an extension of context-free grammars. *J. Assoc. Comp. Mach.*, 15:647, 1968.

[2] A. V. Aho. Nested stack automata. *J. Assoc. Comp. Mach.*, 16:383, 1969.

[3] V. M. Alekseyev and M. V. Jacobson. Symbolic dynamics. *Phys. Rep.*, 25:287, 1981.

[4] C. P. Bachas and B.A. Huberman. Complexity and relaxation of hierarchical structures. *Phys. Rev. Let.*, 57:1965, 1986.

[5] C. H. Bennett. Thermodynamics of computation - a review. *Intl. J. Theo. Phys.*, 21:905, 1982.

[6] C. H. Bennett. On the nature and origin of complexity in discrete, homogeneous locally-interacting systems. *Found. Phys.*, 16:585, 1986.

[7] R. E. Blahut. *Principles and Practice of Information Theory*. Addision-Wesley, 1987.

[8] A. A. Brudno. Entropy and the complexity of the trajectories of a dynamical system. *Trans. Moscow Math. Soc.*, 44:127, 1983.

[9] G. Chaitin. On the length of programs for computing finite binary sequences. *J. ACM*, 13:145, 1966.

[10] G. Chaitin. Randomness and mathematical proof. *Sci. Am.*, May:47, 1975.

[11] N. Chomsky. Three models for the description of language. *IRE Trans. Info. Th.*, 2:113, 1956.

[12] P. Collet, J. P. Crutchfield, and J. P. Eckmann. Computing the topological entropy of maps. *Comm. Math. Phys.*, 88:257, 1983.

[13] P. Collet and J.-P. Eckmann. *Maps of the Unit Interval as Dynamical Systems*. Birkhauser, Berlin, 1980.

[14] J. P. Crutchfield. *Noisy Chaos*. PhD thesis, University of California, Santa Cruz, 1983. published by University Microfilms Intl, Minnesota.

[15] J. P. Crutchfield. Turing dynamical systems. preprint, 1987.

[16] J. P. Crutchfield. Compressing chaos. in preparation, 1989.

[17] J. P. Crutchfield. Information and its metric. In L. Lam and H. C. Morris, editors, *Nonlinear Structures in Physical Systems - Pattern Formation, Chaos and Waves*, Berlin, 1989. Springer-Verlag.

[18] J. P. Crutchfield. Inferring the dynamic, quantifying physical complexity. In N. B. Abraham, A. M. Albano, A. Passamante, and P. E. Rapp, editors, *Measures of Complexity and Chaos*, page 327. Plenum Press, 1990.

[19] J. P. Crutchfield. Reconstructing language hierarchies. In H. A. Atmanspracher, editor, *Information Dynamics*, New York, 1990. Plenum.

[20] J. P. Crutchfield, J. D. Farmer, and B. A. Huberman. Fluctuations and simple chaotic dynamics. *Phys. Rep.*, 92:45, 1982.

[21] J. P. Crutchfield, J. D. Farmer, N. H. Packard, R. S. Shaw, G. Jones, and R. Donnelly. Power spectral analysis of a dynamical system. *Phys. Lett.*, 76A:1, 1980.

[22] J. P. Crutchfield and J. E. Hanson. Chaotic arithmetic automata. *in preparation*, 1989.

[23] J. P. Crutchfield and B. A. Huberman. Fluctuations and the onset of chaos. *Phys. Lett.*, 77A:407, 1980.

[24] J. P. Crutchfield and K. Kaneko. Phenomenology of spatio-temporal chaos. In Hao Bai-lin, editor, *Directions in Chaos*, page 272. World Scientific Publishers, Singapore, 1987.

[25] J. P. Crutchfield and B. S. McNamara. Equations of motion from a data series. *Complex Systems*, 1:417, 1987.

[26] J. P. Crutchfield and N. H. Packard. Noise scaling of symbolic dynamics entropies. In H. Haken, editor, *Evolution of Order and Chaos*, page 215, Berlin, 1982. Springer-Verlag.

[27] J. P. Crutchfield and N. H. Packard. Symbolic dynamics of one-dimensional maps: Entropies, finite precision, and noise. *Intl. J. Theo. Phys.*, 21:433, 1982.

[28] J. P. Crutchfield and N. H. Packard. Symbolic dynamics of noisy chaos. *Physica*, 7D:201, 1983.

[29] J. P. Crutchfield and K. Young. Inferring statistical complexity. *Phys. Rev. Let.*, 63:105, 1989.

[30] D. M. Cvetkovic, M. Doob, and H. Sachs. *Spectra of Graphs*. Academic Press, New York, 1980.

[31] M. J. Feigenbaum. The universal metric properties of nonlinear transformations. *J. Stat. Phys.*, 21:669, 1979.

[32] M.J. Feigenbaum. Universal behavior in nonlinear systems. *Physica*, 7D:16, 1983.

[33] R. Fischer. Sofic systems and graphs. *Monastsh. Math*, 80:179, 1975.

[34] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, New York, 1979.

[35] P. Grassberger. Toward a quantitative theory of self-generated complexity. *Intl. J. Theo. Phys.*, 25:907, 1986.

[36] J. Guckenheimer. Sensitive dependence to initial conditions for one-dimensional maps. *Comm. Math. Phys.*, 70:133, 1979.

[37] J. Guckenheimer and P. Holmes. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*. Springer-Verlag, New York, 1983.

[38] J. E. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley, Reading, MA, 1979.

[39] K. Huang. *Statistical Mechanics*. J. Wiley and Sons, New York, 1963.

[40] M. V. Jacobson. Absolutely continuous invariant measures for one-parameter families of one-dimensional maps. *Comm. Math. Phys.*, 81:39, 1981.

[41] E. T. Jaynes. Where do we stand on maximum entropy? In E. T. Jaynes, editor, *Essays on Probability, Statistics, and Statistical Physics*, page 210. Reidel, London, 1983.

[42] J. G. Kemeny. The use of simplicity in induction. *Phil. Rev.*, 62:391, 1953.

[43] 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.

[44] A. N. Kolmogorov. Entropy per unit time as a metric invariant of automorphisms. *Dokl. Akad. Nauk. SSSR*, 124:754, 1959. (Russian) Math. Rev. vol. 21, no. 2035b.

[45] A. N. Kolmogorov. Three approaches to the concept of the amount of information. *Prob. Info. Trans.*, 1:1, 1965.

[46] A. N. Kolmogorov. Interpolation and extrapolation of stationary time series. In T. Kailath, editor, *Linear Least-Squares Estimation*, page 52. Dowden, Hutchinson, and Ross, New York, 1977.

[47] A. N. Kolmogorov. Stationary sequences in hilbert space. In T. Kailath, editor, *Linear Least-Squares Estimation*, page 66. Dowden, Hutchinson, and Ross, New York, 1977.

[48] A. N. Kolmogorov. Combinatorial foundations of information theory and the calculus of probabilities. *Russ. Math. Surveys*, 38:29, 1983.

[49] A. Lindenmayer and P. Prusinkiewicz. Developmental models of multicellular organisms: A computer graphics perspective. In C. G. Langton, editor, *Artificial Life*, page 221. Addison-Wesley, 1989.

[50] P. Martin-Lof. The definition of random sequences. *Info. Control*, 9:602, 1966.

[51] N. Metropolis, M. L. Stein, and P. R. Stein. On finite limit sets for transformations on the interval. *J. Combin. Th.*, 15A:25, 1973.

[52] J. Milnor and W. Thurston. On iterated maps of the interval. Princeton University preprint, 1977.

[53] M. Minsky. *Computation: Finite and Infinite Machines*. Prentice-Hall, Englewood Cliffs, New Jersey, 1967.

[54] M. Misiurewicz. Absolutely continuous measures for certain maps of an interval. Preprint IHES M-79-293, 1979.

[55] P. J. Myrberg. Iteration reellen polynome zweiten grades iii. *Ann. Akad. Sc. Fennicae A, I*, 336/3:1, 1963.

[56] D. S. Ornstein. Ergodic theory, randomness, and chaos. *Science*, 243:182, 1989.

[57] N. H. Packard. *Measurements of Chaos in the Presence of Noise*. PhD thesis, University of California, Santa Cruz, 1982.

[58] N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw. Geometry from a time series. *Phys. Rev. Let.*, 45:712, 1980.

[59] H. Pagels and S. Lloyd. Complexity as thermodynamic depth. *Ann. Phys.*, 188:186, 1988.

[60] A. Renyi. On the dimension and entropy of probability distributions. *Acta Math. Hung.*, 10:193, 1959.

[61] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:462, 1978.

[62] J. Rissanen. Universal coding, information, prediction, and estimation. *IEEE Trans. Info. Th.*, IT-30:629, 1984.

[63] C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, Champaign-Urbana, 1962.

[64] R. Shaw. Strange attractors, chaotic behavior, and information flow. *Z. Naturforsh.*, 36a:80, 1981.

[65] R. Shaw. *The Dripping Faucet as a Model Chaotic System*. Aerial Press, Santa Cruz, California, 1984.

[66] Ja. G. Sinai. On the notion of entropy of a dynamical system. *Dokl. Akad. Nauk. SSSR*, 124:768, 1959.

[67] R. J. Solomonoff. A formal theory of inductive control. *Info. Control*, 7:224, 1964.

[68] H. N. V. Temperley. *Graph Theory and Applications*. Halsted Press, 1981.

[69] H. O. A. Wold. *A Study in the Analysis of Stationary Times Series.* Almqvist and Wiksell Forlag, 1954.

[70] S. Wolfram. Computation theory of cellular automata. *Comm. Math. Phys.*, 96:15, 1984.

[71] S. Wolfram. Intractability and undecidability in theoretical physics. In S. Wolfram, editor, *Theory and Applications of Cellular Automata.* World Scientific, Singapore, 1986.

[72] S. Wolfram. *Theory and Applications of Cellular Automata.* World Scientific Publishers, Singapore, 1986.

[73] W. H. Zurek. Thermodynamic cost of computation, algorithmic complexity, and the information metric. preprint, 1989.

PART IV
Finitary ϵ-Machines: A Sofic Primer

Abstract[56]

We analyze a number of prototypical finitary ϵ-machines in order to illustrate the formal development presented in a companion exposition. The examples include periodic and chaotic machines; those that are subshifts of finite type and also those that are strictly Sofic systems. In the latter case, we present examples that have chaotic and nonchaotic transient subprocesses. Topological and stochastic versions are investigated. The analyses include estimating the entropy, complexity, and Renyi spectrum. The associated semigroup graph is constructed for each example. Finally, these are compared with both probability and conditional probability (branching) histograms of the symbol sequences each process produces.

[56] This Part is a preprint of a paper by J.P. Crutchfield and K. Young

Introduction

In [6] we introduced ϵ-machines as a way to look at how nonlinear systems process information and especially how and what they compute. Following the lead of a previous approach[12] to investigating nonlinear systems, a central requirement was that all of the notions developed must be reconstructive. That meant in particular that the computational structures had to be estimated from data streams produced by nonlinear dynamical systems. One could not use, for example, *a priori* knowledge of the equations of motion.

It is no surprise that many of the appropriate concepts, methods of analysis, and inference paradigms come neither from dynamical systems theory nor from physics. Rather they are found in theoretical computer science, especially the classic work done in the '50's and '60's on finite and infinite automata. One result of this melding of nonlinearity and computation is the conviction that theoretical computer science has much to contribute not only to dynamical systems, but also to physics, generally, and to statistical mechanics, specifically. We are not alone in this opinion and urge the reader to compare alternative approaches to "physical computation".[1, 5, 15, 8, 2, 13, 17]

Our approach has been constructive. It emphasizes, on the one hand, reconstructive techniques that apply in principle to any data series and, on the other hand, detailed analysis of distinct computational structures. The latter includes an attempt to parallel and, hopefully, extend the classification of stationary processes found in ergodic theory. As one step in this larger effort, the present work and its predecessor concern the study of different types of finitely described processes: finitary ϵ-machines. In order to bring out their computational nature and relate this to existing descriptions, we found it useful to have a set of examples analyzed from different approaches. The following presents a useful set of these. It is something of a primer for those who find the formalization in formal language theory, symbolic dynamics, and ergodic theory, somewhat less than immediately revealing.

This primer illustrates the formal techniques described in [4] by using them to analyze examples of abstract dynamical systems. The examples have been chosen to illustrate a range of deterministic behaviors; arguably one that is a very large fraction of the behavior types occurring in stationary dynamical systems. This collection of examples is a workbook; and we assume the reader is familiar with the exposition just cited. It encapsulates a number of calculations performed by hand and by machine. The following is a very brief notational review, necessary before the examples can be considered.

Finitary ϵ-Machines

A general ϵ-machine of a process is its minimal deterministic description using the least powerful computational model. Computational models with access to only a finite amount of memory have a tremendous amount of structure that can be analyzed. These are referred to as finite automata (FA) or finite state machines. FA are deterministic (DFA) when the state to state transitions are uniquely specified by the transition symbols; and nondeterministic otherwise (NFA).[9] When FA are augmented to include probabilistic structure, they are called stochastic (non)deterministic finite automata. As a shorthand we refer to stochastic deterministic finite automata as ϵ-machines.[57]

The graphical representation of a finitary ϵ-machine is an edge-labeled, directed graph $G = \{\mathbf{V}, \mathbf{E}\}$ with a finite number of vertices $\mathbf{V} = \{v_i\}$ representing optimal forecasting states for the process, introduced and discussed in [6], and a finite number of directed edges $\mathbf{E} = \{e_i\}$ connecting the vertices, each edge representing a transition between states, and labeled by a symbol $s \in \mathbf{A}$ where $\mathbf{A}$ is the alphabet of transition symbols.

Transition probabilities $p(v|v';s)$ also are assigned to the graph's edges, where v and v' represent vertices and s represents the symbol labeling the edge that connects them. The statistical structure of an ϵ-machine is given by a parametrized stochastic connection matrix

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

that is the sum over each symbol $s \in \mathbf{A}$ in the alphabet $\mathbf{A}$ of the state transition matrices

$$T_\alpha^{(s)} = \left\{ e^{\alpha \log p(v_i|v_j;s)} \right\}$$

for the vertices $v_i \in \mathbf{V}$. We will distinguish two subsets of vertices: 1) $\mathbf{V}_t$, or those associated with transient states, and 2) $\mathbf{V}_r$, or those associated with recurrent states. We will be concerned

[57] The ϵ is a reminder that the date stream, from which the machine was reconstructed, was obtained from a finite (ϵ) resolution measurement of a continuous observable.

in this paper with T_α at the particular parameter values $\alpha = 0$ and values $\alpha = 1$. T_0 is called the connection matrix of G and describes the degree of connectivity (i.e. number of edges) between vertices. T_1 is called the Markov matrix or the stochastic connection matrix and describes the probabilities of transitions between the states represented by the vertices $v_i \in \mathbf{V}$.

G and T_1 define a Markov process on the states represented by the vertices and Markov process theory gives the vertex probabilities p_v as the components of the left eigenvector π_l of the Markov matrix T_1.

We will refer to the recurrent part of a minimal DFA as a de Bruijn graph, or de Bruijn machine, and $|V_r|$ represents the number of recurrent vertices. The de Bruijn graph is used to calculate entropies and complexities.

We refer to length L subsequences $s^L = \{s_i \ldots s_j \ldots s_{i+L-1}\}$ of alphabet symbols $s \in \mathbf{A}$, from an allowable sequence for the the given ϵ-machine, as L-cylinders. The topological entropy of an ϵ-machine is a measure of the asymptotic growth rate of the number of legal L-cylinders and is given in the present context by

$$\begin{aligned} h_0 &= \lim_{L\to\infty} \frac{\log N(L)}{L} \\ &= \log_2 \lambda_{max} \end{aligned}$$

where λ_{max} is the principal eigenvalue of the connection matrix T_0.

The information in a probability distribution $p(s^L)$ defined on the L-cylinders $\{s^L\}$ for a given length L is given by

$$H(L) = -\sum_{s^L \in \{s^L\}} p(s^L) \log p(s^L)$$

The metric entropy of an ϵ-machine is a measure of the asymptotic rate of change of information in the cylinder distributions and is given in the present context by

$$\begin{aligned} h_\mu &= \lim_{L\to\infty} \frac{H(L)}{L} \\ &= \sum_{v\in V} p_v \sum_{v'\in V} p\Big(v|v';s\Big) \log p\Big(v|v';s\Big) \end{aligned}$$

where the edge probabilities $p(v|v';s)$ and vertex probabilities p_v are as defined above. The natural measures of complexity in the topological and metric cases are given by

$$C_0 = \log|\mathbf{V}|$$

and

$$C_\mu = -\sum_{v\in\mathbf{V}_r} p_v \log p_v$$

where $|\mathbf{V}|$ is the total number of vertices of the ϵ-machine. For derivations and more detailed discussions of the above quantities see [6] and [7]

Although we will not explore this in detail here, for the complete characterization of the information fluctuations in the cylinder distributions for cylinders of arbitrary length for a general system, it is necessary to consider the above quantities in the case of arbitrary α in which case we have the general Renyi entropies and complexities defined by

$$H_\alpha(L) = (1-\alpha)^{-1}\log\sum_{s^L\in\{s^L\}} p(s^L)^\alpha$$

and

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

That it is necessary to study these quantities for arbitrary α in the general case, is strictly analogous to the fact that to characterize a general probability distribution one must study an arbitrary number of moments of the distribution.

Another important representation in the case of general ϵ-machines, is in terms of the semigroups naturally associated with a given ϵ-machine. A semigroup Γ is a set $\{g_i\}$ endowed with a binary operation which is associative, i.e. preserves translation invariance of the operation. In distinction to a group, a semigroup contains an absorbing element a such that

$$ag_i = g_i a = a \quad \forall g_i \in \{g_i\}$$

and

$$\exists g_i, g_j \text{ distinct from } a \text{ such that } g_i g_j = a$$

It is the absorbing element which serves to determine illegal sequences for an ϵ-machine. A semigroup for which one of the elements is an identity element e such that

$$eg_i = g_i e = g_i \quad \forall g_i \in \{g_i\}$$

is called a monoid. A semigroup is said to be finitely generated if there is a finite subset of semigroup elements

$$\{g_j\} \subset \{g_i\}$$

such that any element of the semigroup Γ is given by a product of elements of the subset $\{g_j\}$. The semigroup naturally associated with an ϵ-machine is the matrix semigroup generated via matrix multiplication by the symbol transition matrices (or state transition matrices of order 0)

$$T_0^{(s)}$$

That the symbol transition matrices serve as generators of the semigroup makes it clear that we can represent the structure of the semigroup in terms of a semigroup graph G_Γ for which each semigroup element represents a vertex

$$v_i = g_i$$

and there is an edge between vertices v_i and v_j labeled with an s if and only if

$$g_i T_o^{(s)} = g_j$$

As in the general case for directed graphs G_Γ contains a transient set of vertices as well as a recurrent set of vertices and we shall see that the structure of the transient set is crucial for the complete classification of ϵ-machines.We note also that each element of the semigroup represents a set of legal sequences for the ϵ-machine, grouped according to unique sets of preceding and succeeding sequences.

In [6] and [7] we considered general ϵ-machine reconstruction for which the class of reconstructed machines is much larger than the class of DFA's. In this paper, however, we only consider finitary machines, i.e. DFA's. This means that the underlying statistical structure of the process is describable as a finite order Markov process in the sense that the entropies and complexities can be calculated, in the manner specified above, from the eigenvalues and eigenvectors of the finite connection and Markov matrices T_0 and T_1, characterizing the finitary machine.

The general representation of finitary processes is as a combination of periodic and random components, denoted $\mathbf{P}_t$ and $\mathbf{B}_t$, respectively. These components can be decomposed by an analysis of the eigenvalue spectrum of the stochastic connection matrix T_1 of the process. A k-periodic process will have k distinct eigenvalues on the unit circle in the complex plane. This leads to a straightforward grouping of the vertices of the DFA into k distinct sets. The graph resulting from the consideration of these sets as vertices is a k-partite graph and its connection matrix is what we refer to as $\mathbf{P}_t$. Once $\mathbf{P}_t$ has been identified a representation of $\mathbf{B}_t$ is generated by considering the k-block "return map" for the set of vertices comprising any of the k elements of $\mathbf{P}_t$. This procedure is a strict, graph theoretic, analog of detecting and filtering periodic components from an arbitrary signal.

The simplest type of finitary system is one with only a $\mathbf{P}_t$ component. This type of system is represented here by the generic period three process. Next in order of complexity are the subshifts of finite type (SSFT) which are not only statistically describable as finite order Markov processes, but for which the language of the process is describable as a topological Markov process, as defined in [14]. This property implies that the process has a finite memory in a sense made clear in [7] and [11]. The golden mean system, described below, is an example of such a system. Finitary systems whose languages are not describable as topological Markov processes are known as strictly sofic systems (SSS). The canonical example of such a system is Weiss's even system[16]. As will be seen in the examples SSS are uniquely characterized by

the existence of cycles in the transient structure of their semigroup graphs G_{1^r}.

To this end we have included as examples, SSS with various types of transient semigroup structure: 1) the even system and a system representing the dynamics of the logistic map at a Misiurewicz parameter value which exhibit transient cycles and 2) a system with a chaotic transient subprocess that generates an uncountable number of Cantor sets. Also included as an example is a system representing the dynamics of the logistic map in the band merging cascade, specifically the parameter value at which two bands merge to one.

For each example there is a brief description followed by a representation of the statistical structure of the process in terms of cylinder histograms with cylinder probabilities $p(s^L)$ for cylinders of various length L .These cylinder probabilities are estimated by their frequency of occurrence

$$p(s^L) \simeq \frac{N(s^L)}{N(L)}$$

in a simulation of the process represented by the ϵ-machine and where $N(s^L)$ is the number of occurrences of the cylinder s^L in the simulation and $N(L)$ is the total number of sequences of length L occurring in the simulation. The first histograms are just the standard cylinder histograms plotted separately for each length. Next the cylinders of all lengths less than a specific length L are represented as a tree $T = \{\mathbf{n}, \mathbf{l}\}$ consisting of nodes $\mathbf{n} = \{n_i\}$ and directed, labeled links $\mathbf{l} = \{l_i\}$ connecting the nodes in a hierarchical structure with no closed paths and the probabilities $p(s^L)$ of the cylinders at any particular level (i.e of a particular length L) are represented as different gray scale values. Lastly the branching structure of the probabilities on the tree, given by

$$p(n_j|n_i; s) = \frac{p(s^L = s_i s_{i+1} \ldots s_{i+L})}{p(s^{L-1} = s_i s_{i+1} \ldots s_{i+L-1})}$$

where

$$s_{i+L} = s$$

and n_j is the terminal node of s^L and n_i is the terminal node of s^{L-1}, is made manifest by representing each tree node with a gray scale value given by its relative branching probability from the node above it in the tree.

Following the histograms, the topological and statistical properties of the ϵ-machines are presented as

1. the connection and symbol matrices, T_0 and T_1,
2. the spectra of the connection matrices, $\mathbf{Sp}[T_\alpha]$,
3. the entropies, h_α, and
4. the complexities C_α.

Both the right and left semigroup representations are then constructed from the symbol transition matrices $\left\{T_0^{(s)}\right\}$ (e.g. in the binary cases considered here we use T_0^0 and T_0^1) which are a minimal generating set for the semigroup. We adopt the convention of labeling these two matrices as the first two elements of the semigroup Γ, i.e.

$$g_0 = T_0^0$$

and

$$g_1 = T_0^1$$

The semigroup is obtained by simply multiplying the generators in lexicographical order

$$T_0, T_1, T_0T_0, T_0T_1, T_1T_0, T_1T_1, T_0T_0T_0, \ldots$$

until no new semigroup elements occur. The right multiplication semigroup representation is the forward time semigroup representation of the process, i.e. the process obtained by parsing the data in forward time, whereas the left multiplication semigroup representation is the reverse time semigroup representation. The terms left and right refer to the direction of concatenation of symbols used in obtaining longer legal strings from shorter ones. The general structure of the semigroup consists of principal components and transient elements. The principal components

contain the recurrent group elements. In the semigroup graph the principal components are each identical copies of the recurrent portion of the minimal (forward or reverse) machine for the process, as discussed in [10]. If the principal components for the right and left representations are identical the process is time reversal symmetric. If there exist cycles or chaos (positive entropy subgraph) in the transient part of the semigroup the process is a SSS as was shown in [4]. We note here that if the semigroup contains a nontrivial identity element, which corresponds to a state of total ignorance for the process, then the process is a SSS. In this case the semigroup is a monoid. We can always augment the semigroup by adding the identity element if it does not occur in the semigroup construction for a given process. This makes the semigroup a monoid, trivially. In this case the identity will constitute what we will refer to as a trivial transient state, representing the initial state of total ignorance.

Finally, the formal language properties, given as lists of accepted and forbidden finite-length words, are presented. These are organized according to the semigroup elements.

The Period Three System

Overview

For completeness we included this simple system which has only a $\mathbf{P_t}$ component. Note that in the tree representation of the process there are asymptotically only three cylinders with uniform probabilities, corresponding to the three phases of the period three process. Although the minimal process has only three vertices, corresponding to the recurrent vertices of the reconstructedϵ-machine, the latter has transients that represent the phase locking portion of the process. Figure 49 shows the reconstructed ϵ-machine for the period three process.

The reconstructed DFA has five states, with two transient states and three recurrent states, the recurrent states corresponding to the DeBrujin graph states.

Cylinder Statistics

Figures 50,51, and 52 describe the cylinder statistics of the period three system. Note the three recurrent cylinders discussed in the overview and representing the three singular points of the distribution for the period three system.

Minimal Machine

Topological

The connection matrix and transition matrices for the topological DeBrujin machine are

$$T_0 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, T_0^0 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, T_0^1 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

The eigenvalue spectrum and left and right eigenvector corresponding to the maximal eigenvalue are

$$\mathbf{Sp}[T_0] = \left\{1, -\frac{1}{2} + \frac{\sqrt{3}}{2}i, -\frac{1}{2} - \frac{\sqrt{3}}{2}i\right\}$$

$$\pi_r = \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{pmatrix}$$

$$\pi_l = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$$

We see that the spectrum consists of three values on the unit circle implying from the above discussion that the process is, as expected, periodic with period three and since there is no branching the only component in this case is the periodic one, $\mathbf{P_t}$.

The topological entropy and complexity are

$$h_0 = 0$$

$$C_0 = \log_2 3 \simeq 1.58496$$

Thus to specify the phase of this process one needs approximately 1.6 bits of information.

Stochastic

The connection matrix, transition matrices, spectrum, entropy, and complexity for the stochastic machine are identical to those for the topological machine, which is always the case for strictly periodic processes.

Semigroup

The elements of the semigroup G_S are

$$g_0 = T^0 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad g_1 = T^1 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

$$g_2 = g_0 g_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \quad g_3 = g_1 g_0 = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$g_4 = g_1g_1 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad g_5 = g_0g_1g_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$g_6 = g_1g_0g_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \; g_7 = g_1g_1g_0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$g_8 = g_1g_0g_1g_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, g_9 = g_1g_1g_0g_1 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$g_{10} = e$$

The left representation is

$$g_0g_0 = e, \; g_0g_1 = g_2, \; g_0g_2 = e, \; g_0g_3 = e, \; g_0g_4 = g_5$$

$$g_0g_5 = e, \; g_0g_6 = e, \; g_0g_7 = g_0, \; g_0g_8 = e, \; g_0g_9 = g_2$$

$$g_1g_0 = g_3, \; g_1g_1 = g_4, \; g_1g_2 = g_6, \; g_1g_3 = g_7, \; g_1g_4 = e$$

$$g_1g_5 = g_8, \; g_1g_6 = g_9, \; g_1g_7 = e, \; g_1g_8 = g_4, \; g_1g_9 = e$$

$$L = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

The right representation is

$$g_0g_0 = e,\ g_1g_0 = g_3,\ g_2g_0 = e,\ g_3g_0 = e,\ g_4g_0 = g_7$$

$$g_5g_0 = g_0,\ g_6g_0 = e,\ g_7g_0 = e,\ g_8g_0 = g_3,\ g_9g_0 = e$$

$$g_0g_1 = g_2,\ g_1g_1 = g_4,\ g_2g_1 = g_5,\ g_3g_1 = g_6,\ g_4g_1 = e$$

$$g_5g_1 = e,\ g_6g_1 = g_8,\ g_7g_1 = g_9,\ g_8g_1 = e,\ g_9g_1 = g_4$$

$$R = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Figure 53 is the right semigroup graph for the period three system.

The semigroup contains eleven elements, including the absorbing element e, nine of which are recurrent and one of which is transient. In both the left and right representations there are three principal components, each with three elements. In this example, and for all periodic processes, the principal components for the left and right representations are identical.

Language properties

The allowed words, W_S, grouped according to semigroup elements are

$$g_0 = \{0(110)^*\}$$

$$g_1 = \{1\}$$

$$g_2 = \{01(101)^*\}$$

$$g_3 = \{10(110)^*\}$$

$$g_4 = \{11(011)^*\}$$

$$g_5 = \{(011)^*\}$$

$$g_6 = \{(101)^*\}$$

$$g_7 = \{(110)^*\}$$

$$g_8 = \{1(011)^*\}$$

$$g_9 = \{1(101)^*\}$$

The forbidden words, F_S, are

$$g_{10} = e = \{0(1(111)^* + 1(11)^{**}0)\}$$

The Golden Mean System

Overview

As mentioned in the introduction, the golden mean system is a simple example of a SSFT. It is usually considered in conjunction with the even system, to be described next, because they both have the same underlying statistical structure, i.e. the Markov process represented by their connection matrix is the same. The different edge labelings of this underlying Markov process however lead to vastly different properties for these two systems; one being a SSFT and the other a SSS. This difference is not reflected in the topological or statistical structure, though, and only becomes manifest in the semigroup structure of the processes; these systems therefore provide a good example of the importance of semigroup structure in distinguishing finitary processes with finite memory from those with infinite memory. Figure 54 shows the reconstructed ϵ-machine for the golden mean process

Cylinder Statistics

We consider cylinder statistics for the golden mean system in both cases for which we have equal branching probabilities in the stochastic machine, and those for which have unequal branching.

Figures 55,56, and 57 describe the cylinder statistics of the golden mean system with equal branching probabilities and figures 58,59, and 60 describe those with unequal branching probabilities. In both cases the figures illustrate the regular self-similar fractal support of the cylinder histogram as has been discussed in [4] but the unequal branching transition diagram gives a better view of the evolution of the cylinder distribution with increasing cylinder length.

Minimal Machine

Topological

The connection matrix and transition matrices for the topological DeBrujin machine are

$$T_0 = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, \; T_0^0 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \; T_0^1 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$$

The eigenvalue spectrum and left and right eigenvector corresponding to the maximal eigenvalue are

$$\mathbf{Sp}[T_0] = \left\{ \frac{\sqrt{5}+1}{2}, \frac{1-\sqrt{5}}{2} \right\}$$

$$\pi_r = \begin{pmatrix} \frac{\sqrt{5}-1}{2} \\ \frac{3-\sqrt{5}}{2} \end{pmatrix}$$

$$\pi_l = \begin{pmatrix} \frac{\sqrt{5}-1}{2} & \frac{3-\sqrt{5}}{2} \end{pmatrix}$$

The topological entropy and complexity are

$$h_0 = \log_2 \left(\frac{\sqrt{5}+1}{2} \right) \simeq 0.6942$$

$$C_0 = 1$$

The reconstructed DFA has 2 states, both recurrent.

There is no periodic component,$\mathbf{P_t}$, for this system hence the connection matrix itself represents the $\mathbf{B_t}$ component of the machine.

Stochastic

In the stochastic case we first consider the process for which all branching ratios are equal, and since we are dealing with binary languages, equal to one half.

The connection matrix and transition matrices for the equal branching stochastic DeBrujin machine are

$$T_1 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \end{pmatrix}, \; T_1^0 = \begin{pmatrix} 0 & \frac{1}{2} \\ 0 & 0 \end{pmatrix}, \; T_1^1 = \begin{pmatrix} \frac{1}{2} & 0 \\ 1 & 0 \end{pmatrix}$$

The eigenvalue spectrum and left and right eigenvector corresponding to the maximal eigenvalue are

$$\mathbf{Sp}[T_1] = \left\{1, \; \frac{1}{2}\right\}$$

$$\pi_r = \begin{pmatrix} \frac{2}{3} \\ \frac{1}{3} \end{pmatrix}$$

$$\pi_l = \begin{pmatrix} \frac{2}{3} & \frac{1}{3} \end{pmatrix}$$

The metric entropy and complexity are

$$h_\mu = \frac{2}{3}$$

$$C_1 = -\frac{1}{3}(2 - 3\log_2 3) \simeq 0.9183$$

We next consider the case of unequal branching.The connection matrix and transition matrices for the stochastic DeBrujin machine in this case are

$$T_1 = \begin{pmatrix} \frac{7}{10} & \frac{3}{10} \\ 1 & 0 \end{pmatrix}, \; T_1^0 = \begin{pmatrix} 0 & \frac{3}{10} \\ 0 & 0 \end{pmatrix}, \; T_1^1 = \begin{pmatrix} \frac{7}{10} & 0 \\ 1 & 0 \end{pmatrix}$$

The eigenvalue spectrum and left and right eigenvector corresponding to the maximal eigenvalue are

$$\mathbf{Sp}[T_1] = \left\{1, -\frac{3}{10}\right\}$$

$$\pi_r = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}$$

$$\pi_l \simeq (\, 0.769231 \quad 0.230769 \,)$$

The metric entropy and complexity are

$$h_\mu \simeq 0.677916$$

$$C_1 \simeq 0.779349$$

Semigroup

The elements of the semigroup G_S are

$$g_0 = T^0 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad g_1 = T^1 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$$

$$g_2 = g_0 g_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad g_3 = g_1 g_0 = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix},$$

$$g_4 = e$$

The left representation is

$$g_0 g_0 = e,\ g_0 g_1 = g_2,\ g_0 g_2 = e,\ g_0 g_3 = g_0$$

$$g_1 g_0 = g_3,\ g_1 g_1 = g_1,\ g_1 g_2 = g_1,\ g_1 g_3 = g_3$$

$$L = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}$$

The right representation is

$$g_0 g_0 = e,\ g_1 g_0 = g_3,\ g_2 g_0 = g_0,\ g_3 g_0 = e$$

$$g_0 g_1 = g_2,\ g_1 g_1 = g_1,\ g_2 g_1 = g_2,\ g_3 g_1 = g_1$$

$$R = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

Figures 61, and 62 are the right and left semigroup graphs, respectively, for the golden mean system.

The semigroup has no nontrivial transients and two disconnected (unless we add the identity) principal components. Since the principal components are identical for the right and left representations the golden mean process is time reversal symmetric.

Language properties

The allowed words, W_S, grouped according to semigroup elements are

$$g_0 = \{0, 010, 0110, 01010, 01110, ...\}$$

$$g_1 = \{1, 11, 101, 1011, 1101, 1111, ...\}$$

$$g_2 = \{01, 011, 0101, 0111, 01011, 01101, ...\}$$

$$g_3 = \{10, 110, 1010, 1110, 10110, 11010, ...\}$$

The forbidden words, F_S, are

$$g_4 = e = \{(00)^n \quad n = 1, 2, 3, ...\}$$

Thus there is essentially a single restriction, no two consecutive zeros. This generates a single Cantor set in the support of $P(s^n)$ whose dimension is h_0, as was demonstrated in the cylinder histograms.

The Even System

Overview

We next consider the canonical and perhaps simplest example of a SSS, Weiss' even system. The topological and statistical structure of the even system has already been discussed in connection with the golden mean system. Figure 63 shows the reconstructed ϵ-machine associated with the even system. We note the appearance of transient cycles in the semigroup representation of the even system. In this case the cycle is 1^*. By a theorem in [4] this implies that the even system is a SSS. We note that in this case since the identity element is an element of the semigroup, by a theorem of Coven and Paul [3] we would have been able to determine by direct examination that the even system is a SSS. From the analysis of many examples, however, the cases for which a nontrivial identity is an element of the semigroup seem to be rare.

We also demonstrate with this example, the necessity of obtaining the semigroup structure in order to determine correct cylinder statistics. If we tried to determine the cylinder statistics simply from the structure of the minimal machine we would be combining multiple machine branching probabilities to obtain single tree branching probabilities and this would lead to inconsistencies in cylinder probability assignments. To illustrate this we show the semigroup graph for the even system, the structure of which will be discussed below, with branching probabilities in figure 64, and the appropriate probabilities on the tree to level four in figure 65. Note that distinct semigroup elements are combined in the minimal or reconstructed machines and this "loss of information" leads to the fact that one needs the full semigroup structure to obtain consistent cylinder statistics for the system. We note that this is the result of having nontrivial transient semigroup elements and hence is not important for finite order Markov processes; this may be the reason that, to the best of our knowledge, semigroup structure has not been emphasized in the statistics literature.

Cylinder Statistics

As we did for the golden mean system we consider cylinder statistics for equal and unequal branching.

Figures 66,67, and 68 describe the cylinder statistics of the even system with equal branching probabilities and figures 69,70, and 71 describe those with unequal branching probabilities. The histograms are clearly more complex than those for the golden mean system. In fact new Cantor sets in the support are generated at successively longer cylinder lengths. The reason for this can be seen most clearly in terms of the formal language properties of the even system and will be described below.

Minimal Machine

Topological

All of the topological properties for the even system are identical to those of the golden mean system.

Stochastic

As for the topological properties, the stochastic properties of the even system are identical to those of the golden mean system, and we again stress that this emphasizes the fact that determining the semigroup structure of a system is necessary for determining whether or not it is a SSFT or a SSS.

Semigroup

The elements of the semigroup G_S are

$$g_0 = T^0 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad g_1 = T^1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

$$g_2 = g_0 g_1 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad g_3 = g_1 g_0 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix},$$

$$g_4 = g_1 g_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad g_5 = g_1 g_0 g_1 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

$$g_6 = e$$

The left representation is

$$g_0 g_0 = g_0, \; g_0 g_1 = g_2, \; g_0 g_2 = g_2$$

$$g_0 g_3 = e, \; g_0 g_4 = g_0, \; g_0 g_5 = e$$

$$g_1 g_0 = g_3, \; g_1 g_1 = g_4, \; g_1 g_2 = g_5$$

$$g_1 g_3 = g_0, \; g_1 g_4 = g_1, \; g_1 g_5 = g_2$$

$$L = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

The right representation is

$$g_0 g_0 = g_0, \; g_1 g_0 = g_3, \; g_2 g_0 = e$$

$$g_3 g_0 = g_3, \; g_4 g_0 = g_0, \; g_5 g_0 = e$$

$$g_0 g_1 = g_2,\ g_1 g_1 = g_4,\ g_2 g_1 = g_0$$

$$g_3 g_1 = g_5,\ g_4 g_1 = g_1,\ g_5 g_1 = g_3$$

$$R = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Figures 72, and 73 are the right and left semigroup graphs, respectively, for the even system.

The semigroup has two transient elements and two principal components each containing two elements. There is a two cycle $(11)^*$ in the transient portion of the semigroup indicating that the even system is a SSS. Since the principal components are identical for the right and left representations, the even system, like the golden mean system, is time reversal symmetric.

Language properties

The allowed words, W_S, grouped according to semigroup elements are

$$g_0 = \{0, 00, 000, 011, 110, ...\}$$

$$g_1 = \{1, 111, 11111, ...\}$$

$$g_2 = \{01, 001, 0001, 0111, ...\}$$

$$g_3 = \{10, 100, 1110, 1011, ...\}$$

$$g_4 = \{11, 1111, 111111, ...\}$$

$$g_5 = \{101, 11101, 10111, ...\}$$

The forbidden words, F_S, are

$$g_6 = e = \left\{01^{2n+1}0\right\} \quad n = 1, 2, 3, ...$$

i.e. strings with an odd number of successive 1's are forbidden.

In this case, as opposed to that of the golden mean, we get new restrictions at each cylinder length. Thus when looking at $P(s^n)$ for successively longer L, new forbidden words are discovered. Each generates its own Cantor set in the support of $P(s^n)$ and in the limit, $L \to \infty$, an infinite number of Cantor sets are generated.

Band Merging 2–>1

Overview

As an illustration of the structure of DFAs generated by the Misiurewicz parameter values for the band merging cascade in the logistic map [7] we discuss the case of two bands merging to one. The minimal DFA for this case has multiple edges between two states so we apply the edge graph operator as discussed in [4] to obtain a graph without this. Figure 74 shows the corresponding reconstructed ϵ-machine.

This is a simple example of a process with both $\mathbf{P}_t$, and $\mathbf{B}_t$ components. We see that the spectrum of the stochastic machine contains the two values (1,-1) on the unit circle indicating, as stated in the introduction, that the process has a period two component with an associated 2-partite graph. When the periodic part of the process is "filtered" the $\mathbf{B}_t$ component that remains is a simple 2-block Bernoulli process.

As viewed by other authors [3] this process would be considered a SSS but we will regard this as a trivial sort of soficity (in this sense any process with a periodic component is sofic.) We restrict our attention to soficity associated with cyclic or chaotic behavior in the transient semigroup elements.

Cylinder Statistics

Figures 75,76 ,and 77 describe the cylinder statistics of the band merging $2 \rightarrow 1$ system with equal branching probabilities and figures 78,79, and 80 describe those with unequal branching probabilities.

Minimal (Single-Edged) Machine

Topological

The connection matrix and transition matrices for the recurrent part of the topological machine are

$$T_0 = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \; T_0^0 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \; T_0^1 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

We note here that the connection matrices for the minimal DeBrujin graph are

$$T_0 = \begin{pmatrix} 0 & 2 \\ 1 & 0 \end{pmatrix}, \; T_0^0 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \; T_0^1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

The eigenvalue spectrum and left and right eigenvector corresponding to the maximal eigenvalue are

$$\mathbf{Sp}[T_0] = \left\{\sqrt{2}, 0, -\sqrt{2}\right\}$$

$$\pi_r = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{4} \\ \frac{1}{4} \end{pmatrix}$$

$$\pi_l = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$$

The topological entropy and complexity are

$$h_0 = \frac{1}{2}$$

$$C_0 = 1$$

We note that there are two recurrent states and one transient state in the reconstructed minimal DFA but there are three recurrent states in the single edge graph obtained from the minimal graph via the edge operator.

$\mathbf{P_t}$ corresponding to the 2-partite graph is

$$\mathbf{P_t} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

and the corresponding 2-block $\mathbf{B_t}$ is

$$\mathbf{B_t} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

Stochastic

The connection matrix and transition matrices for the stochastic machine are

$$T_1 = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \; T_1^0 = \begin{pmatrix} 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \; T_1^1 = \begin{pmatrix} 0 & 0 & \frac{1}{2} \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

The connection matrices for the equal branching stochastic minimal DeBrujin graph are

$$T_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \; T_1^0 = \begin{pmatrix} 0 & \frac{1}{2} \\ 0 & 0 \end{pmatrix}, \; T_1^1 = \begin{pmatrix} 0 & \frac{1}{2} \\ 1 & 0 \end{pmatrix}$$

The eigenvalue spectrum and left and right eigenvector corresponding to the maximal eigenvalue are

$$\mathbf{Sp}[T_1] = \{1, 0, -1\}$$

$$\pi_r = \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{pmatrix}$$

$$\pi_l = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$$

The averaged vertex probabilities for the stochastic minimal DeBrujin machine are

$$p_1 = \frac{1}{2}, \quad p_2 = \frac{1}{2}$$

The metric entropy and complexity are

$$h_\mu = \frac{1}{2}$$

$$C_1 = 1$$

We note here that for systems such as this, whose minimal machine has multiple edges, we must be careful to compute the complexities and metric entropy using the minimal DeBrujin machine despite it's having multiple edges. The spectral radius and topological entropy of the process are invariant under the edge graph operator which operates on the minimal/, multiple edged graph to produce a single edged graph, necessary for the semigroup construction. The complexities and metric entropy however, are not invariant under this operation and so must be computed using the minimal DeBrujin machine. In fact the very definition of complexity is dependent on the claim that one has produced a minimal representation of the given process.

Semigroup

The elements of the semigroup G_S are

$$g_0 = T^0 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad g_1 = T^1 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

$$g_2 = g_0 g_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad g_3 = g_1 g_0 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

$$g_4 = g_1 g_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \quad g_5 = g_0 g_1 g_1 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$g_6 = g_1 g_0 g_1 = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad g_7 = g_1 g_0 g_1 g_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

$$g_8 = e$$

The left representation is

$$g_0g_0 = e,\ g_0g_1 = g_2,\ g_0g_2 = e,\ g_0g_3 = g_0,\ g_0g_4 = g_5$$

$$g_0g_5 = e,\ g_0g_6 = g_2,\ g_0g_7 = g_5$$

$$g_1g_0 = g_3,\ g_1g_1 = g_4,\ g_1g_2 = g_6,\ g_1g_3 = g_0,\ g_1g_4 = g_1$$

$$g_1g_5 = g_7,\ g_1g_6 = g_2,\ g_1g_7 = g_5$$

$$L = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 \end{pmatrix}$$

The right representation is

$$g_0g_0 = e,\ g_1g_0 = g_3,\ g_2g_0 = g_0,\ g_3g_0 = e,\ g_4g_0 = g_0$$

$$g_5g_0 = e,\ g_6g_0 = g_3,\ g_7g_0 = e$$

$$g_0g_1 = g_2,\ g_1g_1 = g_4,\ g_2g_1 = g_5,\ g_3g_1 = g_6,\ g_4g_1 = g_1$$

$$g_5g_1 = g_2,\ g_6g_1 = g_7,\ g_7g_1 = g_6$$

$$R = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Figure 81 shows the right semigroup graph for the band merging system. Note that there is a transient cycle in the right semigroup graph indicating that this system is a SSS.

Language properties

The allowed words, W_S, grouped according to semigroup elements are

$$g_0 = \{0, 010, 110, ...\}$$

$$g_1 = \{1, 111, ...\}$$

$$g_2 = \{01, 0101, 0111, ...\}$$

$$g_3 = \{10, 1010, 1110, ...\}$$

$$g_4 = \{11, 1111, ...\}$$

$$g_5 = \{011, 01111, 10101, 11011, ...\}$$

$$g_6 = \{101, 10111, 11101, ...\}$$

$$g_7 = \{1101, ...\}$$

The forbidden words, F_S, are

$$g_8 = e = \{0(11)^n \quad n = 0, 1, 2, 3, ...0\}$$

Misiurewicz Machine ($r \approx 3.928$)

Overview

This example is of the DFA produced at a "typical" Misiurewicz parameter value in the logistic map, in this case a parameter value at which four bands merge to five. We see that in this case the reconstructed machine has no transient elements and is identical to the minimal machine. To establish that this machine is in fact a SSS it is necessary to construct the semigroup. We observe that there are transient cycles, similar to those that occur for the even system, in the semigroup graph and hence that the machine is a SSS. Figure 82 shows the reconstructed ϵ-machine.

It is interesting to note that no transient vertices occur in the reconstructed ϵ-machine in this case illustrating the fact that while transients in the reconstructed ϵ-machine imply that a system is a SSS the converse is not true. This is another example of the need to appeal to semigroup structure in determining whether or not a system is a SSS.

Cylinder Statistics

As for the golden mean and even systems we consider both equal and unequal branching.

Figures 83,84, and 85 describe the cylinder statistics of the Misiurewicz system in the case of equal branching probabilities and figures 86,87, and 88 describe the case of unequal branching probabilities.

As expected, both the support and distribution are more complex than that for the even system.

Minimal Machine

Topological

The connection matrix and transition matrices for the topological DeBrujin machine are

$$T_0 = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}, \quad T_0^0 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \quad T_0^1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

The eigenvalue spectrum and left and right eigenvector corresponding to the maximal eigenvalue are

$$\mathbf{Sp}[T_0] \simeq \{1.76929, -0.88465 \pm 0.589743i, 1\}$$

$$\pi_r \simeq \begin{pmatrix} 0.361103 \\ 0.277794 \\ 0.130396 \\ 0.230707 \end{pmatrix}$$

$$\pi_l \simeq (\,0.361103 \quad 0.277794 \quad 0.230707 \quad 0.130396\,)$$

The topological entropy and complexity are

$$h_0 \simeq 0.8232$$

$$C_0 = 2$$

The reconstructed DFA has 4 states, all of which are recurrent and it has no $\mathbf{P_t}$ component.

Stochastic

The connection matrix and transition matrices for the stochastic DeBrujin machine in the case of equal branching are

$$T_1 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}, \quad T_1^0 = \begin{pmatrix} 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \end{pmatrix}, \quad T_1^1 = \begin{pmatrix} \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & \frac{1}{2} & 0 \end{pmatrix}$$

The eigenvalue spectrum and left and right eigenvector corresponding to the maximal eigenvalue are

$$\mathbf{Sp}[T_1] \simeq \{1, -0.551392 \pm 0.332728i, 0.602785\}$$

$$\pi_r = \begin{pmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{pmatrix}$$

$$\pi_l = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$$

The metric entropy and complexity are

$$h_\mu = \frac{3}{4}$$

$$C_1 = 2$$

The connection matrix and transition matrices for the stochastic DeBrujin machine in the case of unequal branching are

$$T_1 = \begin{pmatrix} \frac{3}{10} & \frac{7}{10} & 0 & 0 \\ \frac{3}{10} & 0 & \frac{7}{10} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & \frac{7}{10} & \frac{3}{10} & 0 \end{pmatrix}, \quad T_1^0 = \begin{pmatrix} 0 & \frac{7}{10} & 0 & 0 \\ 0 & 0 & \frac{7}{10} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & \frac{7}{10} & 0 & 0 \end{pmatrix}, \quad T_1^1 = \begin{pmatrix} \frac{3}{10} & 0 & 0 & 0 \\ \frac{3}{10} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & \frac{3}{10} & 0 \end{pmatrix}$$

The eigenvalue spectrum and left and right eigenvector corresponding to the maximal eigenvalue are

$$\mathbf{Sp}[T_1] \simeq \{1, -0.531938 \pm 0.542366i, 0.363876\}$$

$$\pi_r = \begin{pmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{pmatrix}$$

$$\pi_l = \begin{pmatrix} \frac{1}{8} & \frac{7}{24} & \frac{7}{24} & \frac{7}{24} \end{pmatrix}$$

The metric entropy and complexity are

$$h_\mu \simeq 0.624248$$

$$C_1 \simeq 1.93041$$

Semigroup

The elements of the semigroup G_S are

$$g_0 = T^0 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \quad g_1 = T^1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

$$g_2 = g_0 g_0 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad g_3 = g_0 g_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

$$g_4 = g_1g_0 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad g_5 = g_1g_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$g_6 = g_0g_0g_1 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad g_7 = g_0g_1g_0 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$$g_8 = g_0g_1g_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}, \quad g_9 = g_1g_0g_0 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$g_{10} = g_1g_0g_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad g_{11} = g_0g_0g_1g_0 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix},$$

$$g_{12} = g_0g_1g_0g_0 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad g_{13} = g_0g_1g_0g_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

$$g_{14} = g_1g_0g_0g_1 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad g_{15} = g_0g_0g_1g_0g_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

$$g_{16} = g_0g_1g_0g_0g_1 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, g_{17} = e$$

The left representation is

$$g_0g_0 = g_2,\ g_0g_1 = g_3,\ g_0g_2 = e,\ g_0g_3 = g_6$$

$$g_0g_4 = g_7,\ g_0g_5 = g_8,\ g_0g_6 = e,\ g_0g_7 = g_{11}$$

$$g_0g_8 = g_2,\ g_0g_9 = g_{12},\ g_0g_{10} = g_{13},\ g_0g_{11} = e$$

$$g_0g_{12} = g_2,\ g_0g_{13} = g_{15},\ g_0g_{14} = g_{16}$$

$$g_0g_{15} = e,\ g_0g_{16} = g_6$$

$$g_1g_0 = g_4,\ g_1g_1 = g_5,\ g_1g_2 = g_9,\ g_1g_3 = g_{10}$$

$$g_1g_4 = g_7,\ g_1g_5 = g_1,\ g_1g_6 = g_{14},\ g_1g_7 = g_4$$

$$g_1g_8 = g_{10},\ g_1g_9 = g_{12},\ g_1g_{10} = g_{13},\ g_1g_{11} = g_4$$

$$g_1g_{12} = g_9,\ g_1g_{13} = g_{10},\ g_1g_{14} = g_{16}$$

$$g_1g_{15} = g_{10},\ g_1g_{16} = g_{14}$$

$$L = \begin{array}{c} \begin{matrix} 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix} \\ \left(\begin{array}{ccccccccccccccccc} 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{array} \right) \end{array}$$

The right representation is

$$g_0g_0 = g_2,\ g_1g_0 = g_4,\ g_2g_0 = e,\ g_3g_0 = g_7$$

$$g_4g_0 = g_9,\ g_5g_0 = g_7,\ g_6g_0 = g_{11},\ g_7g_0 = g_{12},$$

$$g_8g_0 = g_{11},\ g_9g_0 = e,\ g_{10}g_0 = g_4,\ g_{11}g_0 = g_2$$

$$g_{12}g_0 = e,\ g_{13}g_0 = g_7,\ g_{14}g_0 = g_4$$

$$g_{15}g_0 = g_{11},\ g_{16}g_0 = g_7$$

$$g_0g_1 = g_3,\ g_1g_1 = g_5,\ g_2g_1 = g_6,\ g_3g_1 = g_8$$

$$g_4g_1 = g_{10},\ g_5g_1 = g_1,\ g_6g_1 = g_2,\ g_7g_1 = g_{13}$$

$$g_8g_1 = g_3,\ g_9g_1 = g_{14},\ g_{10}g_1 = g_{10},\ g_{11}g_1 = g_{15}$$

$$g_{12}g_1 = g_{16},\ g_{13}g_1 = g_{13},\ g_{14}g_1 = g_9$$

$$g_{15}g_1 = g_{15},\ g_{16}g_1 = g_{12}$$

$$R = \begin{array}{c} \begin{array}{cccccccccccccccc} 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \\ \left(\begin{array}{cccccccccccccccc} 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{array} \right) \end{array}$$

Figure 89 shows the right semigroup graph for the Misiurewicz system. The existence of transient cycles shows that the Misiurewicz system is a SSS.

Language properties

The allowed words, W_S, grouped according to semigroup elements are

$$g_0 = \{0\}, g_1 = \{1, 111, 11111, ...\}$$

$$g_2 = \{00, 0011, 00100, ...\}, g_3 = \{01, 0111, ...\}$$

$$g_4 = \{10, 1010, 1110, ...\}, g_5 = \{11, 1111, 111111, ...\}$$

$$g_6 = \{001, 00111, ...\}, g_7 = \{010, 110, 01010, ...\}$$

$$g_8 = \{011, 01111, ...\}, g_9 = \{100, 10011, 10100, ...\}$$

$$g_{10} = \{10, 1011, 10101, ...\}, g_{11} = \{0010, 0110, ...\}$$

$$g_{12} = \{0100, 1100, ...\}, g_{13} = \{0101, 1101, 01011, ...\}$$

$$g_{14} = \{1001, 101001, ...\}, g_{15} = \{00101, 01101, ...\}$$

$$g_{16} = \{01001, 11001, ...\}$$

The forbidden words, F_S, are

$$g_{17} = e = \{000, 0000, 0001, 1000, ...\}$$

Transient Chaos Machine

Section 1 Overview

This example was constructed to demonstrate the behavior of a system with chaotic transient elements. As one might expect the cylinder statistics and semigroup structure are much more complex than in the cases examined earlier. This is due in large part to the growth in the number of transient cylinders. Figure 90 shows the reconstructed ϵ-machine.

Cylinder Statistics

Again we will consider both the equal and unequal branching cases.

From the histograms we see that as for the Misiurewicz system both $P(s^n)$ and its support are much more complex than for the golden mean or even systems.

Figures 91,92, and 93 describe the cylinder statistics of the transient chaos system in the case of equal branching probabilities and figures 94,95, and 96 describe the case of unequal branching probabilities.

Minimal (Single-Edged) Machine

Topological

The connection matrix and transition matrices for the recurrent part of the topological

machine are

$$T_0 = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$T_0^0 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad T_0^1 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

We note here that the connection matrices for the minimal DeBrujin graph are

$$T_0 = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \\ 2 & 0 & 0 & 0 \end{pmatrix}, \quad T_0^0 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}, \quad T_0^1 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

The eigenvalue spectrum and left and right eigenvector corresponding to the maximal eigenvalue are

$$\mathbf{Sp}[T_0] \simeq \{1.6041, 0.6080 \pm 1.1903i, -0.9101 \pm 0.7534i, 0, 0\}$$

$$\pi_r \simeq \begin{pmatrix} 0.217522 \\ 0.131408 \\ 0.105397 \\ 0.105397 \\ 0.169069 \\ 0.135603 \\ 0.135603 \end{pmatrix}$$

$$\pi_l \simeq (\,0.28659 \quad 0.17866 \quad 0.11138 \quad 0.11138 \quad 0.13886 \quad 0.08657 \quad 0.08657\,)$$

The topological entropy and complexity are

$$h_0 \simeq 0.6818$$

$$C_0 \simeq 2.8074$$

The minimal, single edged, DFA has seven recurrent states and has only a $\mathbf{B}_t$ component

Stochastic

The connection matrix and transition matrices for the stochastic machine are

$$T_0 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$T_0^0 = \begin{pmatrix} \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad T_0^1 = \begin{pmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

The connection matrices for the equal branching minimal stochastic DeBrujin graph are

$$T_0 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}, \quad T_0^0 = \begin{pmatrix} \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \end{pmatrix}, \quad T_0^1 = \begin{pmatrix} 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \\ \frac{1}{2} & 0 & 0 & 0 \end{pmatrix}$$

The eigenvalue spectrum and left and right eigenvector corresponding to the maximal eigenvalue for the equal branching minimal stochastic DeBrujin graph are

$$\mathbf{Sp}[T_1] \simeq \{1, 0.119492 \pm 0.813835i, -0.738984\}$$

$$\pi_r = \begin{pmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{pmatrix}$$

$$\pi_l = \begin{pmatrix} \frac{2}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \end{pmatrix}$$

The metric entropy and complexity are

$$h_\mu = \frac{4}{5}$$

$$C_1 \simeq 1.92193$$

The connection matrices for the unequal branching minimal stochastic DeBrujin graph are

$$T_0 = \begin{pmatrix} \frac{3}{10} & \frac{7}{10} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}, \; T_0^0 = \begin{pmatrix} \frac{3}{10} & 0 & 0 & 0 \\ 0 & 0 & \frac{3}{10} & 0 \\ 0 & 0 & 0 & 0 \\ \frac{3}{10} & 0 & 0 & 0 \end{pmatrix}, \; T_0^1 = \begin{pmatrix} 0 & \frac{7}{10} & 0 & 0 \\ 0 & 0 & \frac{7}{10} & 0 \\ 0 & 0 & 0 & 1 \\ \frac{7}{10} & 0 & 0 & 0 \end{pmatrix}$$

The eigenvalue spectrum and left and right eigenvector corresponding to the maximal eigenvalue for the unequal branching minimal stochastic DeBrujin graph are

$$\mathbf{Sp}[T_1] \simeq \{1, 0.0739918 \pm 0.905546i, -0.847984\}$$

$$\pi_r = \begin{pmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{pmatrix}$$

$$\pi_l \simeq (\, 0.322581 \quad 0.225806 \quad 0.225806 \quad 0.225806 \,)$$

The metric entropy and complexity are

$$h_\mu \simeq 0.682289$$

$$C_1 \simeq 1.98085$$

Here as in the case of the band merging example above, we must use the stochastic DeBrujin machine for the computation of the complexities and metric entropy.

Semigroup

The semigroup structure is much more complex in this example than the others; the semigroup has eighty one elements. As has been proven in[10] and demonstrated in the

above examples, the principal components are simply identical copies of the minimal machine. Therefore to simplify the discussion of this example we will represent the set of principal components, i.e the minimal graph, as a heptagon, shown in figure 97 ,with each edge of the heptagon representing a particular vertex of the minimal machine. We then explicitly show the semigroup transients in figure 98 , for the right representation, with transient strings entering the appropriate vertex of a particular principal component and with the edge that enters the principal component labeled by the number of the principal component which it enters. We note that in this case there are nine principal components.

Language properties

As the semigroup elements are too numerous to list we will only list the first few forbidden words for this example.

The forbidden words, F_S, are

$$g_{81} = e = \{00100, 00110, 010101, 010111, ...\}$$

Conclusions

We have demonstrated with these examples, methods for analyzing some of the rich structure of finitary systems. Even in relatively simple cases such as the even system, the formal language and semigroup properties generate intricate self-similar structures as seen in the cylinder histograms.This is a general characteristic of strictly sofic systems, i.e systems with complex transient behavior in their semigroup representations. Sofic systems with more complex transients than the even system, i.e. the Misiurewicz system and the transient chaos system, were also examined.

We have provided methods for determining whether or not a given system is strictly sofic and if so how this governs the orbit distribution. In fact, for sofic systems, the semigroup structure is required for the construction of the orbit distribution. For a more detailed discussion and the underlying theory of finitary processes we refer the reader to the long paper [4]

Acknowledgements

The authors wish to thank Eric Friedman and Jim Hanson for useful discussions. They are grateful for the continuing support of Carson Jeffries. KY wishes to thank Jeff Scargle and Dave Donoho for support during the period for which this work was done.

Figures

Figure Captions

Figure 49 The reconstructed ϵ-machine for the period three system.

Figure 50 Cylinder probability histograms for cylinders of length 1 through 9 for the period three system.

Figure 51 Tree representation of cylinder histograms for cylinders of length 1 through 8 with cylinder probabilities given by gray scale values, for the period three system.

Figure 52 Tree representation of cylinder histograms for cylinders of length 1 through 8 with branching probabilities given by gray scale values, for the period three system.

Figure 53 Right semigroup graph for the period three system.

Figure 54 The reconstructed ϵ-machine for the golden mean system

Figure 55 Cylinder probability histograms for cylinders of length 1 through 9 for the golden mean system with equal branching.

Figure 56 Tree representation of cylinder histograms for cylinders of length 1 through 8 with cylinder probabilities given by gray scale values, for the golden mean system with equal branching.

Figure 57 Tree representation of cylinder histograms for cylinders of length 1 through 8 with branching probabilities given by gray scale values, for the golden mean system with equal branching.

Figure 58 Cylinder probability histograms for cylinders of length 1 through 9 for the golden mean system with unequal branching.

Figure 59 Tree representation of cylinder histograms for cylinders of length 1 through 8 with cylinder probabilities given by gray scale values, for the golden mean system with unequal branching.

Figure 60 Tree representation of cylinder histograms for cylinders of length 1 through 8 with branching probabilities given by gray scale values, for the golden mean system with unequal branching.

Figure 61 Right semigroup graph for the golden mean system.

Figure 62 Left semigroup graph for the golden mean system.

Figure 63 The reconstructed ϵ-machine for the even sofic system: no even numbers of ones.

Figure 64 The semigroup graph for the even system with branching probabilities.

Figure 65 The tree of cylinders with cylinder probabilities for the even system.

Figure 66 Cylinder probability histograms for cylinders of length 1 through 9 for the even system with equal branching.

Figure 67 Tree representation of cylinder histograms for cylinders of length 1 through 8 with cylinder probabilities given by gray scale values, for the even system with equal branching.

Figure 68 Tree representation of cylinder histograms for cylinders of length 1 through 8 with branching probabilities given by gray scale values, for the even system with equal branching.

Figure 69 Cylinder probability histograms for cylinders of length 1 through 9 for the even system with unequal branching.

Figure 70 Tree representation of cylinder histograms for cylinders of length 1 through 8 with cylinder probabilities given by gray scale values, for the even system with unequal branching.

Figure 71 Tree representation of cylinder histograms for cylinders of length 1 through 8 with branching probabilities given by gray scale values, for the even system with unequal branching.

Figure 72 Right semigroup graph for the even system.

Figure 73 Left semigroup graph for the even system.

Figure 74 The reconstructed ϵ-machine for the band merging system

Figure 75 Cylinder probability histograms for cylinders of length 1 through 9 for the band merging $2 \rightarrow 1$ system in the case of equal branching.

Figure 76 Tree representation of cylinder histograms for cylinders of length 1 through 8 with cylinder probabilities given by gray scale values, for the band merging $2 \rightarrow 1$ system in the case of equal branching.

Figure 77 Tree representation of cylinder histograms for cylinders of length 1 through 8 with branching probabilities given by gray scale values, for the band merging $2 \rightarrow 1$ system in the case of equal branching.

Figure 78 Cylinder probability histograms for cylinders of length 1 through 9 for the band merging $2 \rightarrow 1$ system in the case of unequal branching.

Figure 79 Tree representation of cylinder histograms for cylinders of length 1 through 8 with cylinder probabilities given by gray scale values, for the band merging $2 \rightarrow 1$ system in the case of unequal branching.

Figure 80 Tree representation of cylinder histograms for cylinders of length 1 through 8 with branching probabilities given by gray scale values, for the band merging $2 \rightarrow 1$ system in the case of unequal branching.

Figure 81 Right semigroup graph for the band merging system

Figure 82 The reconstructed ϵ-machine for the Misiurewicz system

Figure 83 Cylinder probability histograms for cylinders of length 1 through 9 for the Misiurewicz system in the case of equal branching.

Figure 84 Tree representation of cylinder histograms for cylinders of length 1 through 8 with cylinder probabilities given by gray scale values, for the Misiurewicz system in the case of equal branching.

Figure 85 Tree representation of cylinder histograms for cylinders of length 1 through 8 with branching probabilities given by gray scale values, for the Misiurewicz system in the case of equal branching.

Figure 86 Cylinder probability histograms for cylinders of length 1 through 9 for the Misiurewicz system in the case of unequal branching.

Figure 87 Tree representation of cylinder histograms for cylinders of length 1 through 8 with cylinder probabilities given by gray scale values, for the Misiurewicz system in the case of unequal branching.

Figure 88 Tree representation of cylinder histograms for cylinders of length 1 through 8 with branching probabilities given by gray scale values, for the Misiurewicz system in the case of unequal branching.

Figure 89 Right semigroup graph for the Misiurewicz system.

Figure 90 Reconstructed ϵ-machine for the transient chaos system

Figure 91 Cylinder probability histograms for cylinders of length 1 through 9 for the transient chaos system with equal branching.

Figure 92 Tree representation of cylinder histograms for cylinders of length 1 through 8 with cylinder probabilities given by gray scale values, for the transient chaos system with equal branching.

Figure 93 Tree representation of cylinder histograms for cylinders of length 1 through 8 with branching probabilities given by gray scale values, for the transient chaos system with equal branching.

Figure 94 Cylinder probability histograms for cylinders of length 1 through 9 for the transient chaos system with unequal branching.

Figure 95 Tree representation of cylinder histograms for cylinders of length 1 through 8 with cylinder probabilities given by gray scale values, for the transient chaos system with unequal branching.

Figure 96 Tree representation of cylinder histograms for cylinders of length 1 through 8 with branching probabilities given by gray scale values, for the transient chaos system with unequal branching.

Figure 97 Schematic representation of the principal components for the transient chaos system

Figure 98 Right semigroup graph for the transient chaos system

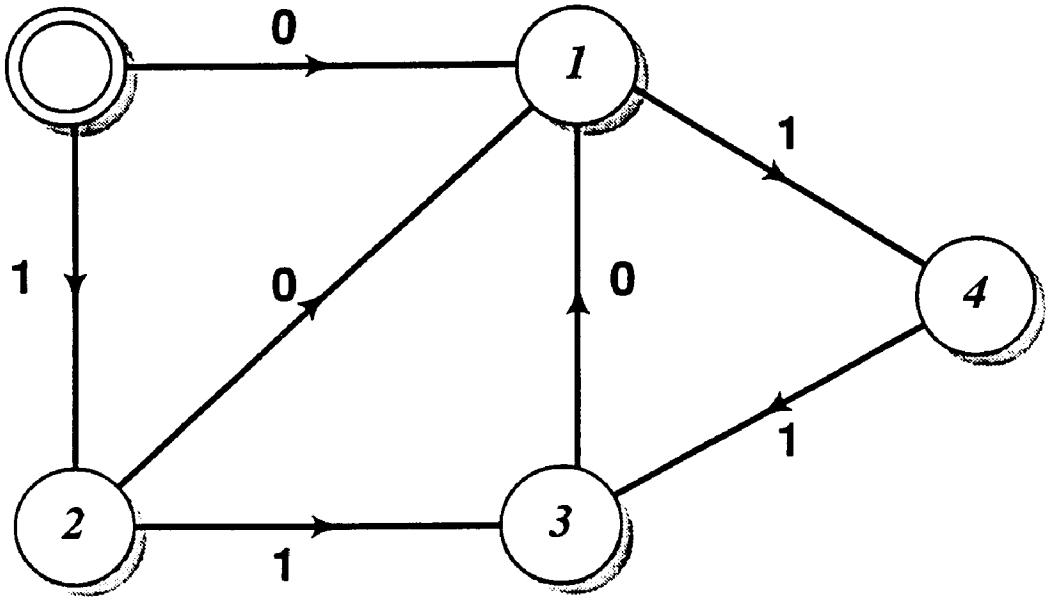


Figure 49

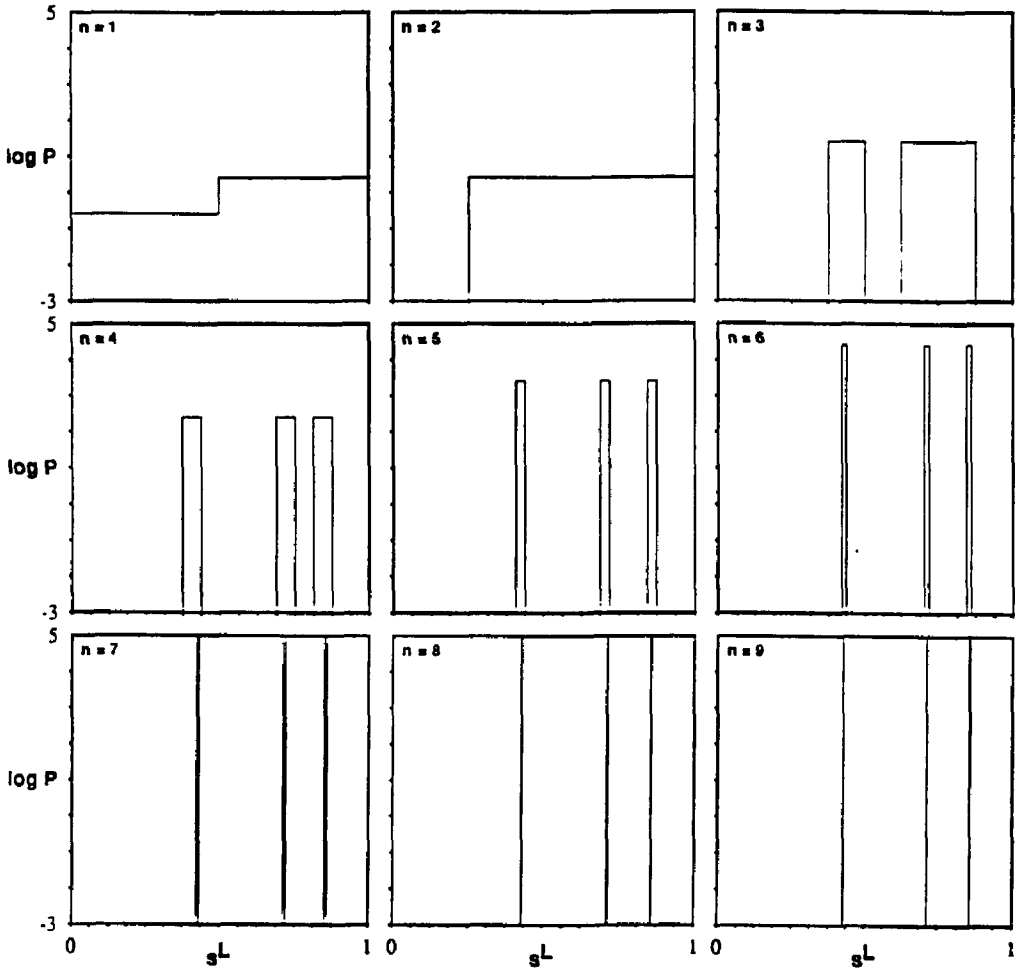


Figure 50

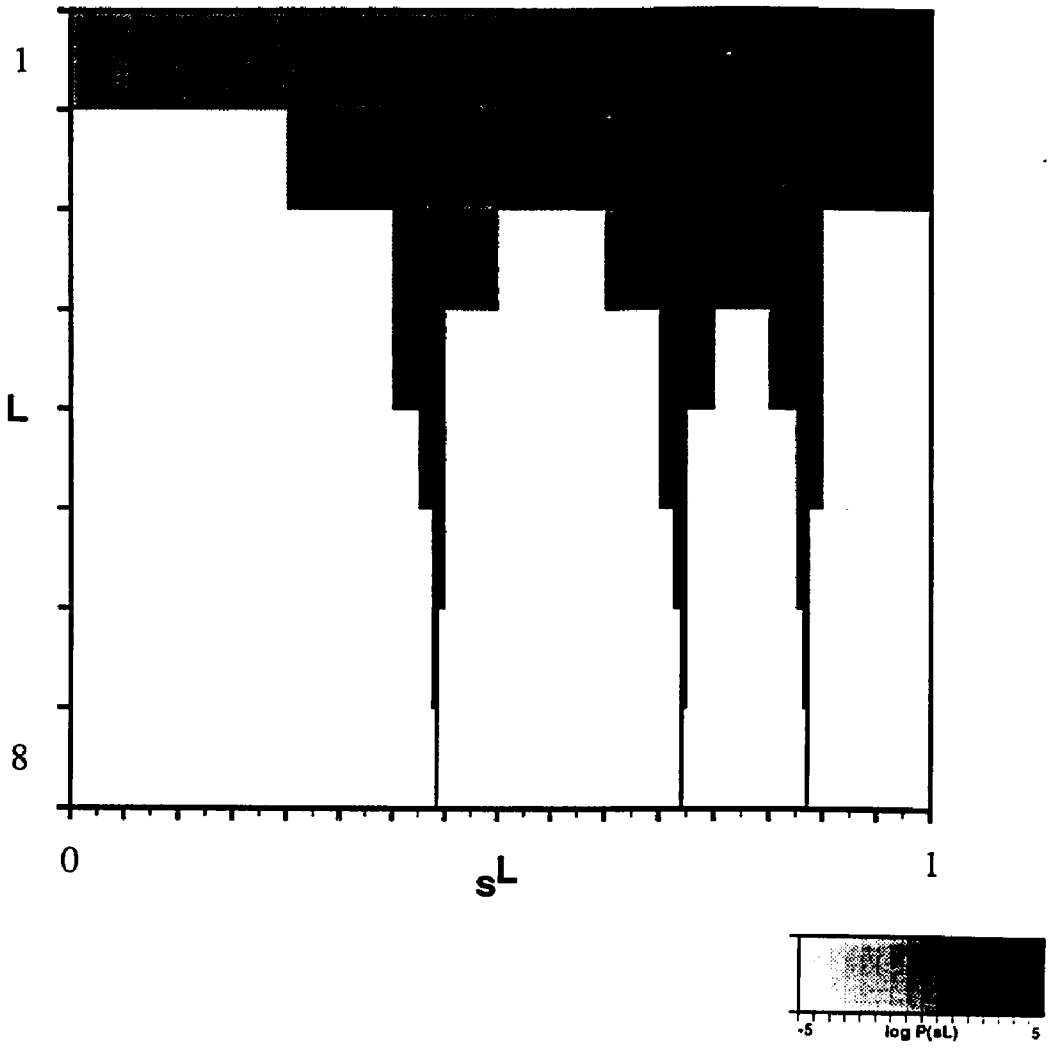


Figure 51

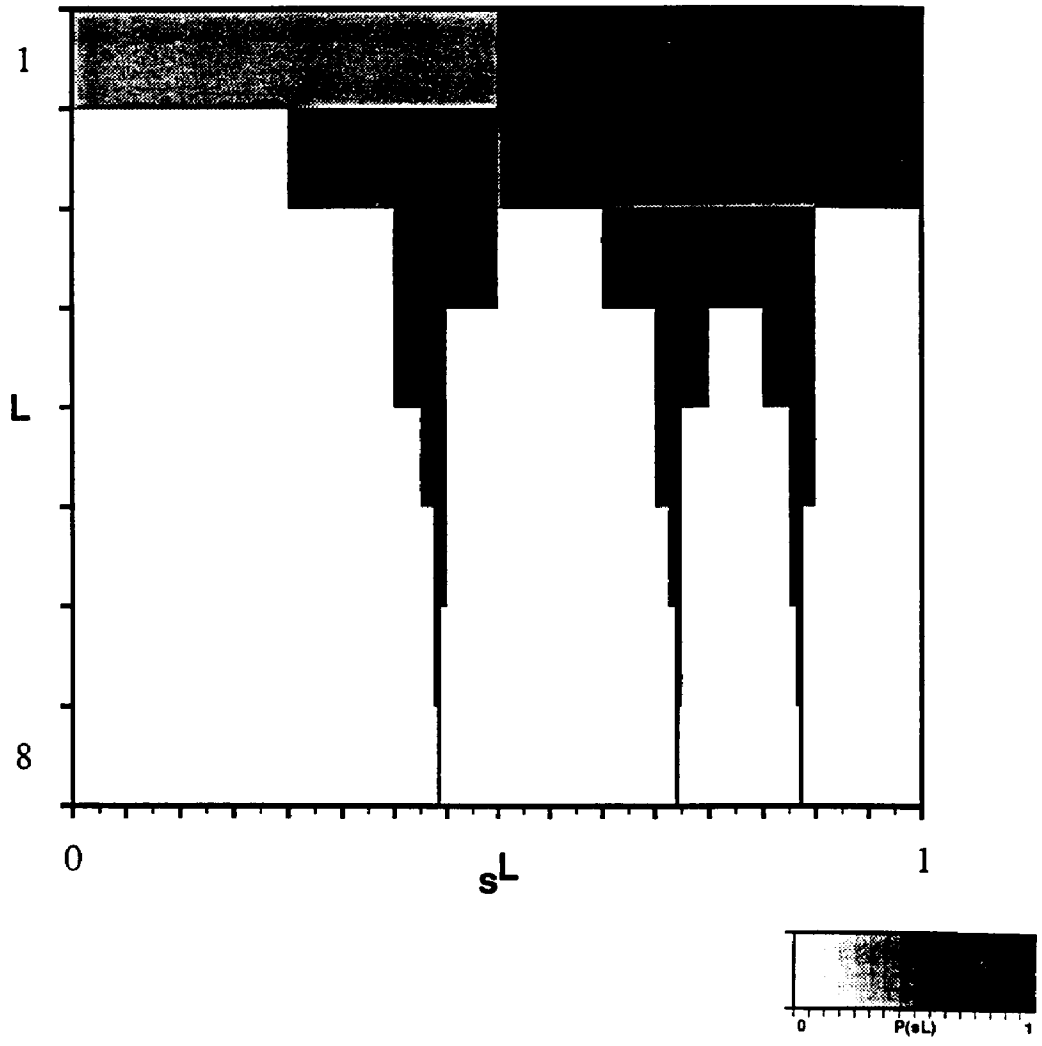


Figure 52

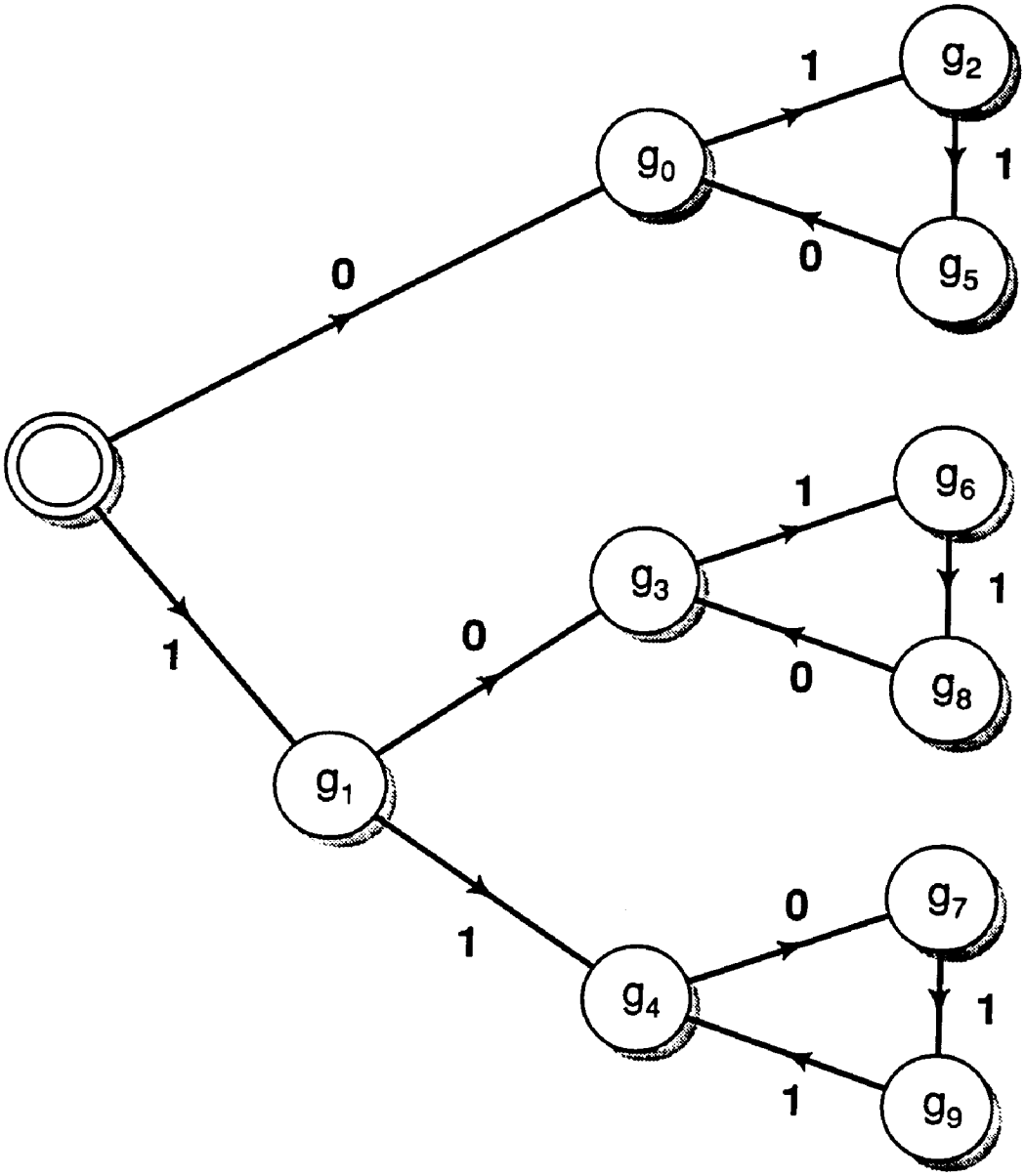


Figure 53

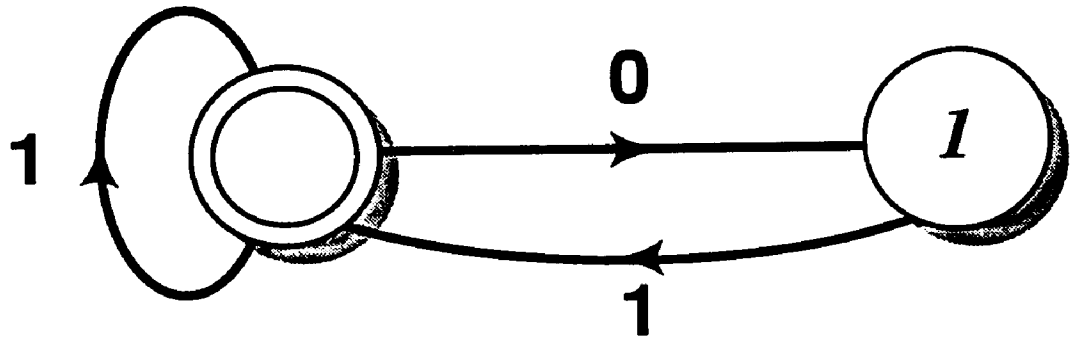


Figure 54

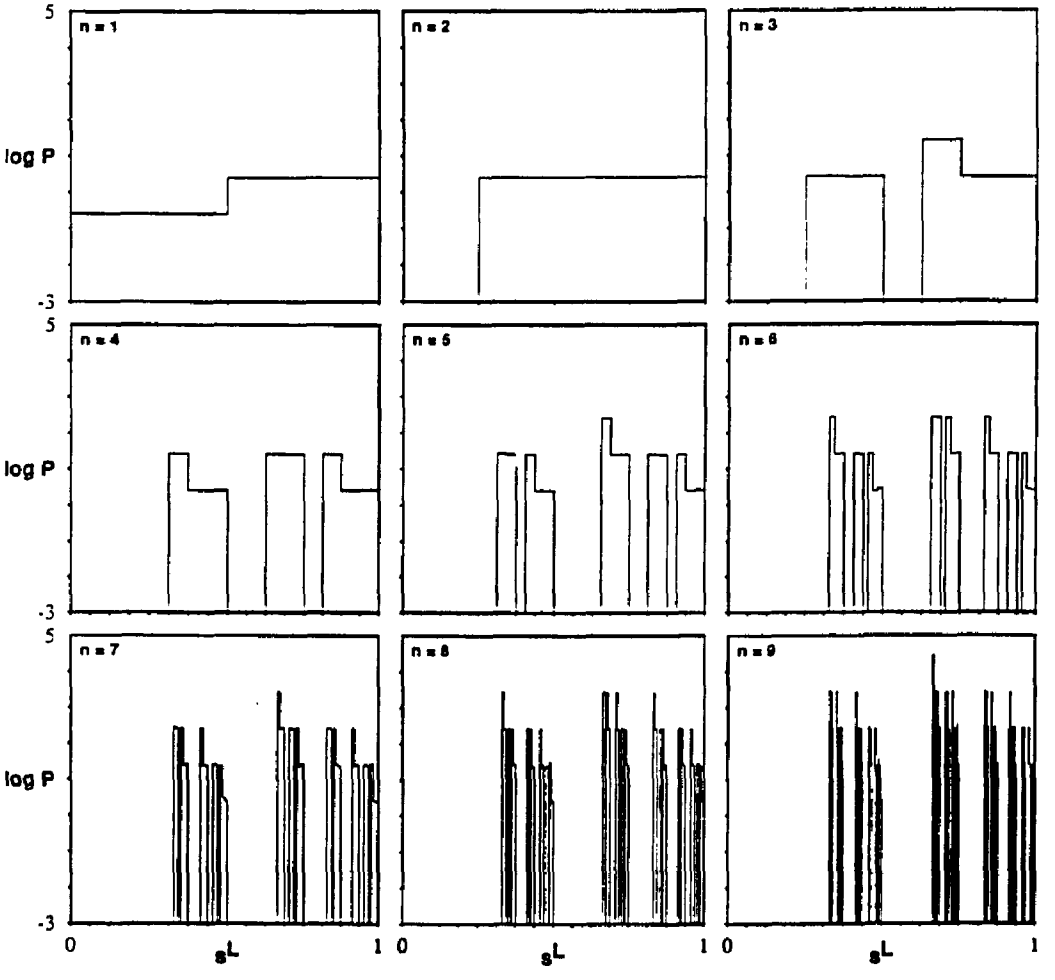


Figure 55

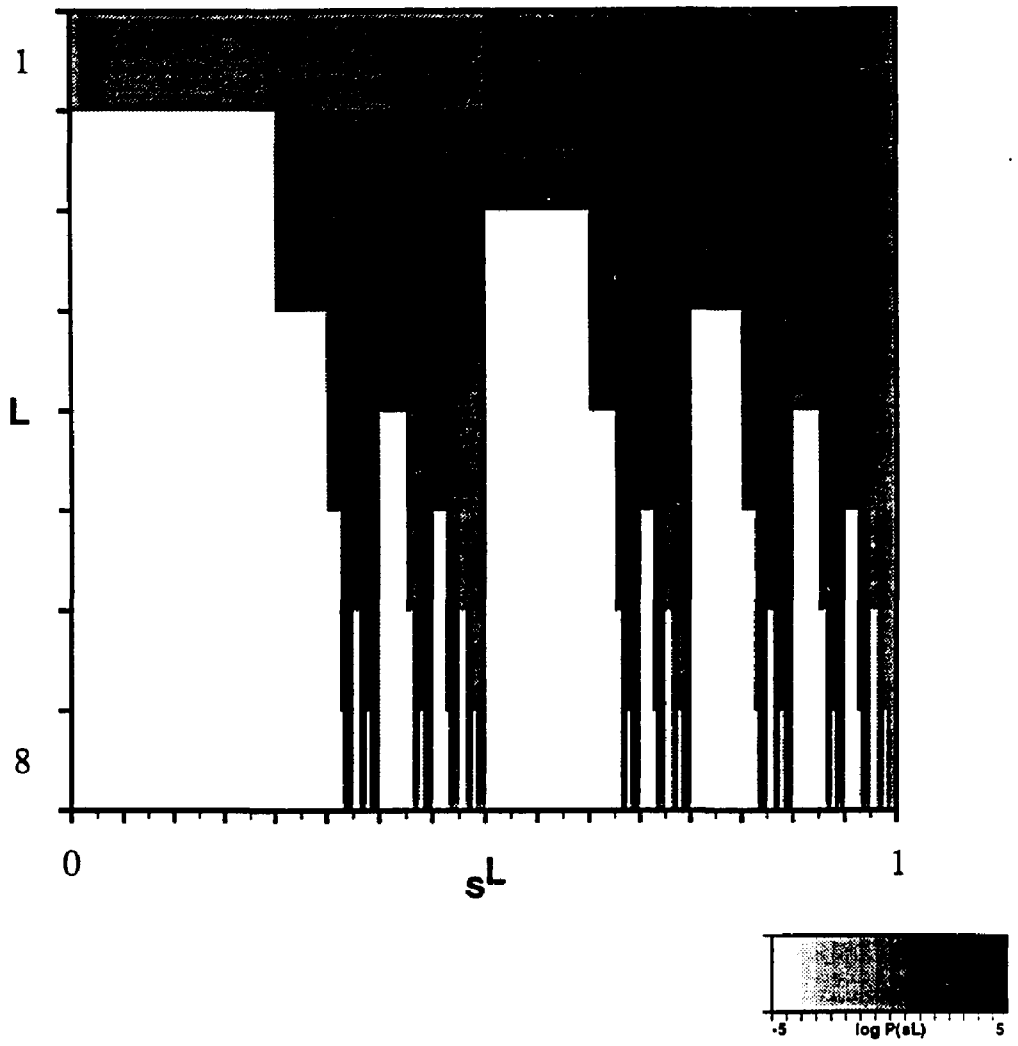


Figure 56

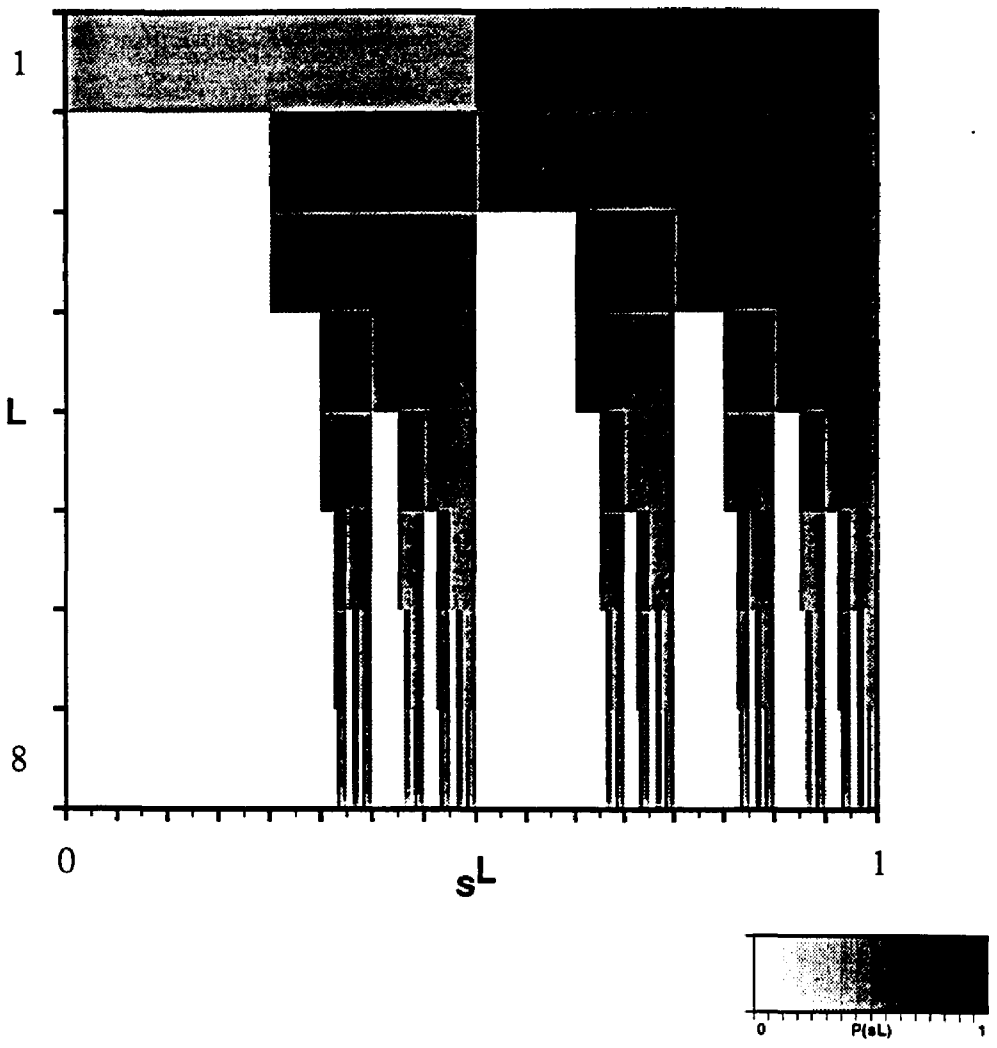


Figure 57

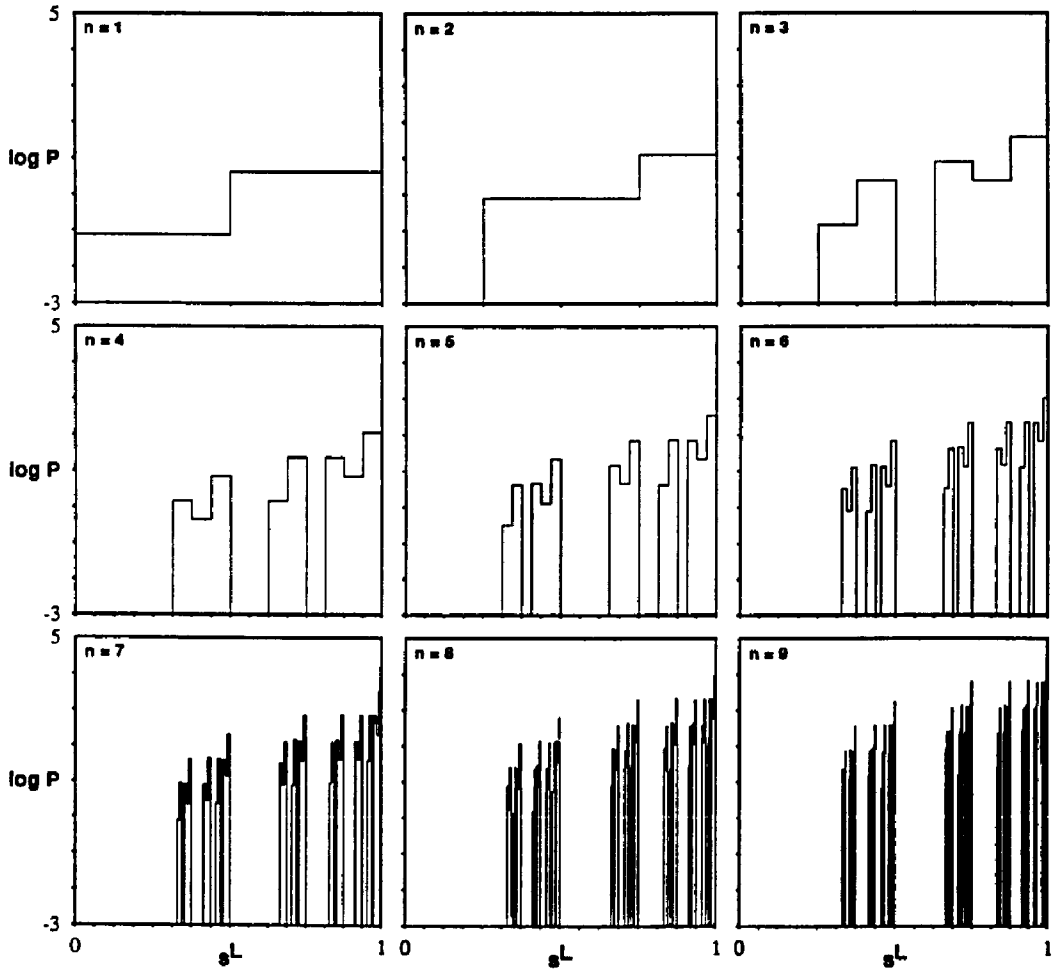


Figure 58

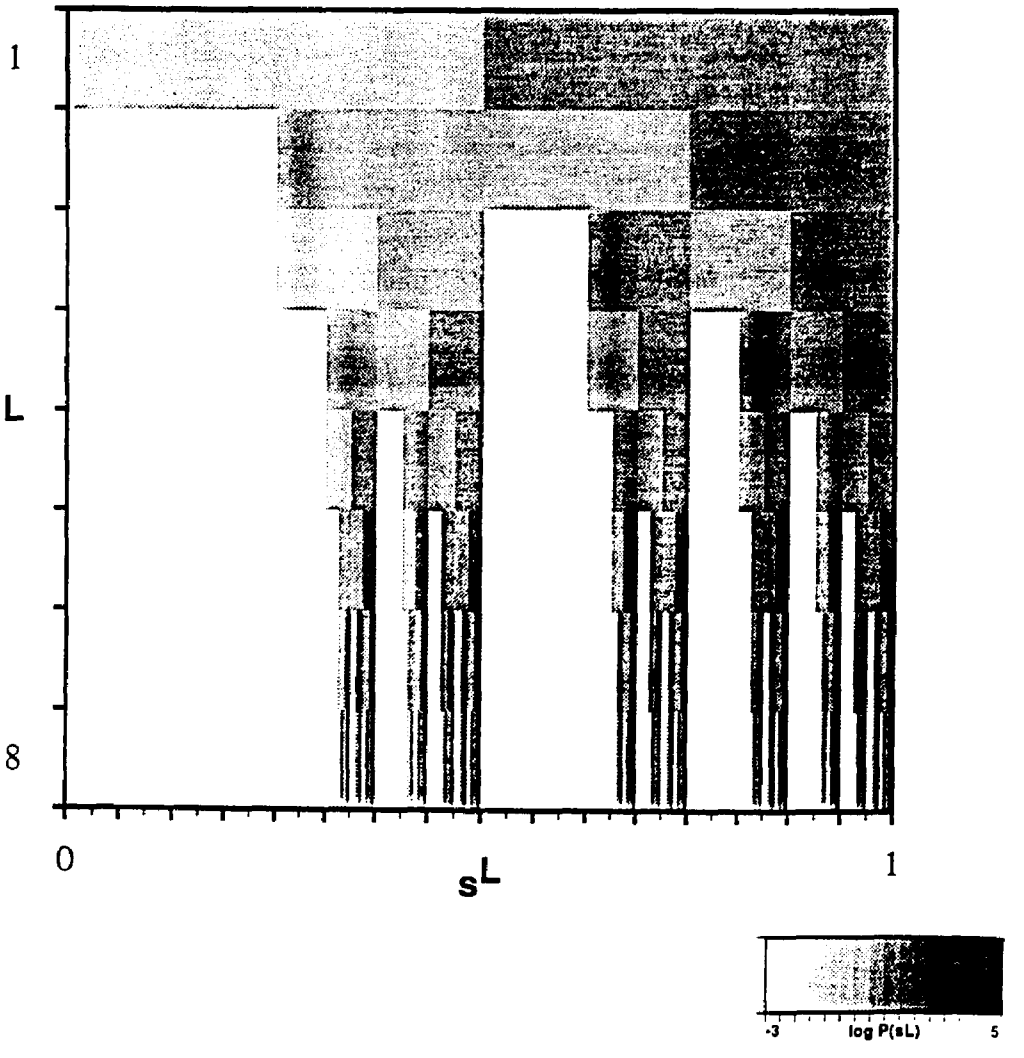


Figure 59

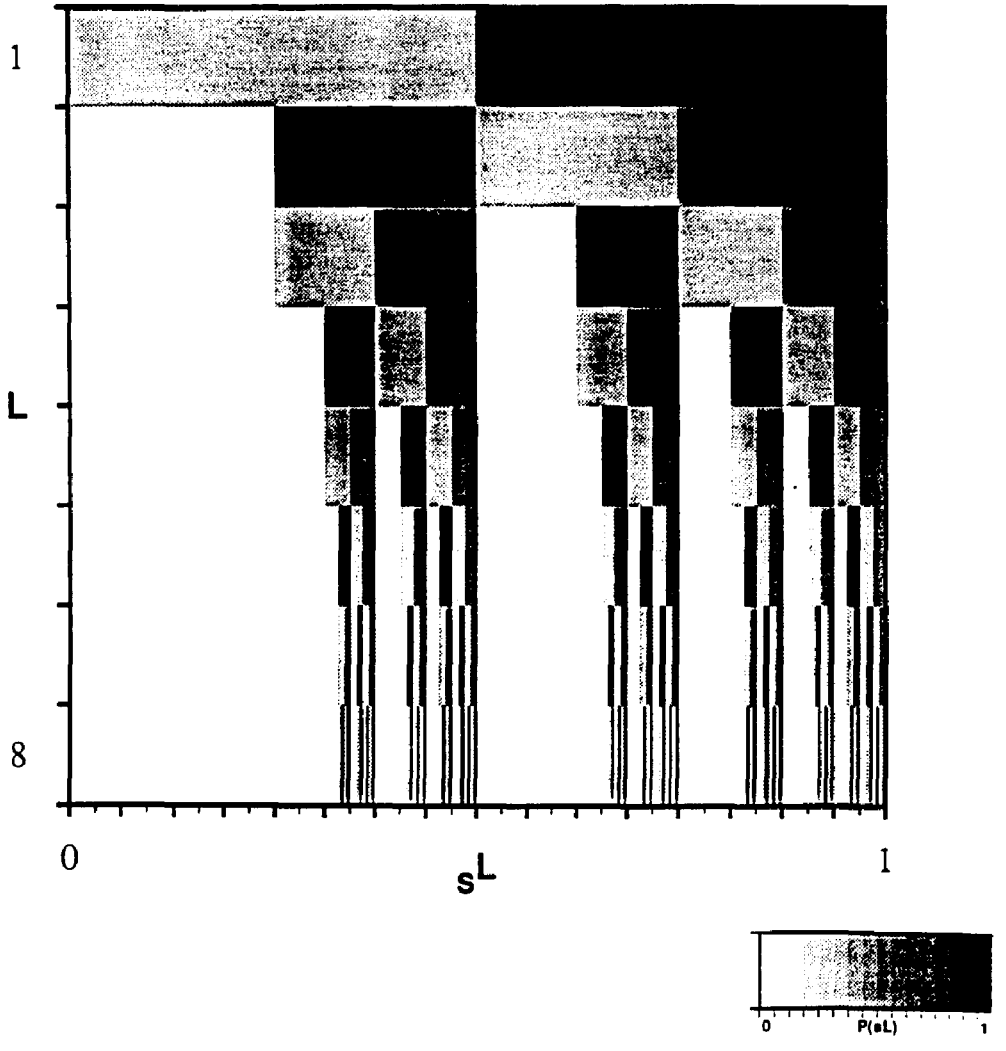


Figure 60

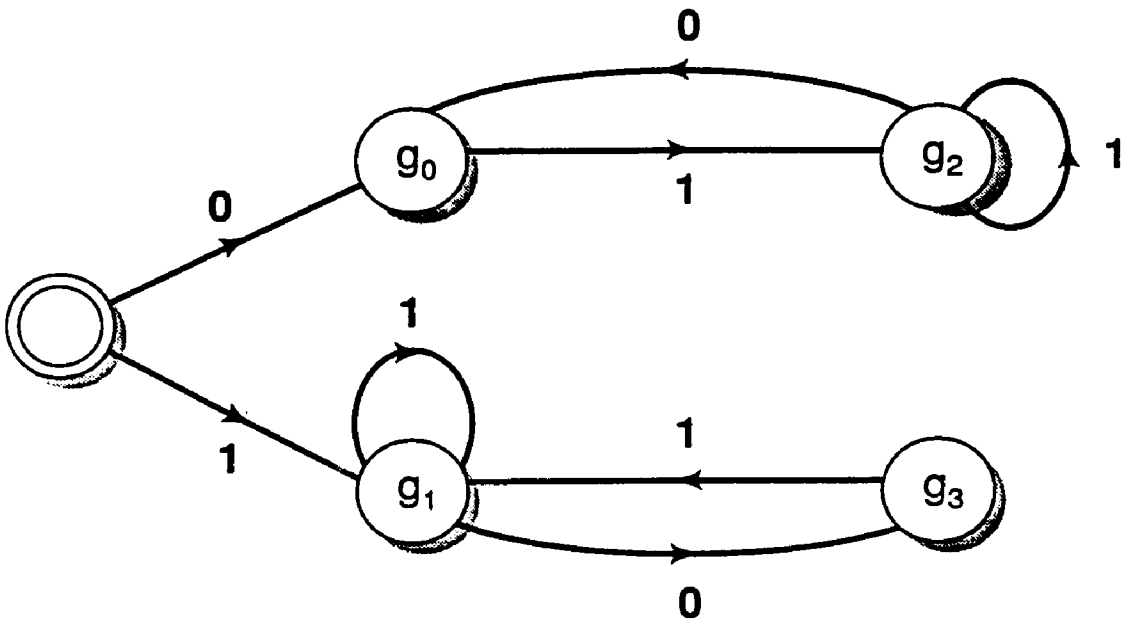


Figure 61

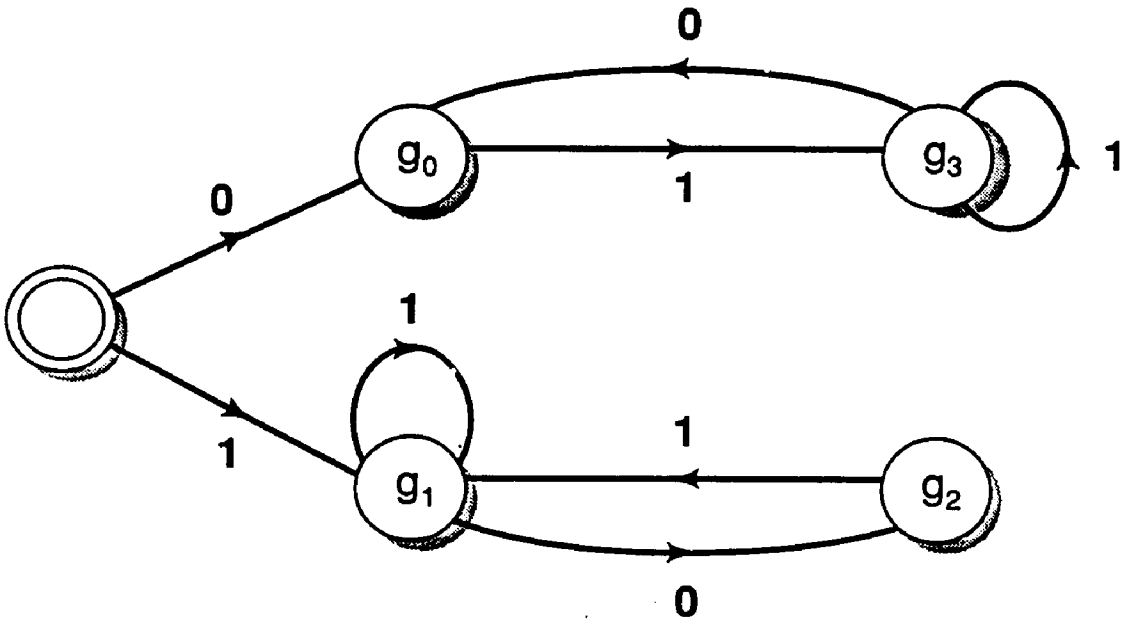


Figure 62

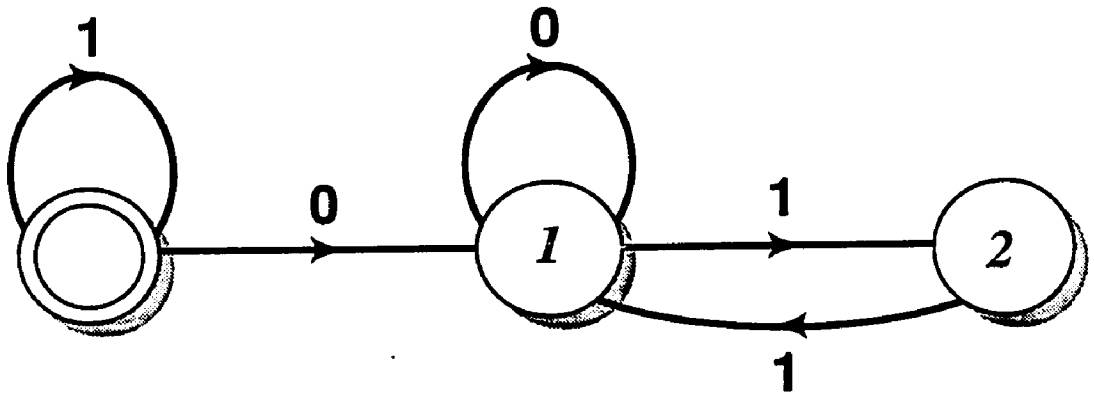


Figure 63

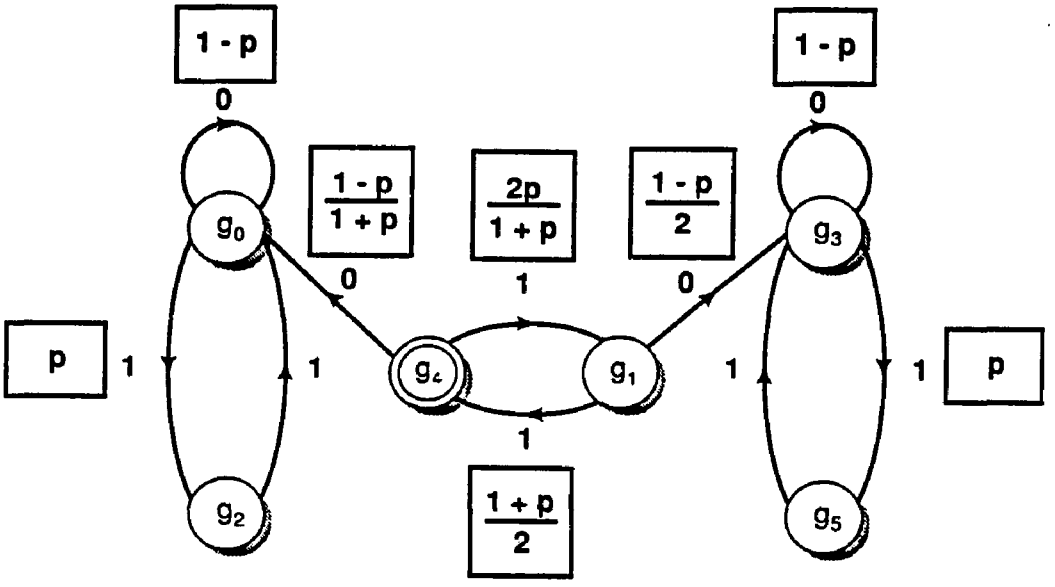


Figure 64

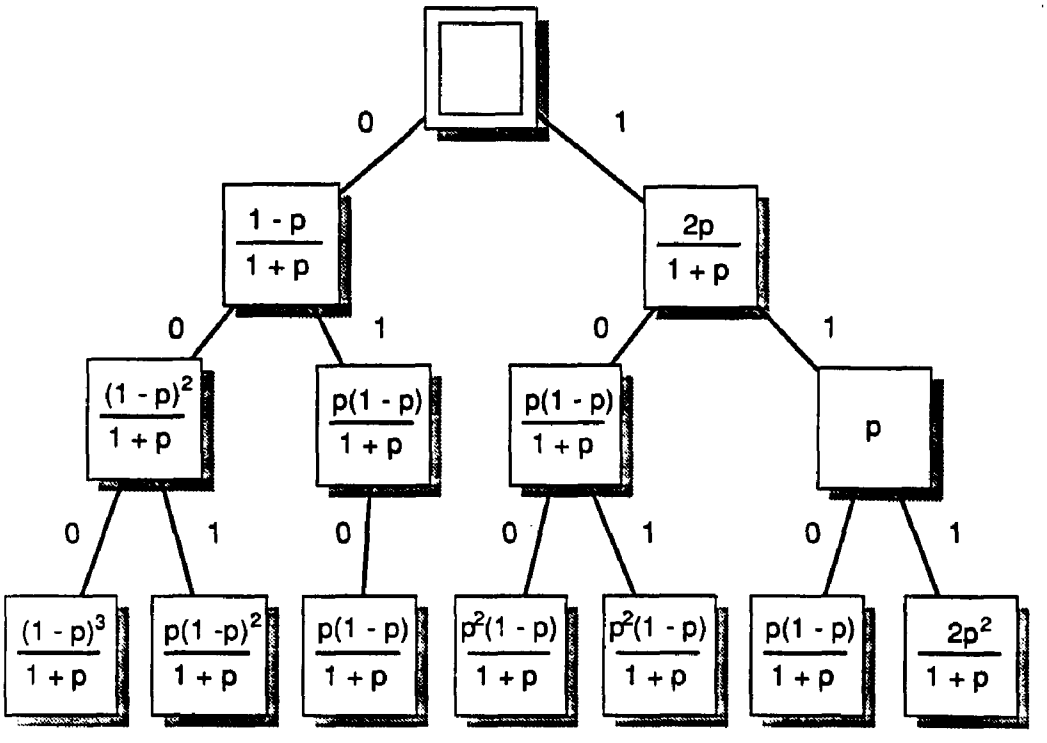

0
1
$\frac{1-p}{1+p}$
$\frac{2p}{1+p}$
0
1
0
1
$\frac{(1-p)^2}{1+p}$
$\frac{p(1-p)}{1+p}$
$\frac{p(1-p)}{1+p}$
p
0
1
0
1
0
1
0
1
$\frac{(1-p)^3}{1+p}$
$\frac{p(1-p)^2}{1+p}$
$\frac{p(1-p)}{1+p}$
$\frac{p^2(1-p)}{1+p}$
$\frac{p^2(1-p)}{1+p}$
$\frac{p(1-p)}{1+p}$
$\frac{2p^2}{1+p}$

Figure 65

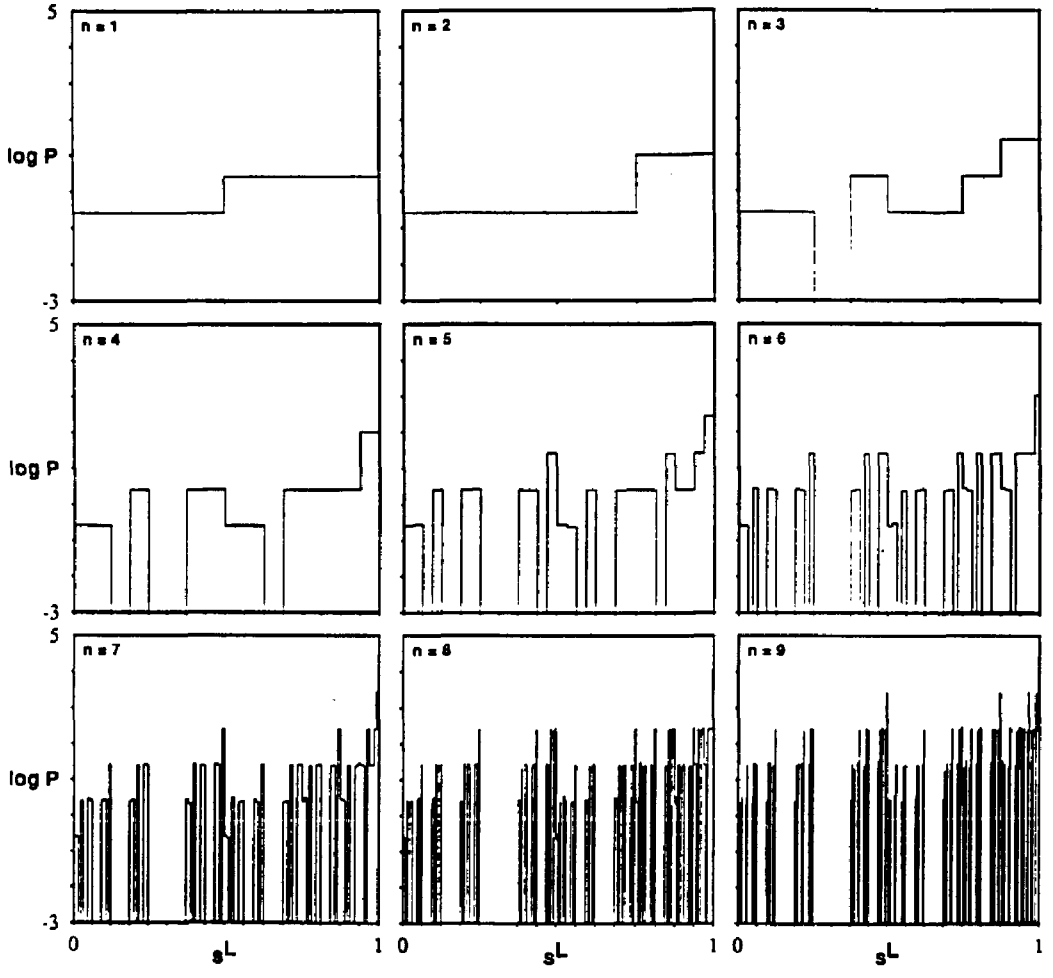


Figure 66

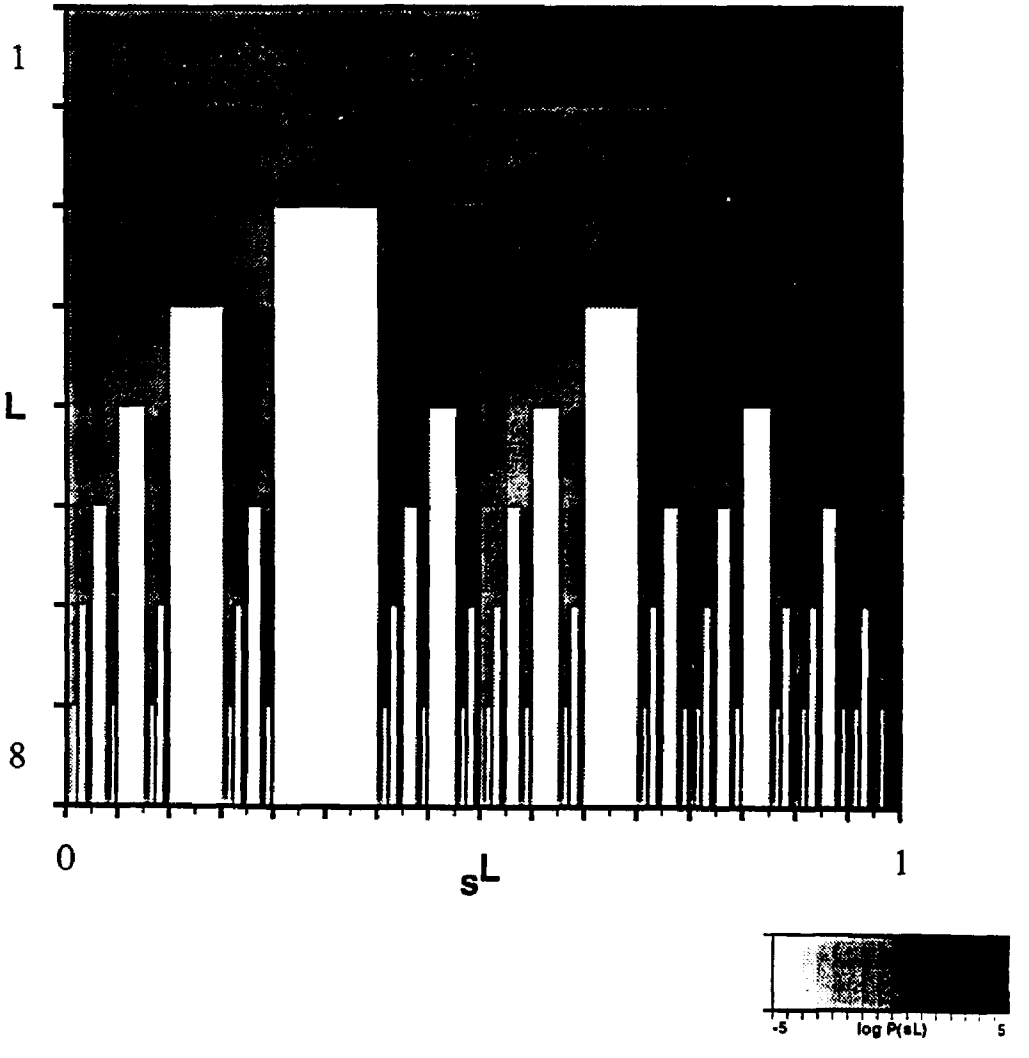


Figure 67

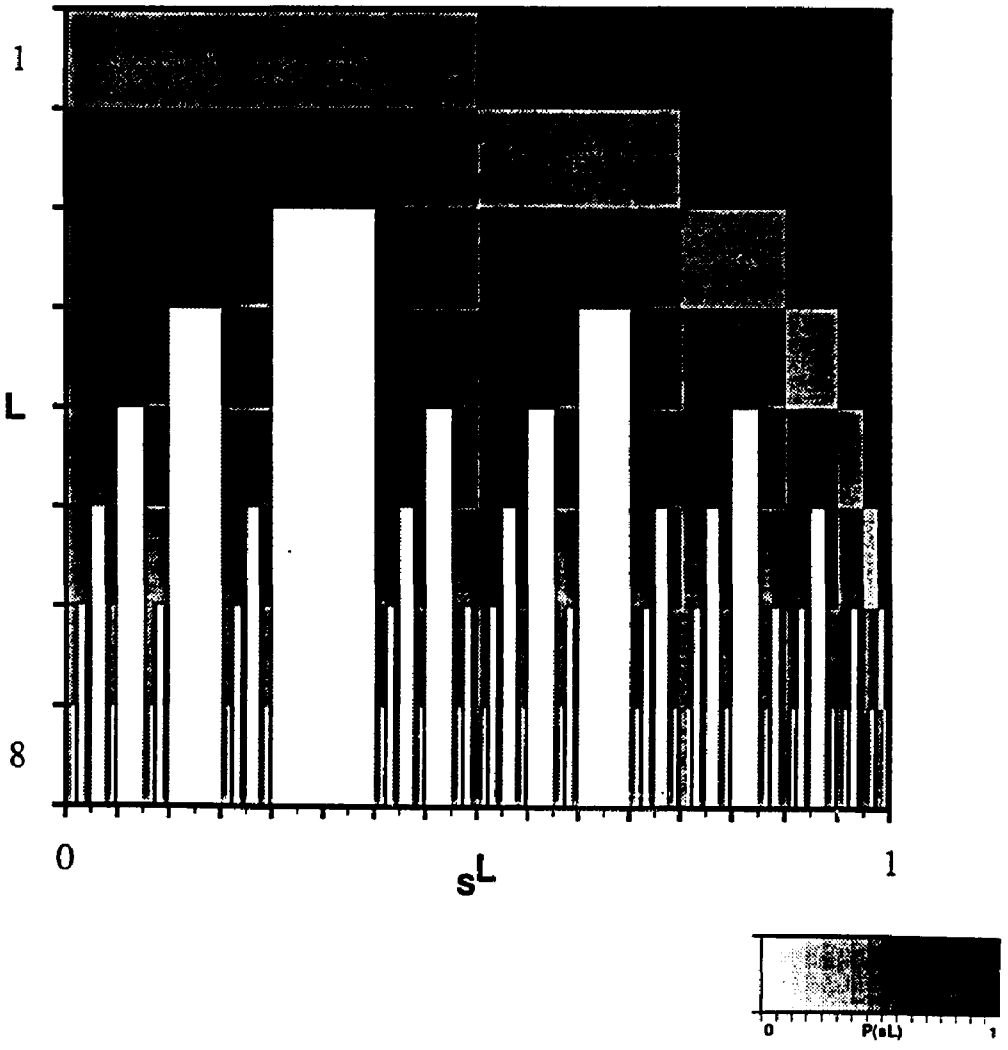


Figure 68

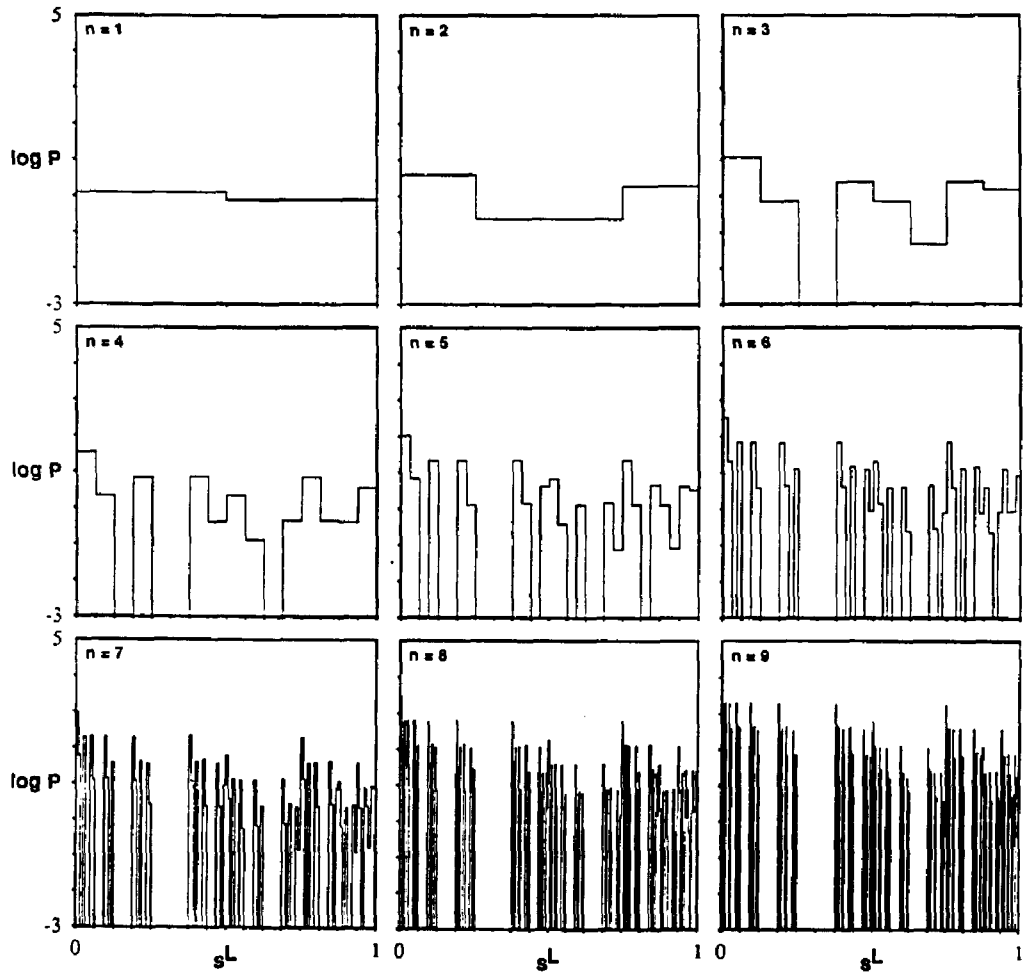


Figure 69

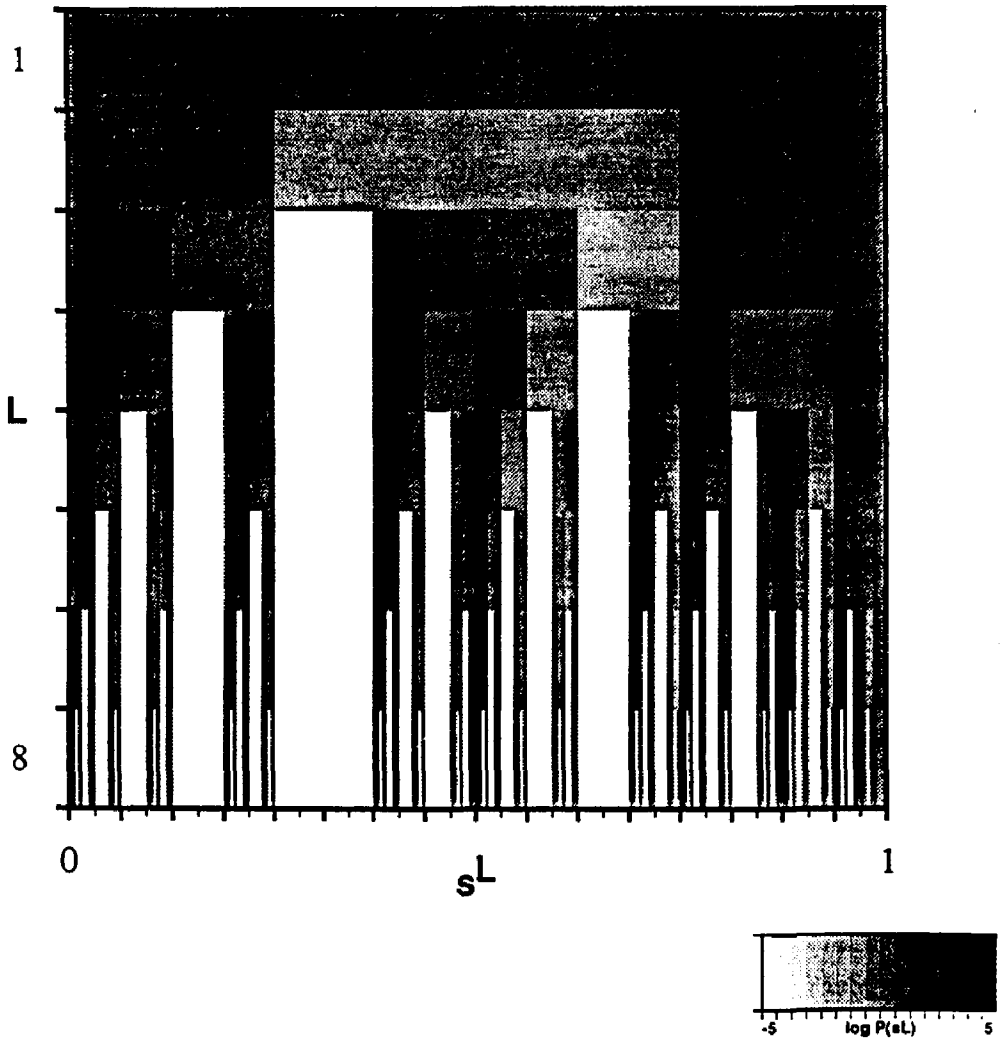


Figure 70

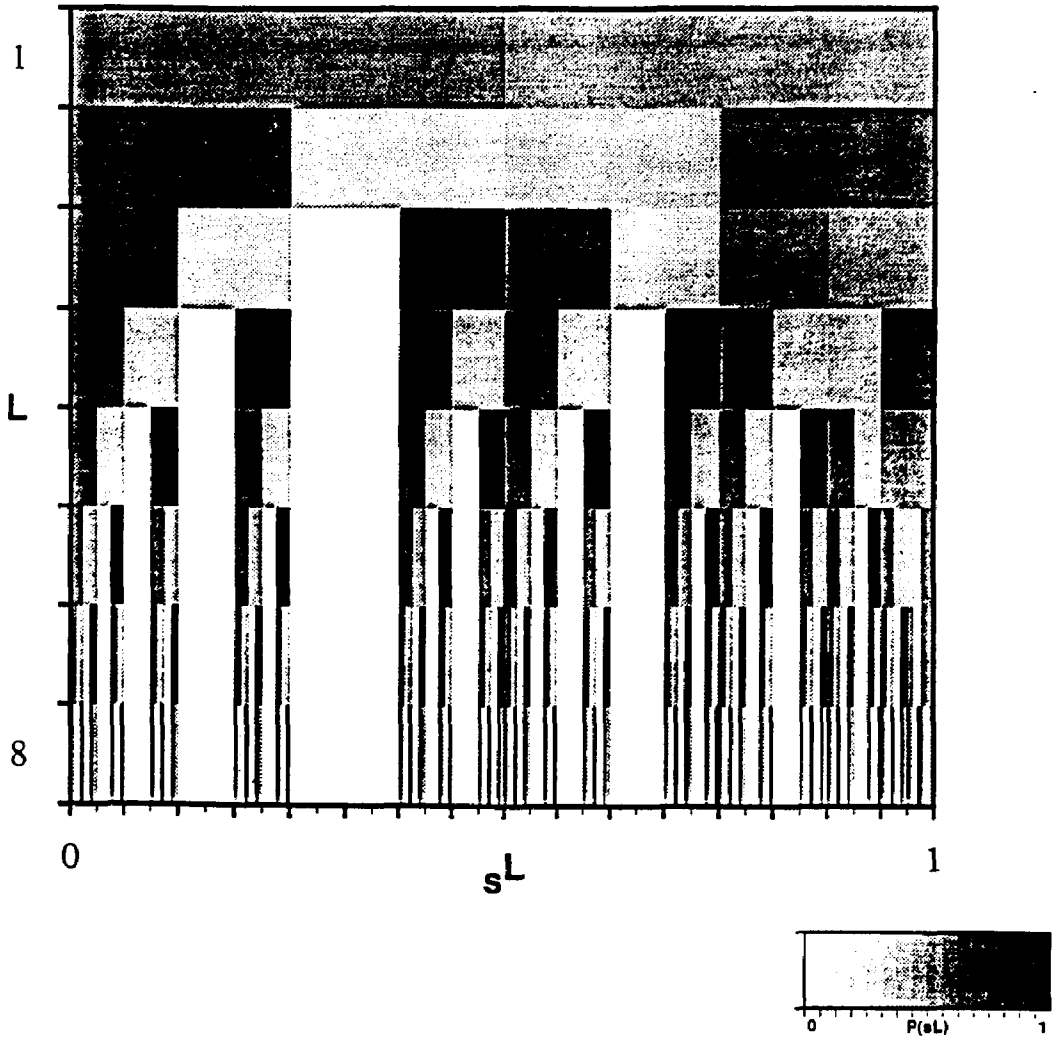


Figure 71

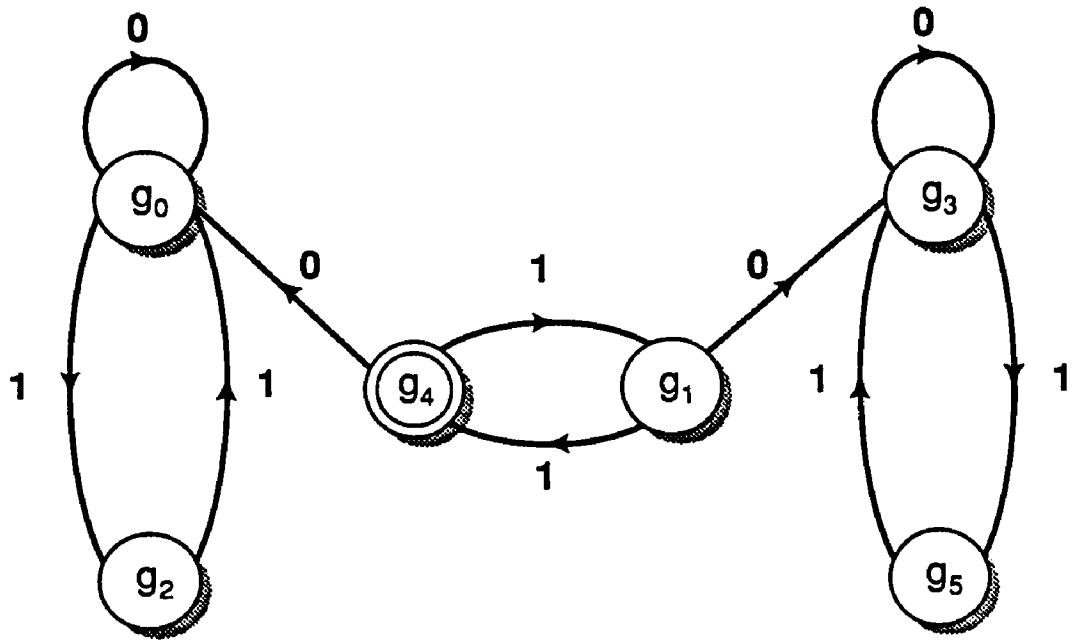


Figure 72

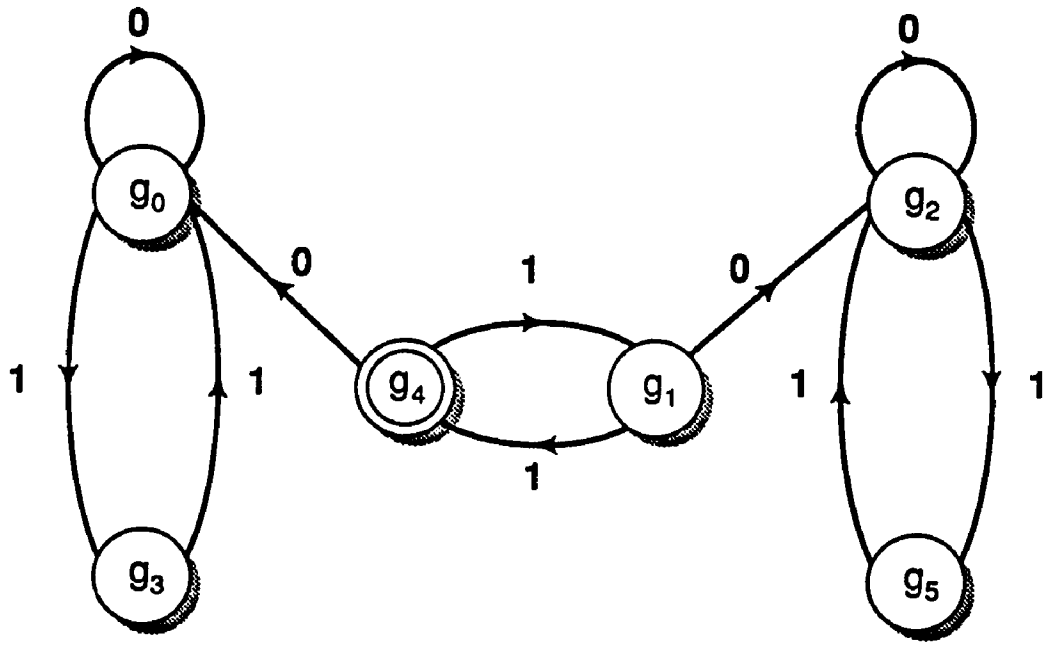


Figure 73

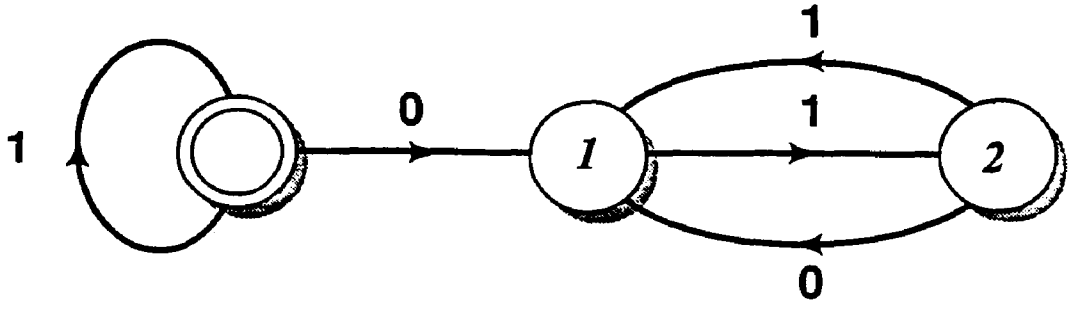


Figure 74

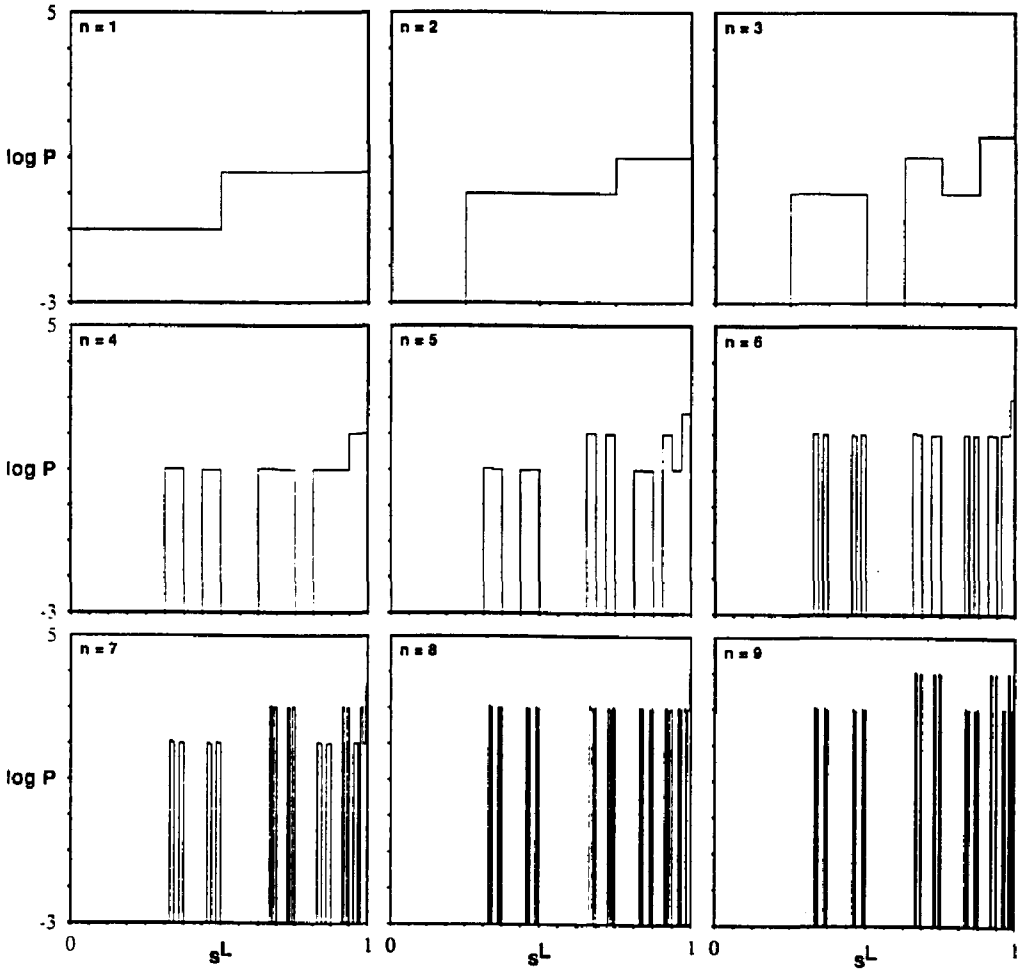


Figure 75

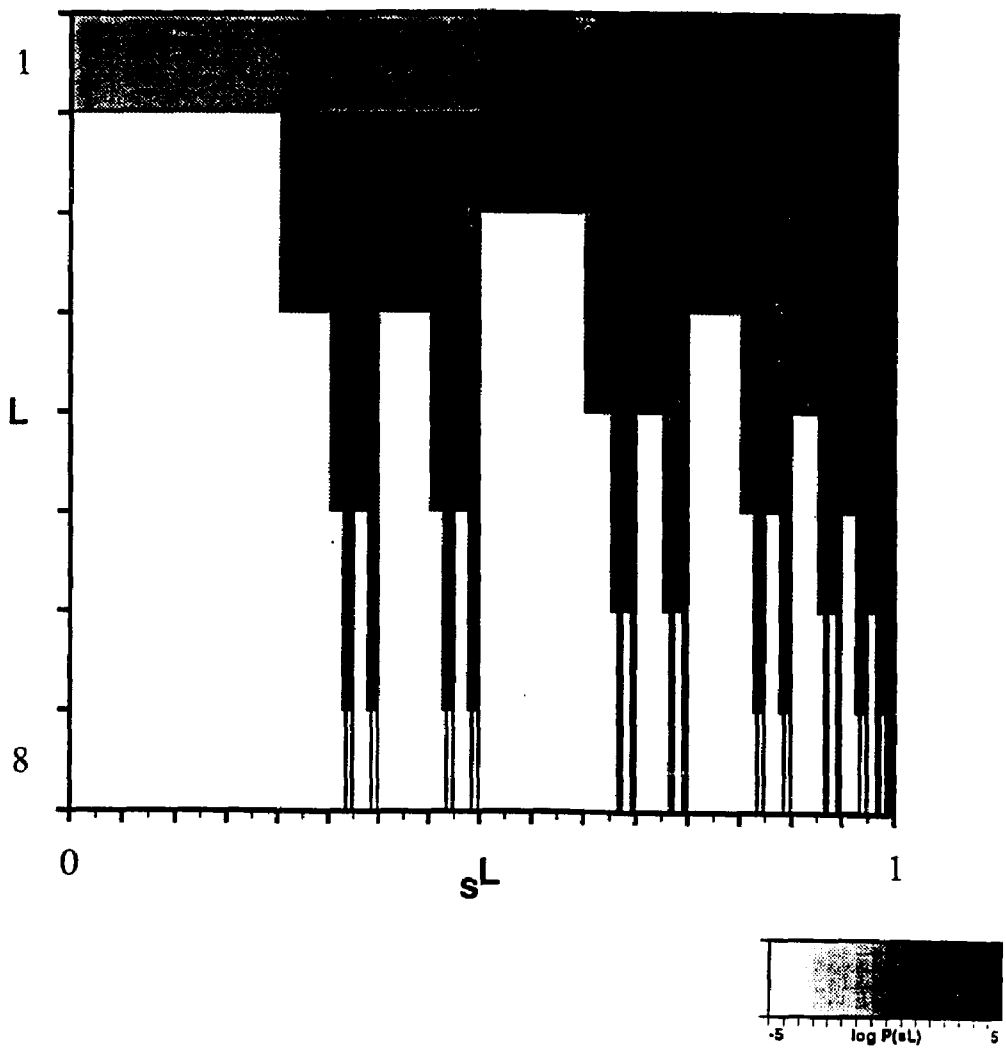


Figure 76

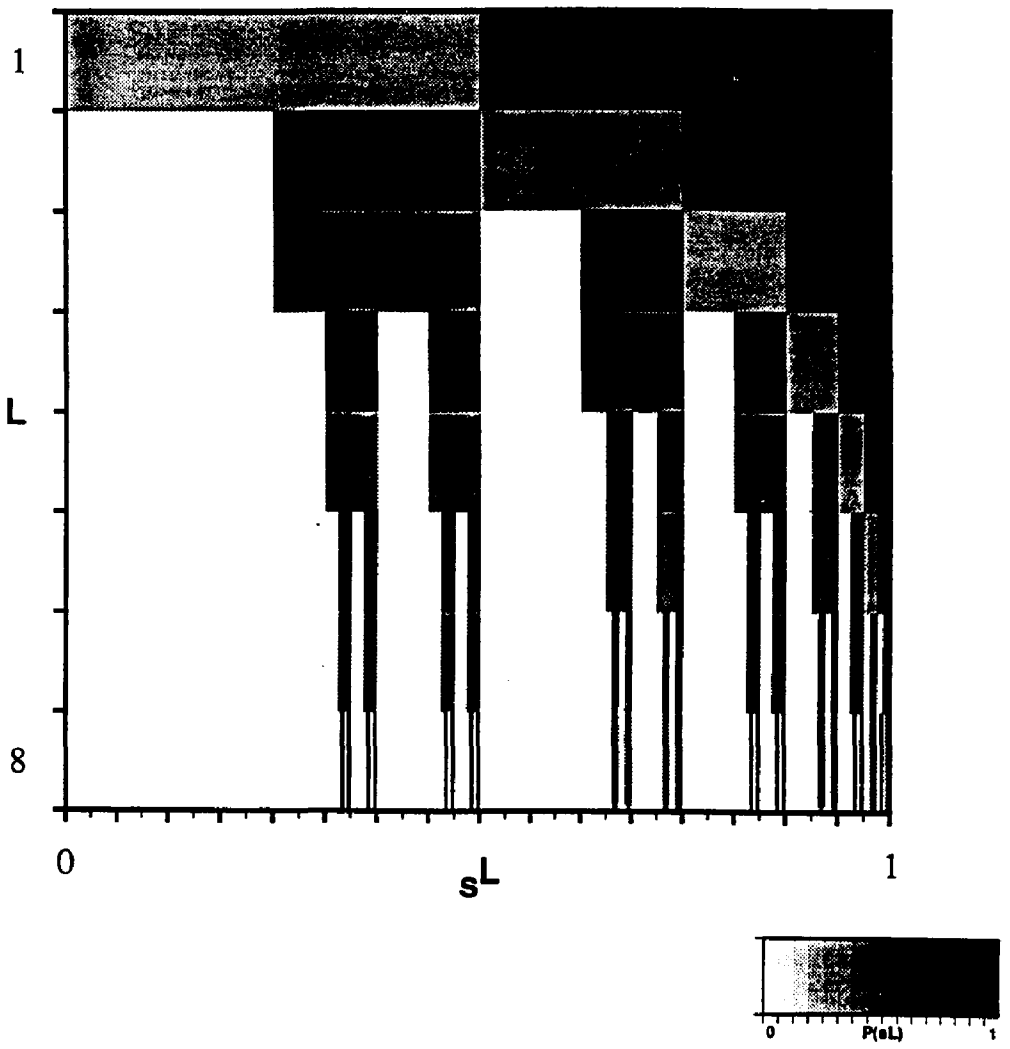


Figure 77

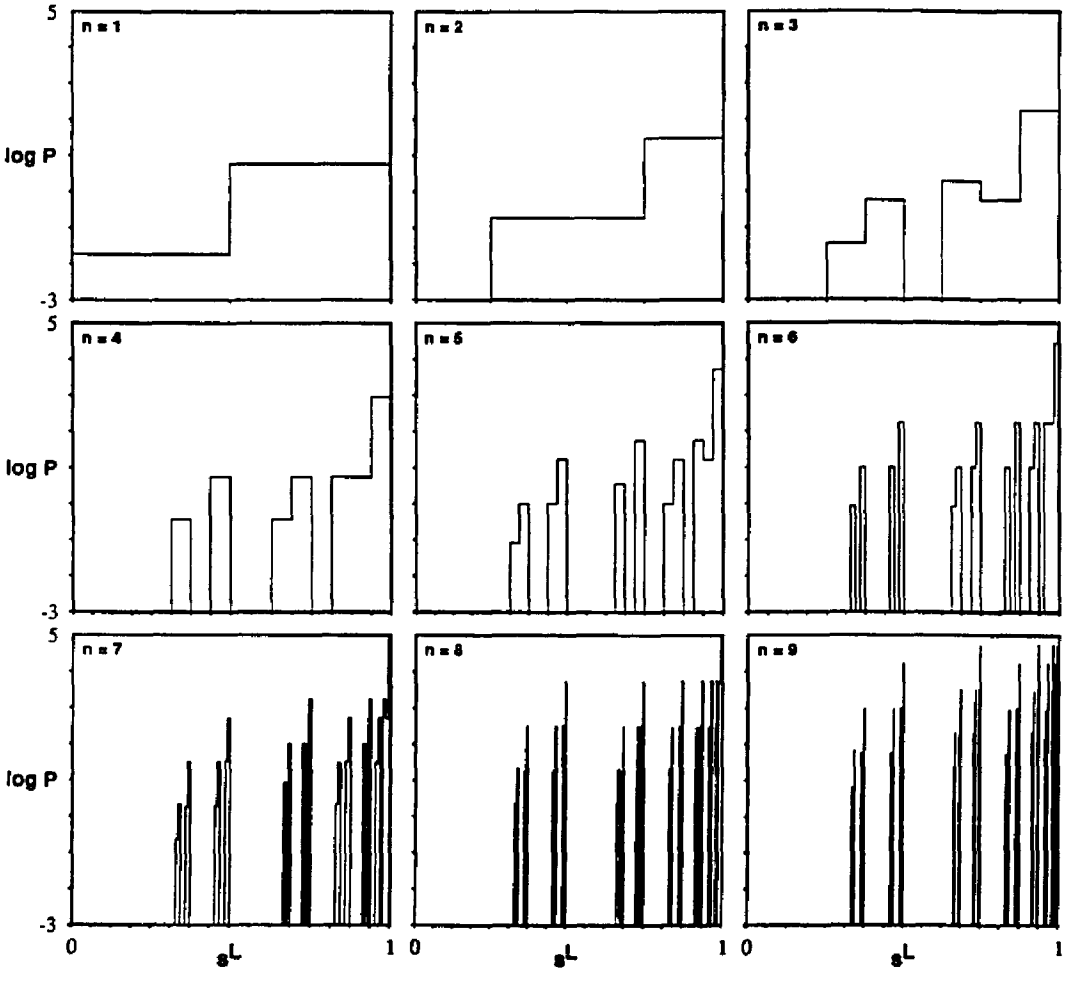


Figure 78

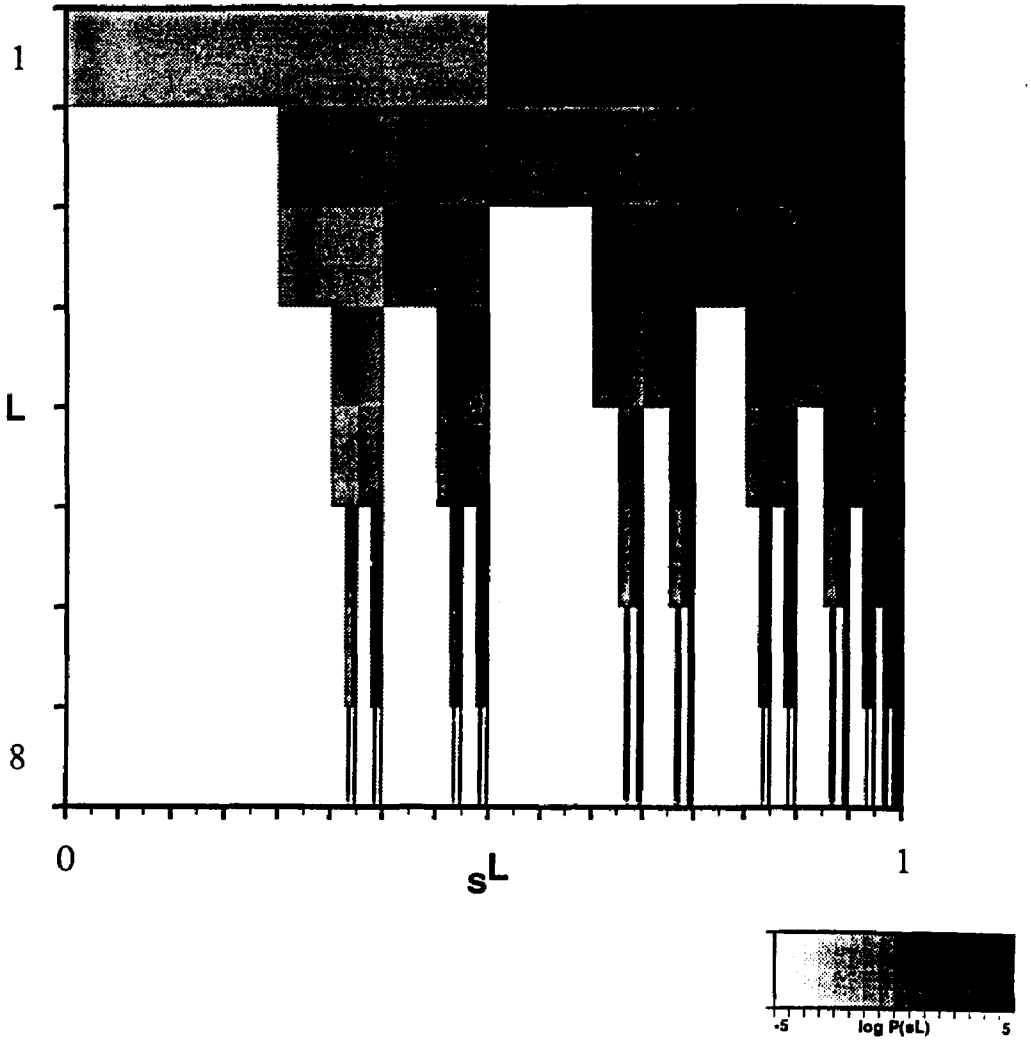


Figure 79

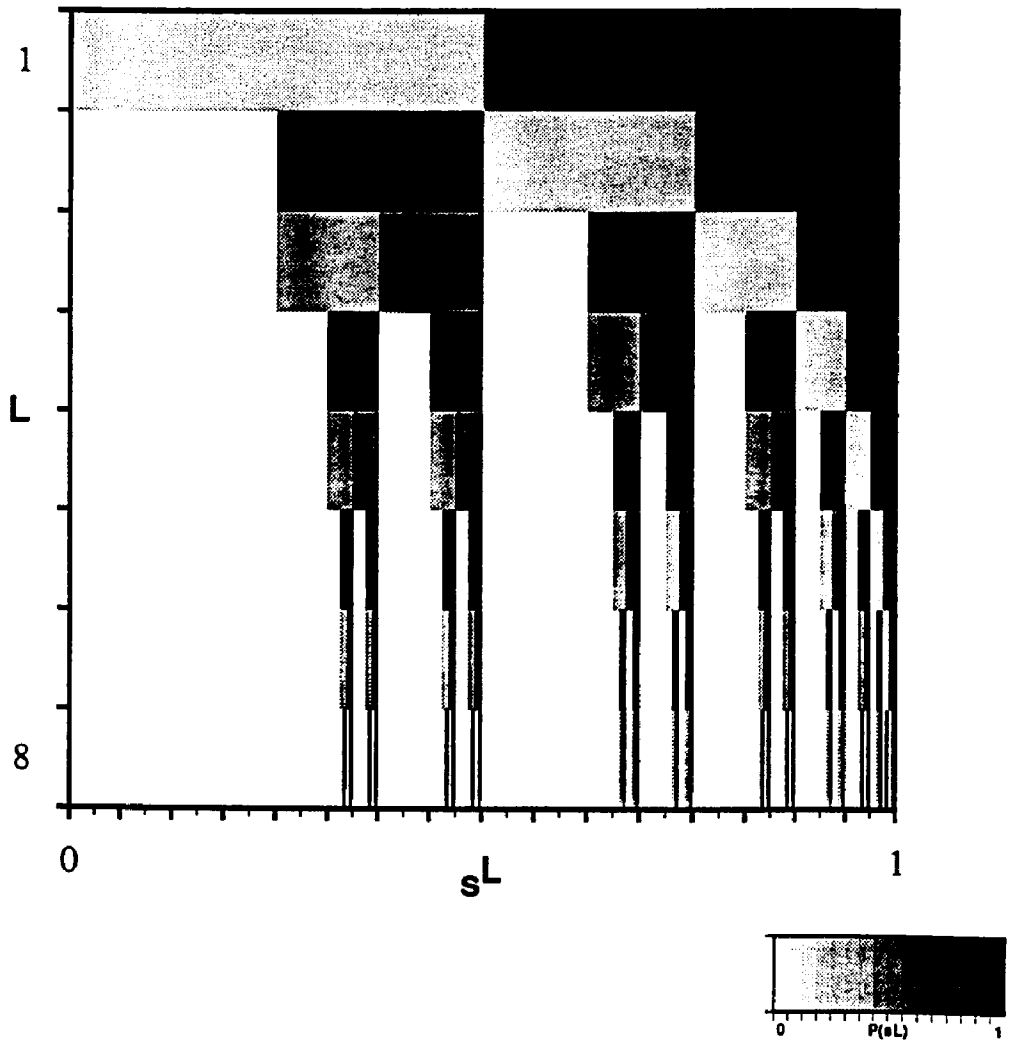


Figure 80

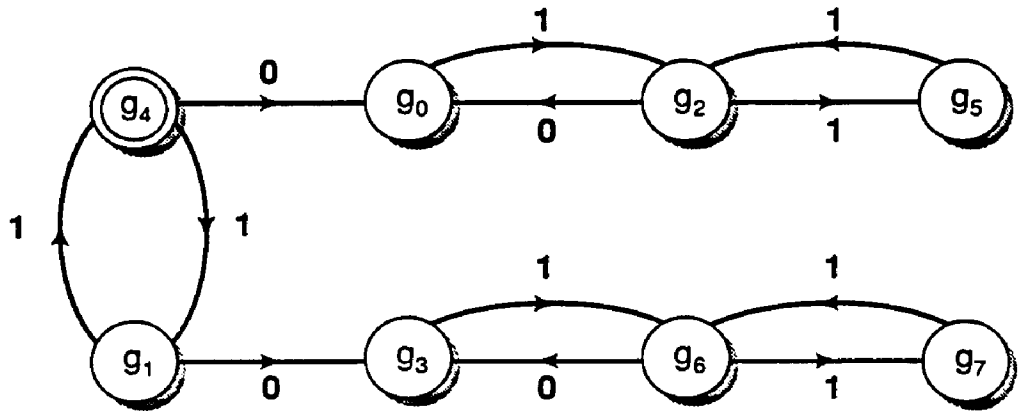


Figure 81

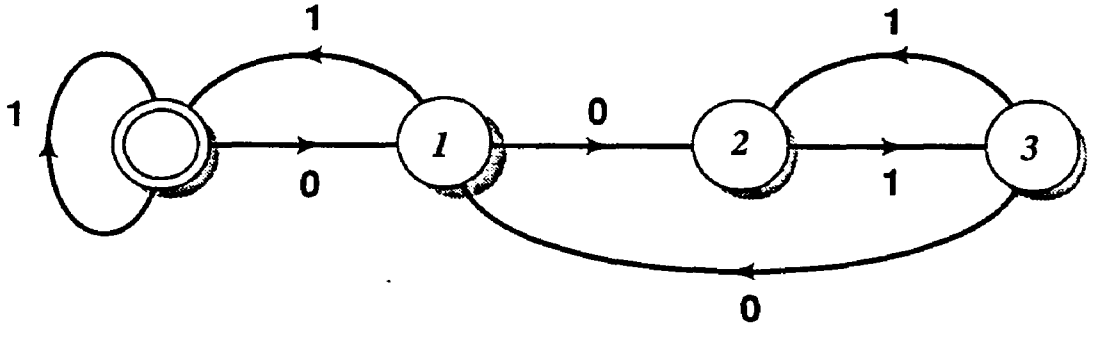


Figure 82

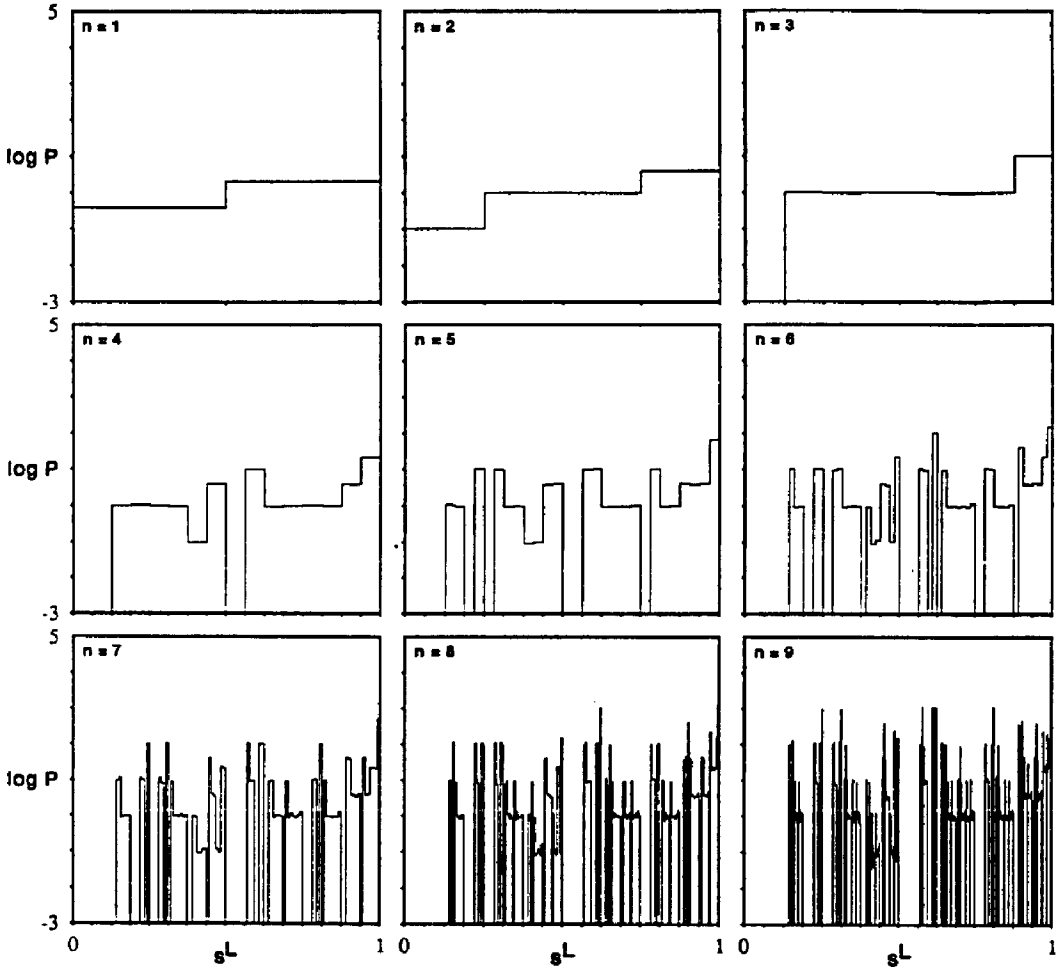


Figure 83

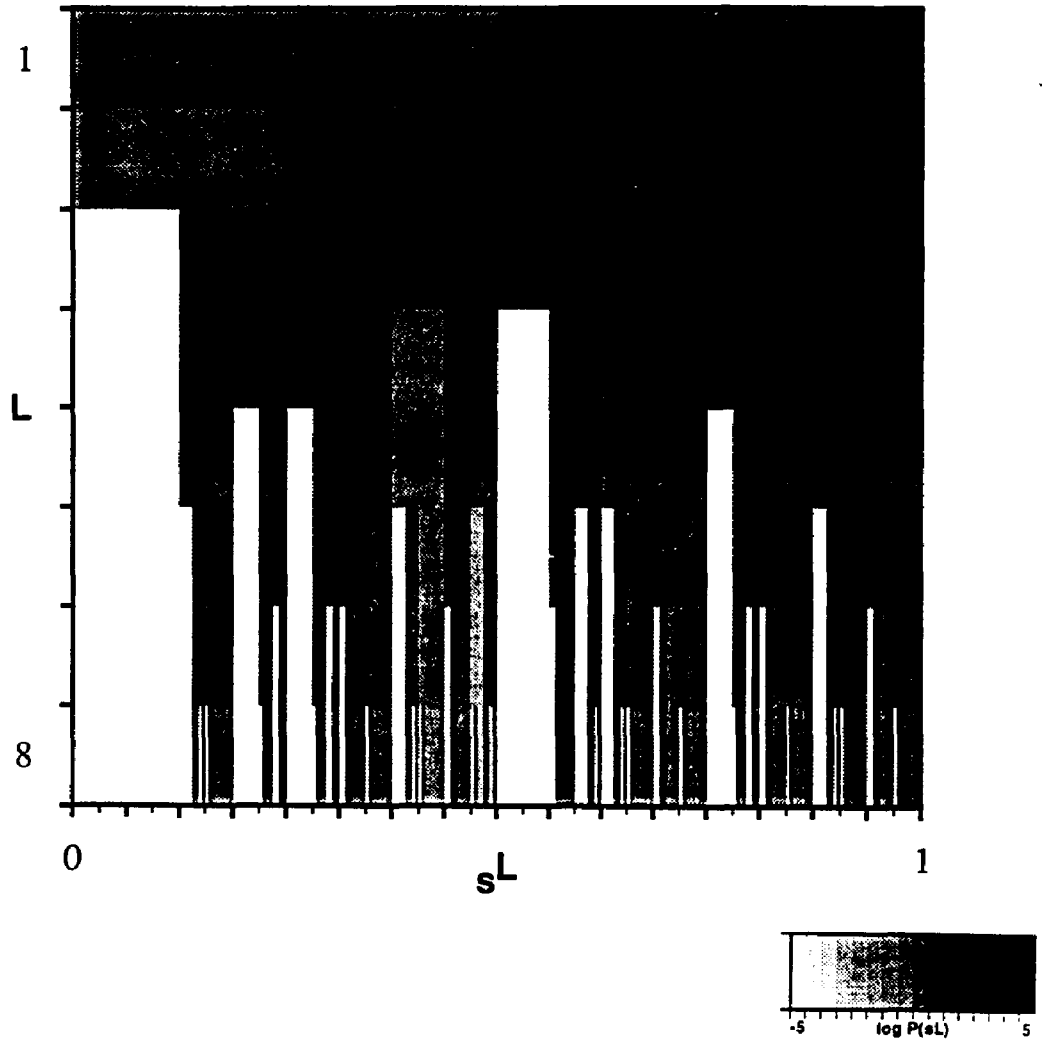


Figure 84

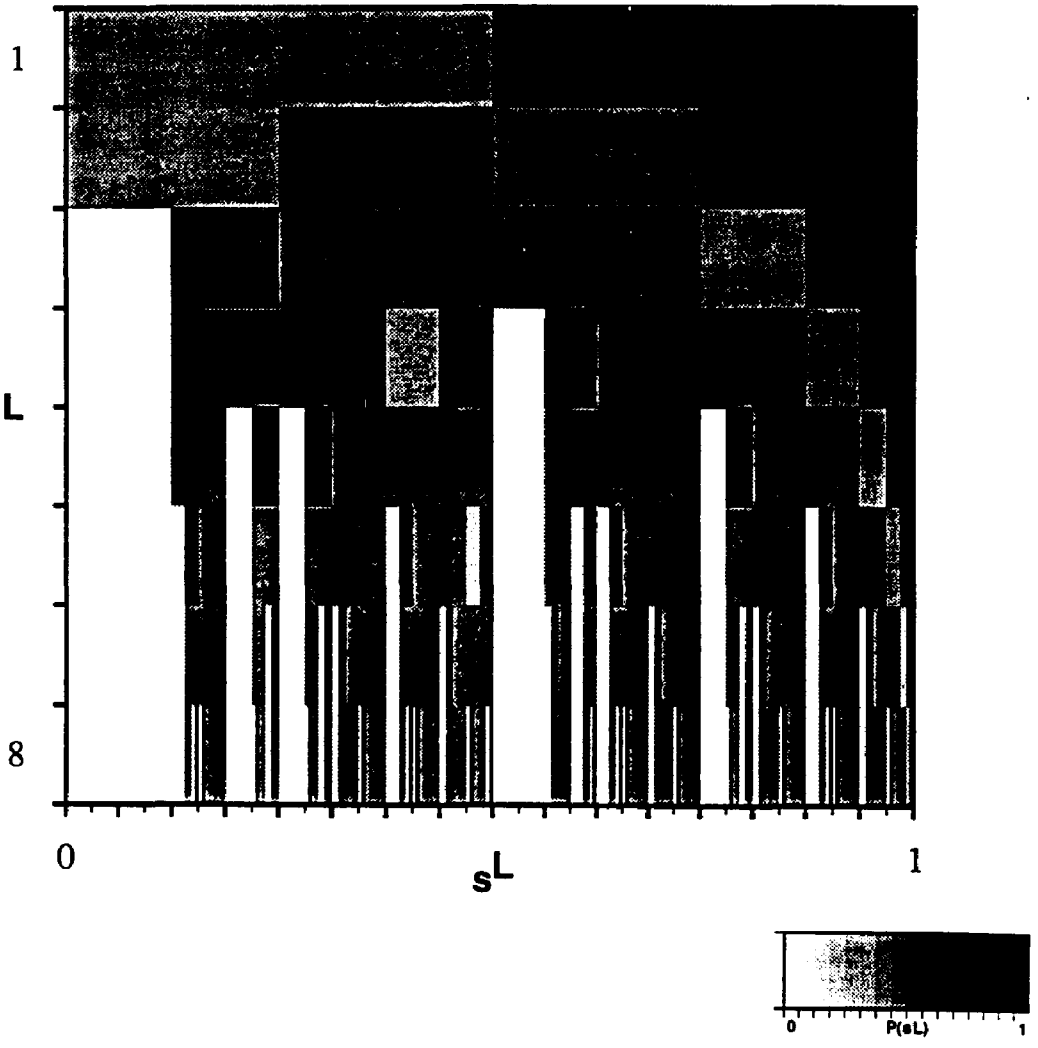


Figure 85

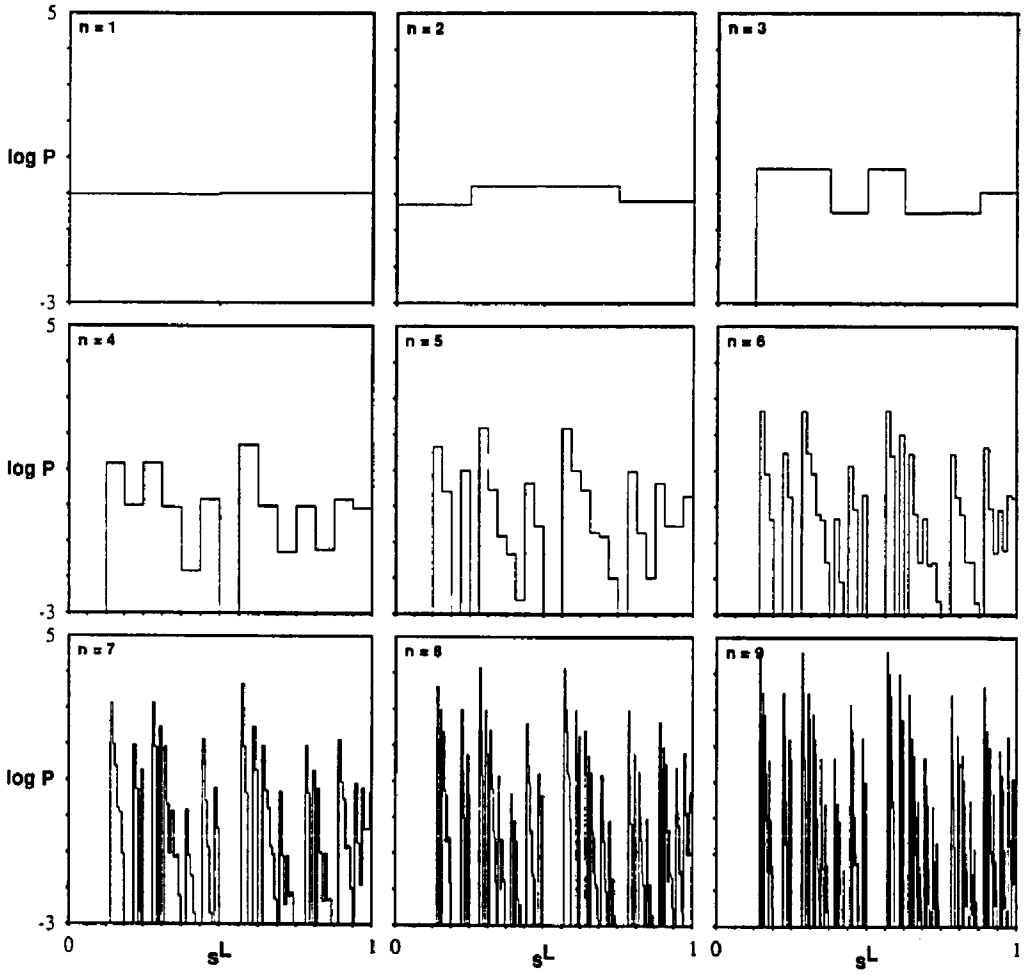


Figure 86

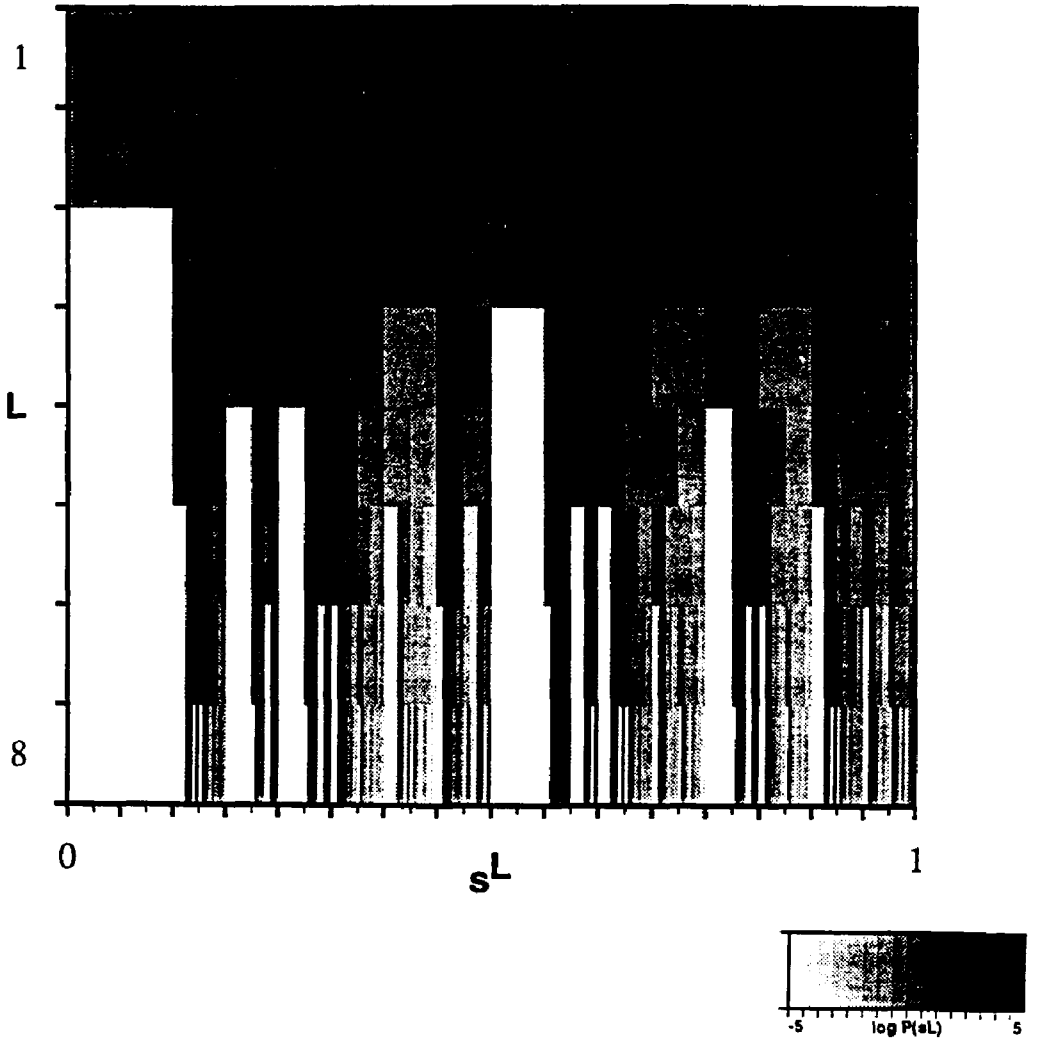


Figure 87

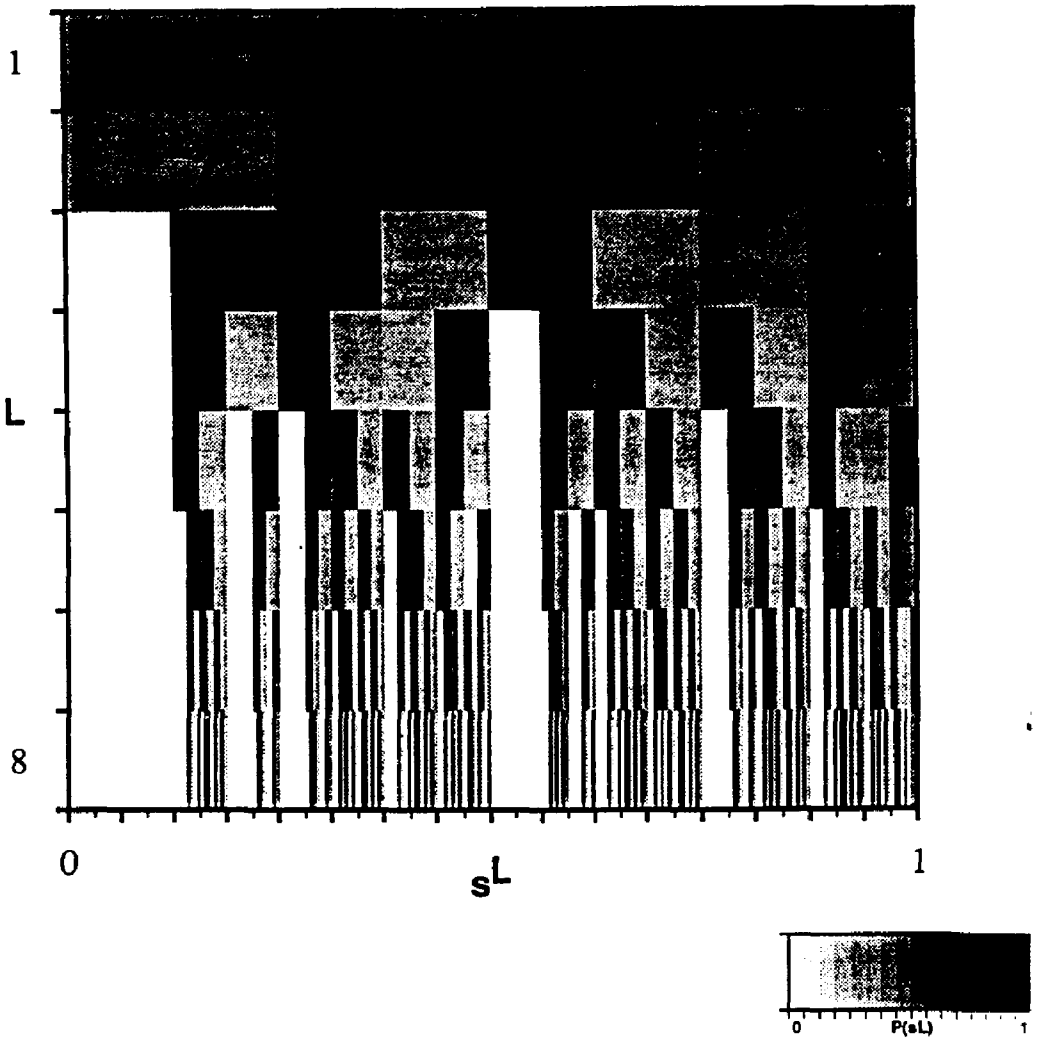


Figure 88

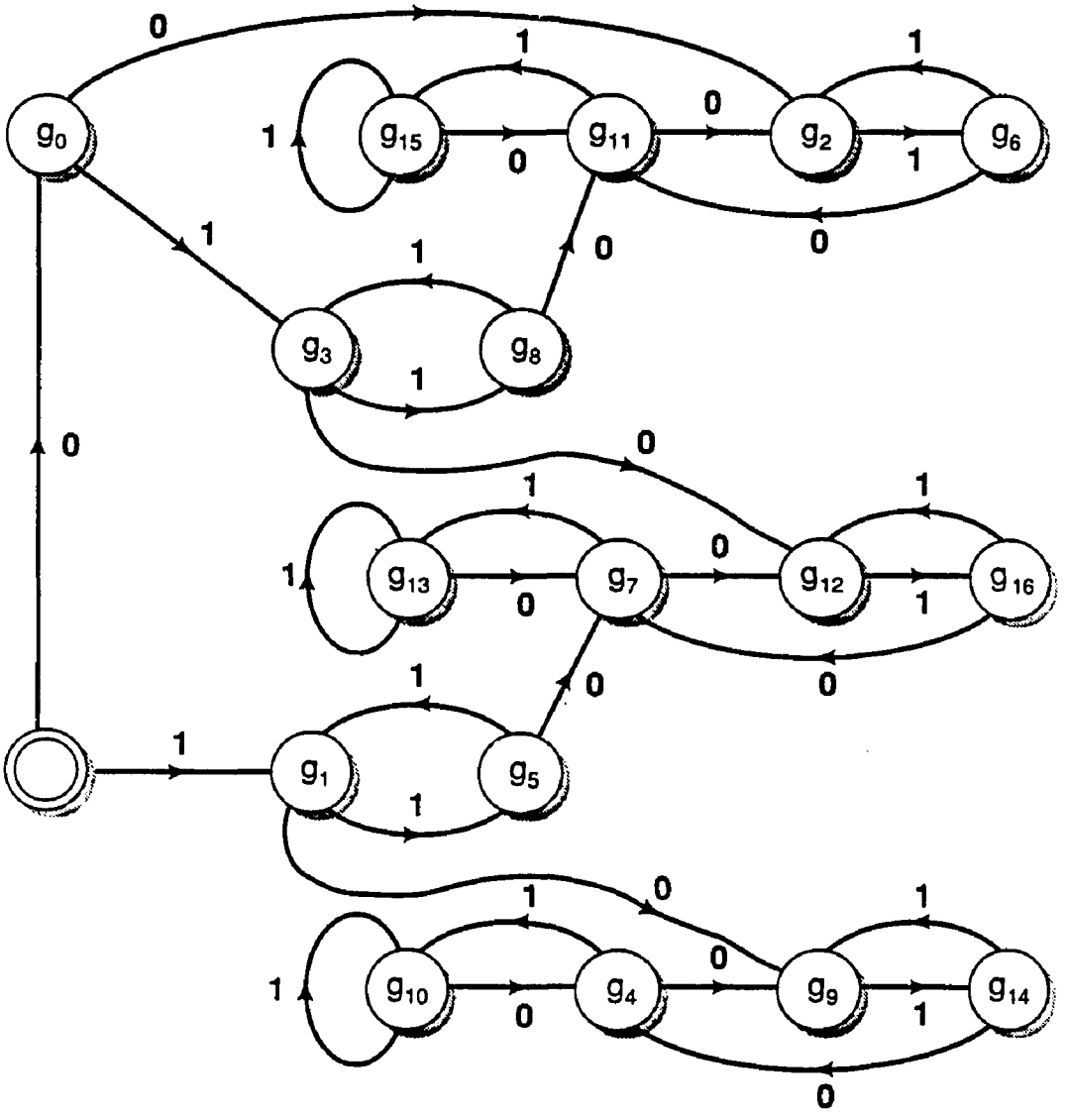


Figure 89

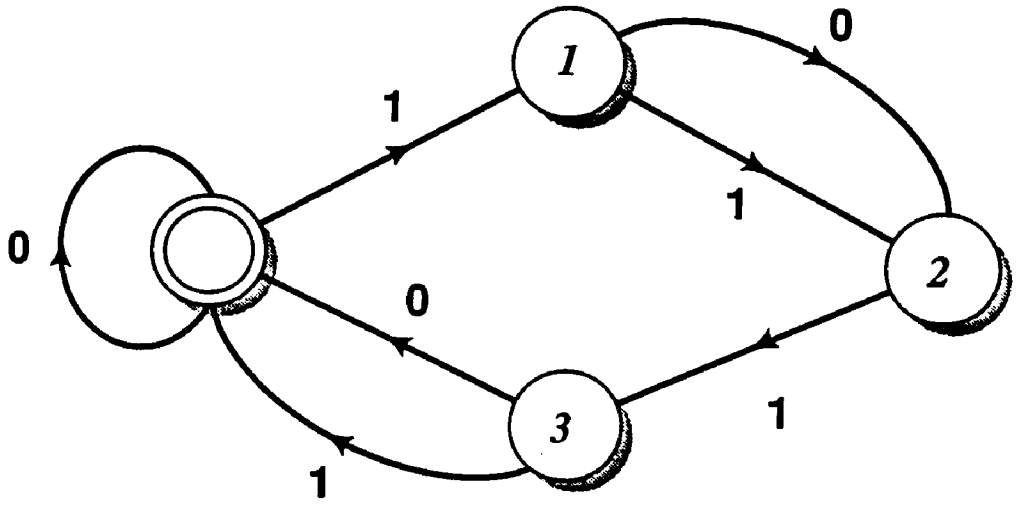


Figure 90

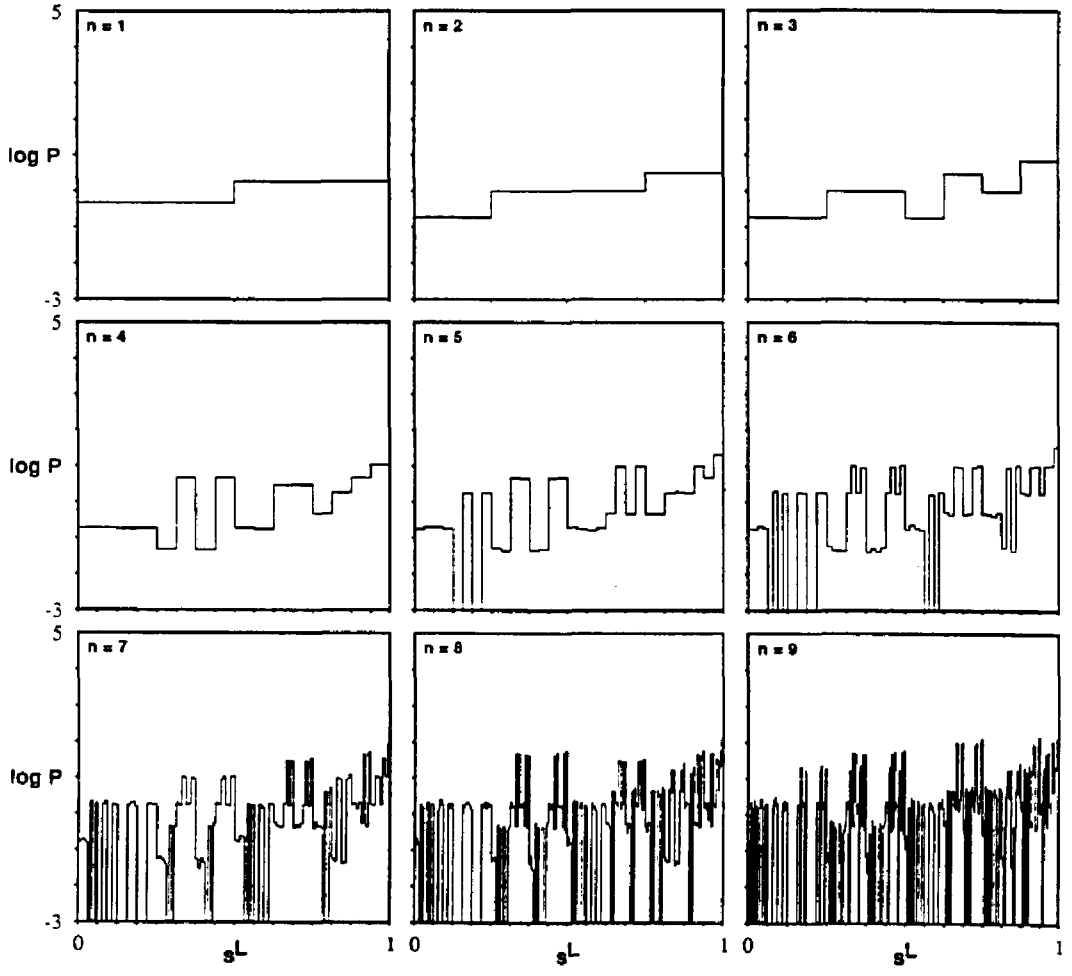


Figure 91

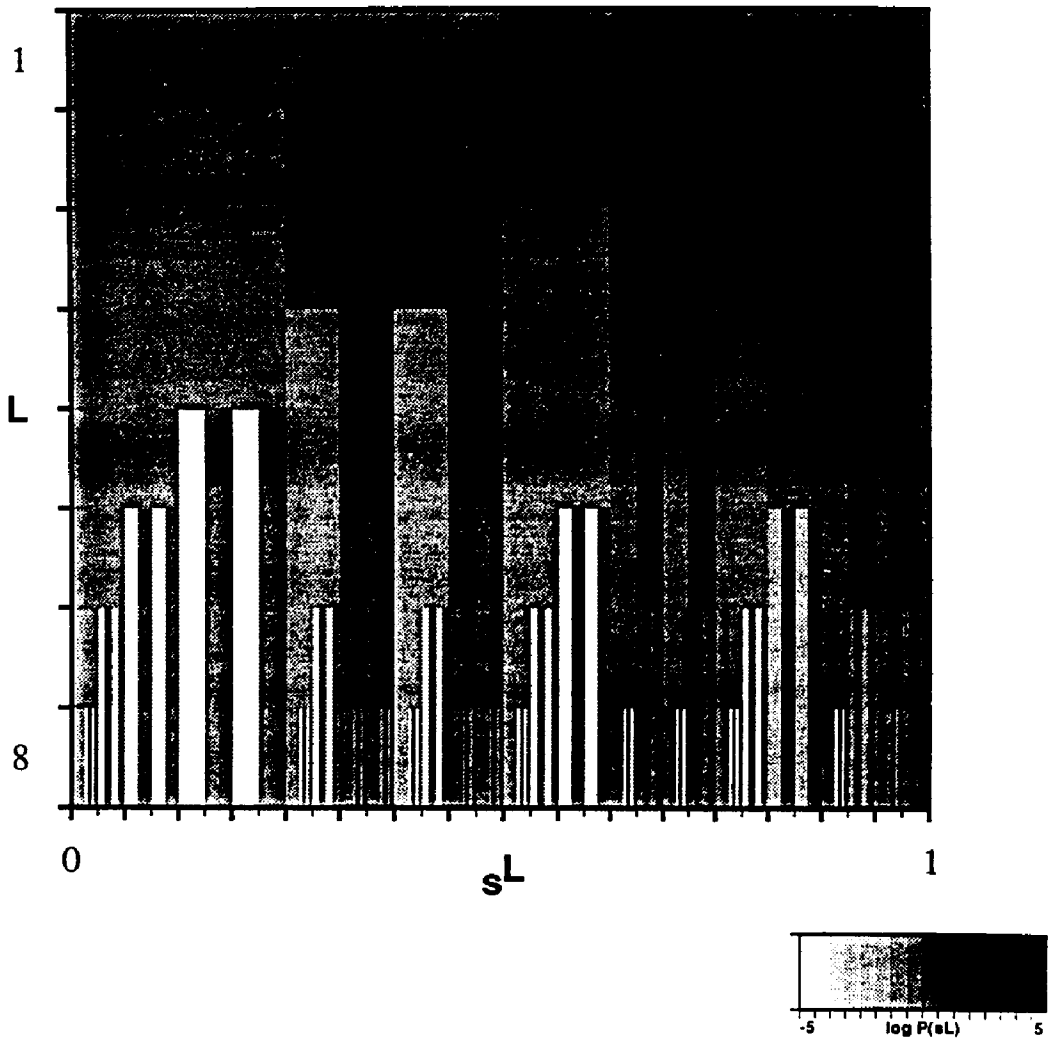


Figure 92

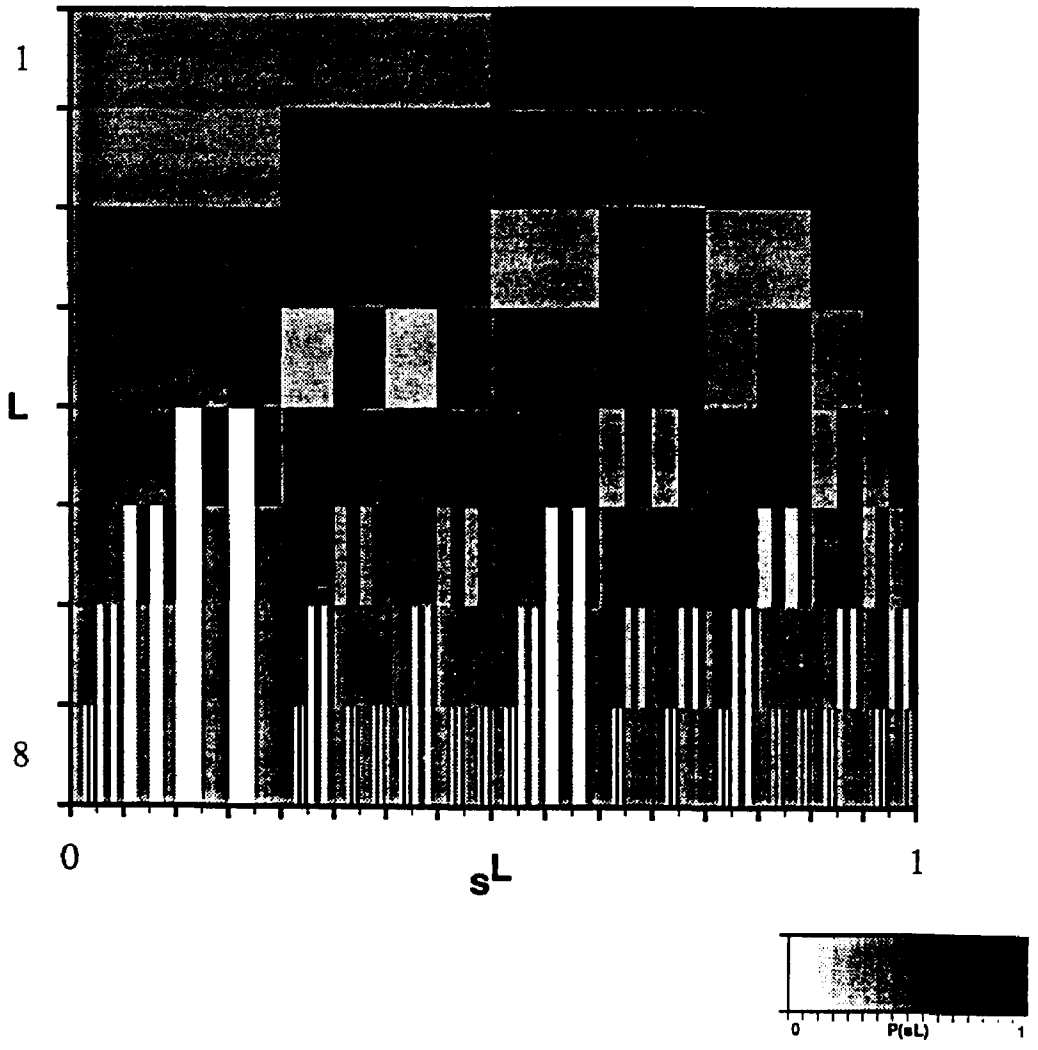


Figure 93

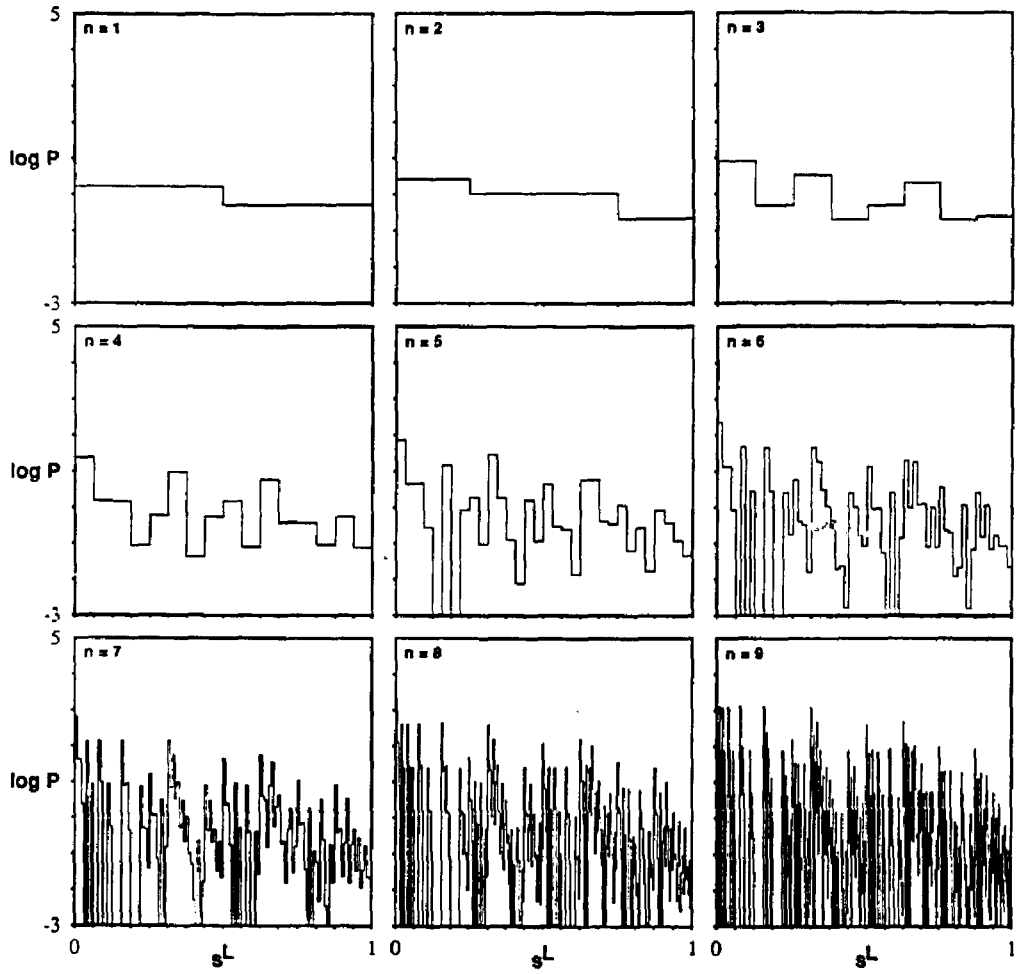


Figure 94

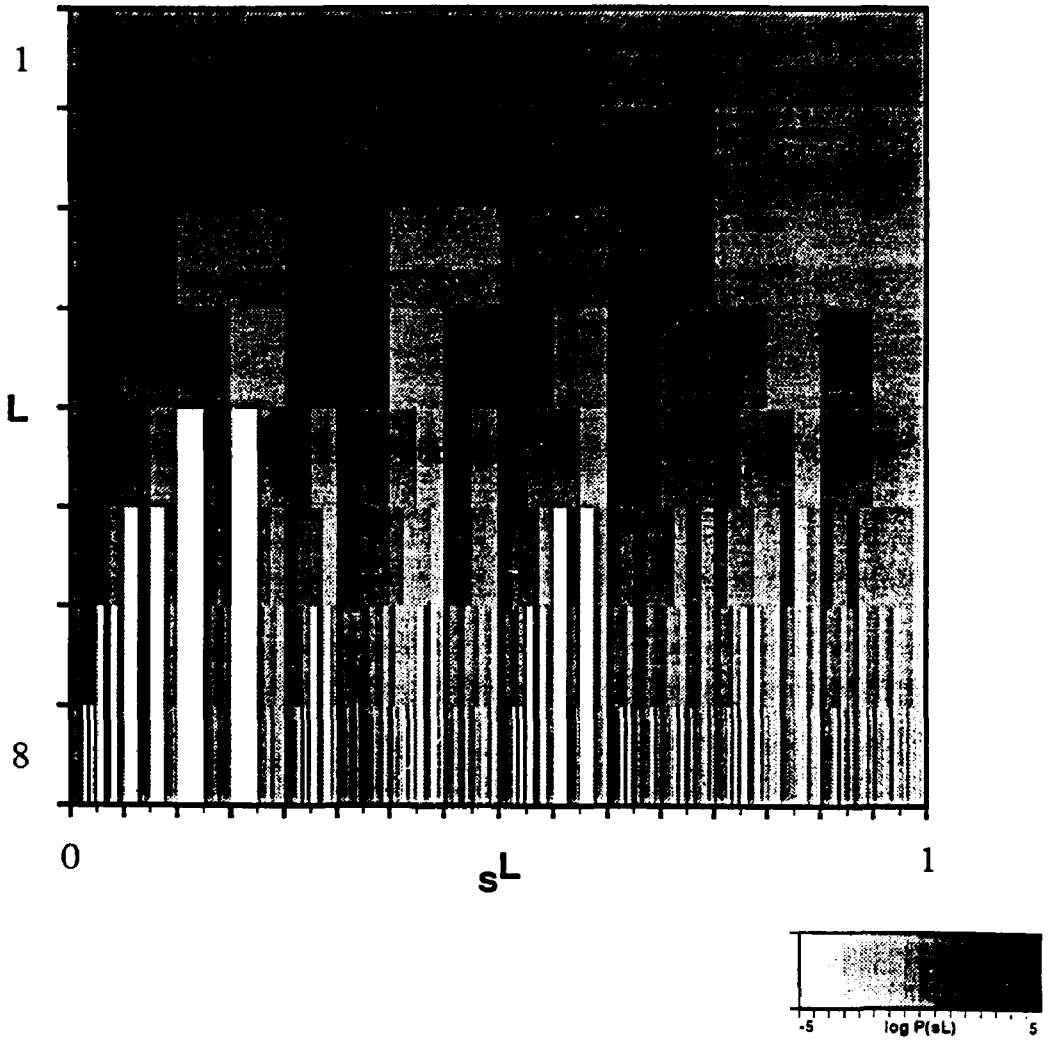


Figure 95

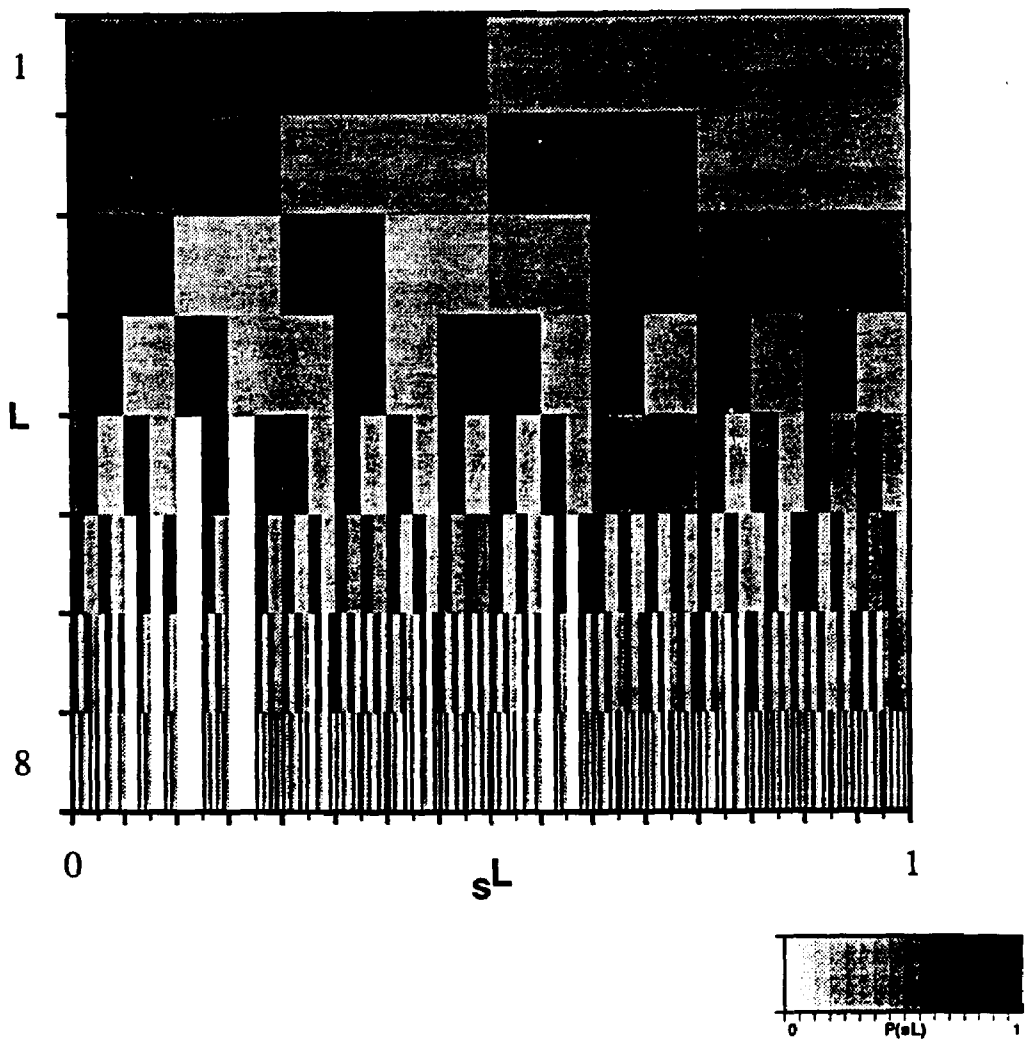


Figure 96

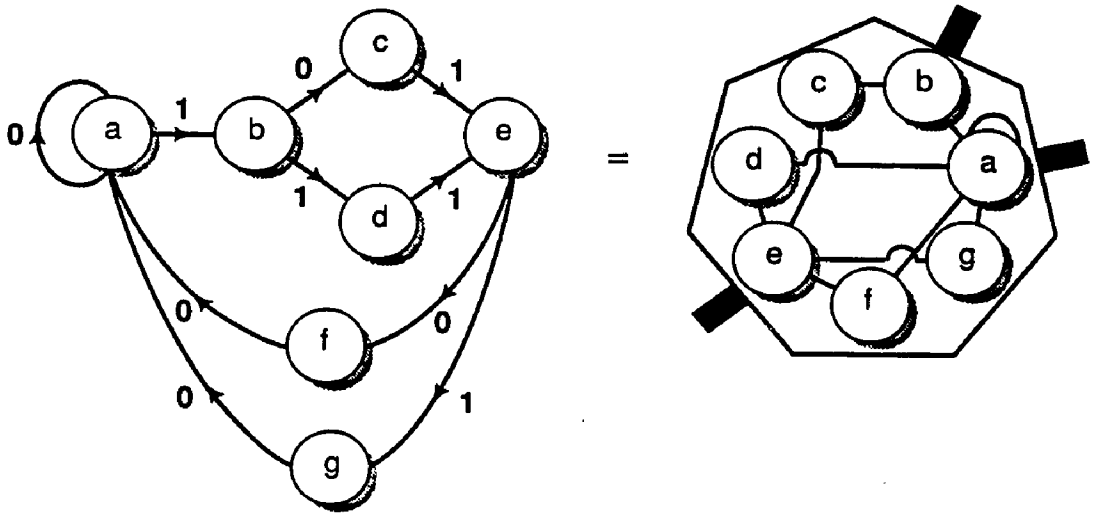


Figure 97

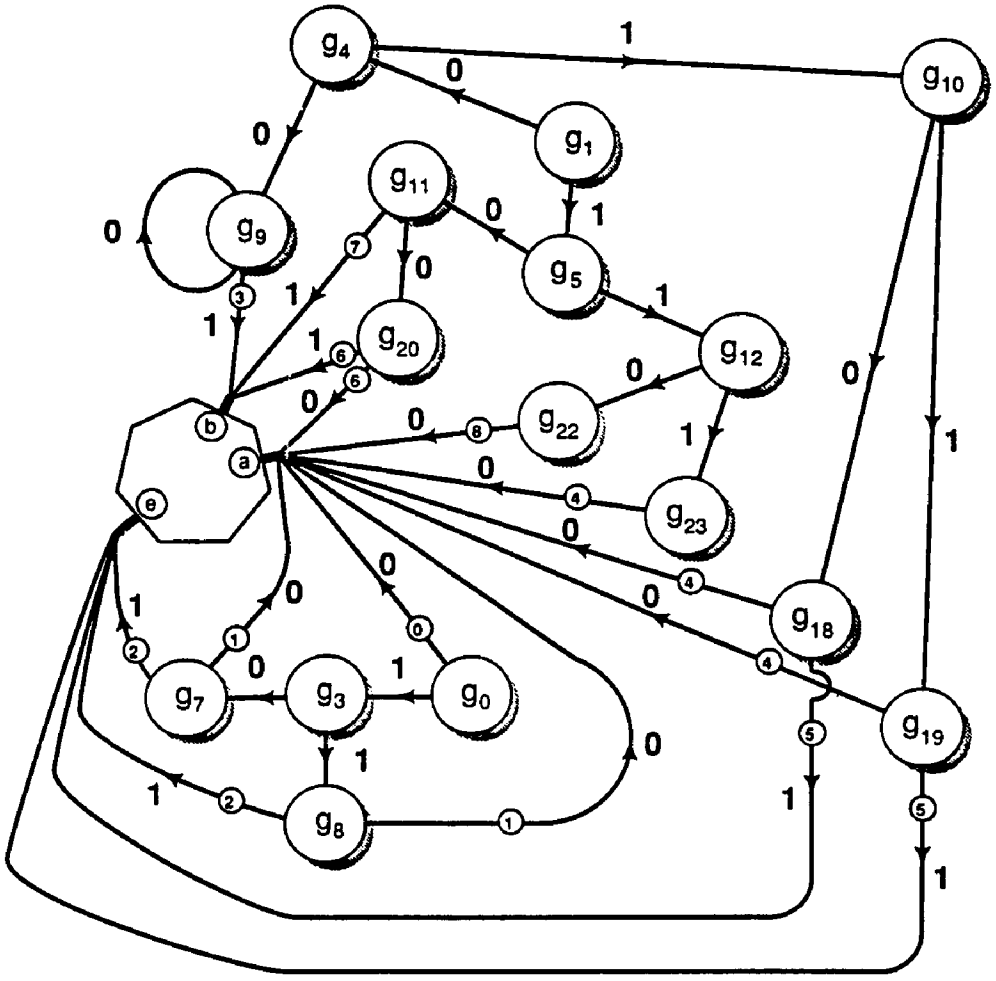


Figure 98

Bibliography

[1] C. H. Bennett. Thermodynamics of computation - a review. *Intl. J. Theo. Phys.*, 21:905, 1982.

[2] C. H. Bennett. Dissipation, information, computational complexity, and the definition of organization. In D. Pines, editor, *Emerging Syntheses in the Sciences.* Addison-Wesley, Redwood City, 1988.

[3] E. M. Coven and M. E. Paul. Finite procedures for sofic systems. *Monastsh. Math.*, 83:265, 1977.

[4] J. P. Crutchfield. Complexity: Order contra chaos. In S. Yasui and T. Yamakawa, editors, *Proceedings of the International Conference on Fuzzy Logic and Neural Networks.* World Scientific Publishers, 1990.

[5] J. P. Crutchfield and N. H. Packard. Symbolic dynamics of one-dimensional maps: Entropies, finite precision, and noise. *Intl. J. Theo. Phys.*, 21:433, 1982.

[6] J. P. Crutchfield and K. Young. Inferring statistical complexity. *Phys. Rev. Let.*, 63:105, 1989.

[7] 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.

[8] P. Grassberger. Toward a quantitative theory of self-generated complexity. *Intl. J. Theo. Phys.*, 25:907, 1986.

[9] J. E. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation.* Addison-Wesley, Reading, MA, 1979.

[10] B. Kitchens and S. Tuncel. Semi-groups and graphs. *Israel. J. Math.*, 53:231, 1986.

[11] B. Marcus. Sofic systems and encoding data. *IEEE Transactions on Information Theory*, 31:366, 1985.

[12] N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw. Geometry from a time series. *Phys. Rev. Let.*, 45:712, 1980.

[13] H. Pagels and S. Lloyd. Complexity as thermodynamic depth. *Ann. Phys.*, 188:186, 1988.

[14] W. Parry. Intrinsic markov chains. *Trans. Amer. Math. Soc.*, 112:55, 1964.

[15] R. Shaw. *The Dripping Faucet as a Model Chaotic System.* Aerial Press, Santa Cruz, California, 1984.

[16] B. Weiss. Subshifts of finite type and sofic systems. *Monastsh. Math.*, 77:462, 1973.

[17] W. H. Zurek. Thermodynamic cost of computation, algorithmic complexity, and the information metric. preprint, 1989.

PART V
ϵ-Machine Spectroscopy

Abstract[58]

The Renyi entropy and complexity spectra give one of the most complete descriptions to date of the statistical fluctuations exhibited by a process. We show how to compute them from a given finitary ϵ-machine. A comparison of early work in information theory and statistical methods for Markov chains with the contemporary approach of large deviation and ergodic theories gives an implementable notion of "typical" behavior. Typicality is perhaps the central insight of the underlying mathematics which addresses the phenomenon that the macroscopic equilibrium state need not contain all allowable states. Using machine reconstruction on a given time series along with these procedures to analyze the estimated machine, fluctuation spectra can be estimated to arbitrary accuracy, even for sequences that have not been observed or occur with very low probability. For applications the important lesson here is that one should estimate the model which properly represents one's prior before estimating derivative statistics that implicitly impose equivalent or stronger restrictions on the representation of the data source. Despite the utility and popularity of entropy and dimension spectra, they are in fact rather incomplete. Simple examples, encountered often in chaotic dynamical systems, demonstrate that they are incapable of capturing important computational properties.

[58] This Part is a preprint of a paper by J.P. Crutchfield and K. Young.

Typicality

When investigating a phenomenon via measurements of a selected observable, one is often concerned with summarizing its properties via various statistics rather than (say) direct study of the underlying mechanism. Indeed, if an easily-computed statistic gives an adequate representation, then one is strongly inclined to view the underlying phenomenon itself as simple. In this way there is a compelling, and at least intuitive, connection between an observer's notion of a phenomenon's complexity and the set of statistics necessary to describe it within a given level of approximation.

Traditionally, one of the simplest statistics employed is the average value of an observable.[59] Under appropriate restrictions on the data source, the law of large numbers gives conditions for the average to exist and so to be a useful statistic. That is, the conditions determine when the statistic converges to the mean value. And, in this case, the statistic provides a meaningful summary of an aspect of the phenomenon.[60]

The existence of an average is far from the whole story. Except for only the most trivial of phenomena, any particular measurement deviates from the mean value. Beyond the convergence of sample averages to a mean value, the question arises of how these fluctuations occur. The degree to which these fluctuations must be taken account of in the description of a process is reflected in the relative importance of higher order moments, i.e. increasing numbers of "useful" statistics, and is the subject of large deviation theory.[15] The mathematical problem there is

[59] Another, closely-related statistic is simply counting events.

[60] It must be acknowledged early on that this is actually a two-way street. The fact is that a statistic's success in capturing a large fraction of the observed behavior rather strongly colors one's view of the underlying mechanisms, for better or worse. More specifically, if for a given data set the average is found to be useful (e.g. highly predictive), then one tends to believe the underlying mechanisms embody the theorem's restrictions, such as (say) stationarity. Obviously, this need not be true. Worse, it can never be established from any finite amount of observation. Even in the face of the given logical facts of this situation, the belief in this type of truth — that derives from the reverse application of a theorem — has been found to be effective. It is the cornerstone of the semantic content of inductive inference and relates theoretical to empirical science. This is one of the most basic examples of the philosophical notion of intentionality.[13]

to understand how sample averages converge to the mean and what range of fluctuations persist asymptotically. One very practical result is the estimation of convergence rates. This can be used to determine, for example, the number of samples required to guarantee in probability a given error in a sample average. In time series analysis, the convergence rates are related to the process's memory capacity.[8]

Perhaps of more fundamental interest, though, is that large deviation theory gives a framework that establishes the notion of "typical" behavior. Consider the example of the sequences of heads and tails generated by a fair coin. A periodic sequence HTHTHTHTHTHTHTHT is less typical than an (apparently) random or partially random one TTHHHTHTHHHTHTTT. Nonetheless they both exhibit a frequency of heads of 1/2. By "typical" then we do not refer to their likelihood of occurrence, since any two sequences of the same length from a fair coin have, by definition, the same probability and may share the same mean. To capture "typicality" one must look at a different property of sequences than direct (averaged count) probability. We do this by looking at the convergence of subword frequencies; that is, the probability of H, T, HH, HT, TH, TT, and so on. If a long sequence has the most uniform population[61] of length L subwords then we call it a typical sequence for the fair coin.

Large deviation theory is not the first articulation of typicality. Boltzmann's and then Gibbs' development of statistical mechanics are studies of just such questions: the existence of a macroscopic average for a given set of microscopic states. The convergence properties, of macroscopic observables, as a function of, for example, system volume and particle number are closely related to the free energies and to thermodynamic entropy. Indeed, the thermodynamics of a macroscopic system is a compact description of the structure of fluctuations that the microscopic system supports and how that structure changes under external and internal conditions.

The concept of an equilibrium state captures the notion of typicality. Those microscopic states that are associated with the equilibrium macroscopic state are typical.

---

[61] By "most uniform" we mean relative to the given population of sequences.

Perhaps the most succinct study of typicality is found in communication theory. It is not too radical a simplification to say that it was Shannon's main insight that only the typical messages need be coded for efficient transmission through a communication channel; or conversely, that the atypical messages can be ignored. The typical-atypical dichotomy is summarized in the Shannon-McMilllan theorem.[3, 21] In essence it says that typical sequences from an information source have maximal entropy rate. One consequence is that they overwhelm the atypical sequences in number. Another is that the relative probability of atypical sequences decays exponentially with increasing length.

Typical behavior can be detected once one clearly states the modeling assumptions. More specifically, once one estimates a particular model within some chosen class then, with respect to that model, typical behavior can be identified with certain properties, such as subword convergence behavior. The following is an investigation of exactly these questions, but in the domain of modeling nonlinear dynamical systems with stochastic automata. In this way, we hope to give concrete form to the preceding notions in the context of time series analysis. Through this development we intend to show the utility of a thermodynamic analysis in understanding the gross features of the microscopic combinatorial complexity of a process. This approach also shows rather clearly its inadequacies in capturing important computational properties.

The main representation, or model class, we shall use is stochastic automata to approximate the probability distribution over subsequences. But first we must establish exactly what elements of probability distributions we wish to approximately model.

Preliminaries

We begin by reviewing some basic definitions. An abstract dynamical system is a measure preserving transformation T of an abstract measure space $(X, \mathcal{B}, \mu)$ where X is a sample space, $\mathcal{B}$ is a σ-algebra of subsets of X and constitutes the set of elementary or atomic events, and μ, the measure, is a normalized set function defined on $\mathcal{B}$, i.e.

$$\text{supp}\mu = \mathcal{B}$$

As an example consider the logistic map. The underlying sample space is the unit interval, i.e. $X = [0, 1]$. The σ-algebra of subsets of X is essentially the set of open subintervals of X. The transformation T is the function $f(x) = rx(1 - x)$ on the unit interval. There are any number of measures μ we could consider, such as the "physical" or BRS measure, [14] the only criterion being that the measure is preserved by T.

We will assume that we have a generating partition P defined on the sample space X, such that there is a one-to-one mapping between sample paths generated by T on P and the events $\mathcal{B}$. For the example of the logistic map a generating partition is given by

$$P = \{[0.0, 0.5), (0.5, 1.0]\}$$

[26] The probability of an event ϵ, i.e. some set of elementary events, is given by

$$Pr(\epsilon) = \int_{\epsilon} d\mu(x)$$

We will consider a labeling of the finite set of elements of P, with a set of symbols we refer to as an alphabet A. We denote the set of infinite sequences of symbols from A as A^{∞}. A data stream s is a string of alphabet symbols produced by the action of the abstract dynamical system T. In what follows we will consider the that the system producing the data stream is stationary, i.e. is not itself dependent on time. A word ω is also string of alphabet symbols

whose length we denote by $|\omega|$. We will usually think of data streams s as composed of sets of words ω. We denote the size of a set S as $||S|| =$ card S.

We will consider the scaling behavior of probability distributions of sequences generated by finitary ϵ-machines introduced previously.[11] These stochastic machines constitute the assumed model class mentioned above in the context of Bayesian parameter estimation. Strictly speaking, the restriction to finitary ϵ-machines determines the support of the modeling prior. These systems are more general than those describable in terms of multinomial scaling, i.e. independently distributed stochastic processes, as discussed in a slightly different context in [2]. And they are more general than finite-state Markov chains.[9, 10] In simple terms, finitary ϵ-machines describe systems with either finite correlation length, i.e. with finite memory, or an extremely limited form of infinite memory that is restricted to orbits with less than full measure.

An important distinction has become apparent recently in the study of complex dynamical systems. This is the distinction between the scaling structure of the support of the distribution, suppμ, and the scaling structure of the distribution itself. At finer and finer scales not only is the structure of the support very complicated, e.g. infinitely-many Cantor sets, but also the way in which probability is distributed over the atomic events. In the following we will give simple examples of processes which exhibit such structure.

We find it useful to consider the scaling behavior of the support and distribution independently. This is accomplished by first considering the scaling properties of the distribution on a uniform support, which constitutes the content of this paper. The approach mirrors that of ϵ-machine reconstruction [11] in which the topological machine describing the observed sequences is built first and then the "μ-machine" describing the structure in the distribution is built over the encoded output of the ϵ-machine. We shall consider, in a sequel to the present paper, the joint scaling structure of the distribution and its support and explore their mutual convergence properties in a form consistent with the present work but slightly different from the approach taken in the recent literature.[20] We will not consider the effects of variation of the

approximation parameters in ϵ-machine reconstruction, i.e. ϵ, the measurement resolution, L, the morph depth, nor N the parse tree depth. Their variation will be addressed elsewhere.

Cylinders

We turn to the study of that scaling structure in distributions $\mu(\omega)$ over infinite measurement sequences ω which is independent of the support's structure. The methods and results used in this section are similar to those used by probabilists in the theory of large deviations.[15] Here we consider a distribution over the possible paths of the stochastic process associated with a finitary process. The elements of the distribution's support supp μ in this case are the process' sample paths $\{w\}$. Restrictions in the range of allowed sequences, e.g. if given measurement events do not occur consecutively, lead to scaling structure in the set of observed sequences and so in the support. The development of Cantor set structure in the support and in the distribution can be seen in the figures of [7]. The following analysis, however, ignores the support's scaling structure by assuming that the sample paths form a set of identical intervals over which the distribution is defined. That is, we will focus on the variation in sequence probability amplitude, i.e. "fluctuations".

In ergodic theory parlance the set of infinite sequences which share a particular length L subsequence $s^L = s_0 s_1 \ldots s_{L-1}, s_i \in \mathbf{A}$ is called an L-cylinder

$$s^L = \{\omega : \omega_0 = s_0, \ldots, \omega_{L-1} = s_{L-1}; s_i \in \mathbf{A}, \omega \in \mathbf{A}^\infty\}$$

Formally the cylinder measure is

$$\mu(s^L) = \frac{N(s^L)}{\|\mathcal{S}^L\|}$$

where $N(s^L) = \|s^L\|$ and $\mathcal{S}^L = \bigcup_{s^L \in \mathbf{A}^L} s^L$ consists of all of the L-cylinders. This definition is formal because for general stochastic processes $N(s^L)$ and $\|\mathcal{S}^L\|$ are infinite although the ratio $\mu(s^L)$ is clearly finite and less than or equal to one. Then the number of distinct L-cylinders observed, i.e. the number of elements s^L that partition the set $\mathcal{S}^L$, is

$$N(L) = \|\{\omega : \omega \in \mathbf{A}^L, \mu(\omega) > 0\}\|$$

For a finite length data stream $\mathbf{s} = s_0 s_1 \ldots s_{k-1}$, the above definitions must be modified. In particular, we associate the occurrences of a given cylinder with the indices within the data stream at which it occurs. In analogy with the cylinder measure, the natural estimator for a word's probability is then

$$Pr(\omega) \approx \frac{N(\omega)}{\|\mathcal{S}^L\|}, \quad |\omega| = L \leq k$$

where $N(\omega)$ is the finite count of ω's appearance in $\mathbf{s}$ and $\|\mathcal{S}^L\| = k - L$ is the total number of L-cylinders. The estimate's accuracy clearly depends on the length of the data stream and the nature of the source.

When it simplifies presentation in the following, each finite measurement sequence of length L, independent of when it is observed, will be referred to as an L-cylinder. Thus, we conflate the sequence s^L with the ergodic theory L-cylinder. We will, hence, denote the subword s^L as distinct from the set which it indexes within $\mathcal{S}^L$.

If we lexicographically order the observed L-subwords $\{s^L\}$, then in order to uniquely identify a single element it requires

$$H_0(L) = \ln N(L)$$

bits. This is the length L total topological (Hartley) entropy of the process. The specific topological entropy, or simply the topological entropy, is the exponential growth rate of the total

$$h = \lim_{L \to \infty} \frac{H_0}{L}$$

If we consider the underlying process as a communication channel that is transmitting sequences from some ideal random source by "filtering" that randomness through its internal structure, then h is called its channel capacity.[25]

To conclude this section we introduce cylinder histogram plots for the example systems with which we will work in this paper. This gives a visual impression of the variety of behaviors

in sequences observed for various standard processes. Figure 99 shows a histogram plot of $\log_2 Pr(s^L)$ vs. $s^L \in [0,1]$ for a data stream of length $k = 10^7$ obtained from a simulation of the flipping of a biased coin with $Pr(s = 0) = .4$ and $Pr(s = 1) = .6$ where $s = 0$ represents tails and $s = 1$ represents heads. The binary strings, s^L, thus obtained for various lengths L are treated as binary fractions in the interval $s^L \in [0,1]$ for the purposes of the histogram plot, and where we have divided the unit interval into 2^L bins. Thus, we use the mapping $\rho : s^L \rightarrow \Re$,

$$s^L \rightarrow \sum_{i=1}^{L} (s^L)_i \, 2^i$$

which makes no assumption about embedding dimension for individual measurements. We show the cylinder histograms for sequences of length $L = 1, 2, ..., 9$.

In figure 100, we show the same set of cylinder histogram plots for a binary data stream of length $k = 10^7$ generated by simulating the dynamics of the logistic map at the Misiurewicz parameter value $r = r_M \approx 3.927737...$ on a binary partition with the partition divider at the critical point of the map.[62]

Figure 101 shows the same set of cylinder histogram plots for a binary data stream of length $k = 10^7$ generated by simulating a system known as the golden mean system whose properties will be described more fully below. The defining property of the golden mean system is that it generates sequences containing only isolated 0's, i.e. ...00... does not occur. The name derives from the fact that the topological entropy for this system is equal to the logarithm of the golden mean: $h = \ln \frac{1+\sqrt{5}}{2}$.

Finally figure 102 shows the same set of cylinder histogram plots for a binary data stream of length $k = 10^7$ this time generated by simulating a system known as the even system whose properties also will be described more fully below. The defining property of the even system is that it generates sequences containing only even runs of 1's between 0's. It can be thought of as a parity recognizing or generating machine.

[62] The Misiurewicz parameter values are those for which band mergings occur in the logistic map $f(x,r) = rx(1-x)$. In the current case we are considering 5 bands merging to 4. This parameter value is formally defined as the root, r, of the equation $f^4(r, x_c) = f^5(r, x_c)$ with $x_c = \frac{1}{2}$.

Starting with the simplest case we see that the fluctuations for the biased coin histogram develop in a simple way as a function of increased cylinder length. For an unbiased coin all sequences would have the same probability but this symmetry is broken by the bias of the coin. The level sets in the histogram correspond to the sets of cylinders having the same ratio of 0's to 1's since this ratio uniquely specifies the probability. The number of cylinders with a particular value of this ratio is given simply by the corresponding binomial coefficient. For instance, since $Pr(s = 1) = .6$, the cylinder with the highest probability for any L is the cylinder consisting of L consecutive 1's. The corresponding binomial coefficient is 1 as can be seen in the histograms, i.e. there is only one cylinder having the maximum probability for any L. For the golden mean system, despite the fact that the support of the distribution does not have full measure as in the case of the biased coin, the structure of the level set distribution again appears to develop in a relatively simple manner. This structure, which is slightly harder to characterize, as we will demonstrate subsequently, is still clearly discernible from the histogram plots. The even system, while still more complicated in support structure and level set distribution, nonetheless still exhibits a discernible self similarity. Finally, the Misiurewicz system exhibits no clearly discernible level set structure in the histograms up to cylinder length $L = 9$. A more careful analysis of the fluctuation spectra of these examples, to follow, will demonstrate these differences clearly.

Scaling in Sequence Distributions

With these examples of complicated sequences and probability distributions in mind, we are ready to investigate the fluctuation spectra given a distribution over sequences. The goal is to glean useful "macroscopic" properties from the population of microscopic sequences. The following considers not only finite L, but also seeks a description of the complete measure found as $L \to \infty$. We develop the thermodynamics of sequence distributions in three steps. The first is a description in terms only of extensive parameters; i.e. total energy, total entropy, cylinder length, and so on. By extensive we simply mean quantities which are a function of cylinder length, analogous to quantities which are a function of volume in standard thermodynamics. We are not concerned in this case with the scaling structure of the distribution, only with the total quantities defined at a particular length L. We shall see in the examples below that certain thermodynamic relationships hold for the extensive quantities defined in this section.

The asymptotic growth rates, which are the quantities of interest in the thermodynamic limit, of the extensive parameters are the basis of the intensive description. This is presented third, since we interpolate another thermodynamic description in which the quantities are called *transtensive*. They are the finite-size versions of the specific thermodynamic parameters, i.e. densities or rates , and correspond to "bare" quantities in renormalization group language. Within each thermodynamics, Legendre transforms introduce intensive parameters that play the same role as those in the thermodynamic limit. This three stage development shows that the intensive parameters obtained via the thermodynamic limit of the transtensive parameters and those obtained via Legendre transforms need not be, but often are, the same.

One of the primary reasons for the use of thermodynamic formalism is that it provides a language for the discussion of notions such as typicality and fluctuation in the context of inferring the structure of a dynamical system. Similar points have been discussed in [24], [20], and

[17] from other perspectives. We will therefore note the connections between thermodynamic descriptions and the problems of inference throughout the following discussion.

Extensive Thermodynamics

Note that a particular sequence can be viewed as the instantaneous configuration of a spin system. The symbols interact just as spins would, except that we specify the strength of interaction via a joint probability $Pr(s^L)$ over all the symbols comprising s^L. The total energy of a sequence is then determined by its probability as follows

$$\mathbf{U}_{s^L} = -J \log_2 Pr(s^L)$$

where J is a constant that sets the scale of the average bond strength in units of energy per bit. In this way, unlikely spin configurations have high energy and likely ones have low energy. This is, in fact, much the same way that Gibbs first characterized the relationship between the energy and probability of a state.[16]

In what follows we wish to model the stochastic nature of our process (which generates a particular cylinder distribution) as an interaction of the system under study with its environment. Hence J appears in the product

$$\beta = \beta' J$$

where β' defines the unit of inverse temperature in units of energy, i.e.

$$\beta' = \frac{1}{kT}$$

. At this point we set J equal to unity and in the following work with the scaled energy $\mathbf{U}_{s^L}/J$ and the dimensionless quantity β, which is a measure of the ratio of the internal energy of the system to the ambient thermal energy. Since we will be working with systems having a finite spectrum, i.e. a finite maximum energy, the interpretation of negative values of β will be much the same as in the case of a finite spin system in a state with a population inversion.[18]

The distribution of energy over the "microstates" s^L is summarized by grouping them into energy macrostates, or energy level sets,

$$\mathcal{S}_{\mathbf{U}}^L = \{\omega : \mathbf{U}_\omega = \mathbf{U}, \omega \in \mathcal{S}^L\}$$

where $\mathbf{U}$ is considered to be a positive, real valued parameter. The total energy of this macrostate is then

$$\mathbf{U}_{\mathcal{S}^L} = \sum_{\omega \in \mathcal{S}_{\mathbf{U}}^L} \mathbf{U}_\omega = -\sum_{\omega \in \mathcal{S}_{\mathbf{U}}^L} \log_2 Pr(\omega)$$

The total information required to specify it, however, differs from this and is given by

$$R(\mathbf{U}, L) = -\log_2 Pr(\mathbf{U}_{\mathcal{S}^L})$$

where

$$Pr(\mathbf{U}_{\mathcal{S}^L}) = \sum_{\omega \in \mathcal{S}_{\mathbf{U}}^L} Pr(\mathbf{U}_\omega)$$

By varying the energy parameter $\mathbf{U}$ one focuses on the constant-$\mathbf{U}$ subsets $\mathcal{S}_{\mathbf{U}}^L$, measuring their informational size via $R(\mathbf{U}, L)$.

Note that $\{\mathbf{U}_\omega : \omega \in \mathcal{S}_{\mathbf{U}}^L\}$ and $R(\mathbf{U}, L)$ can be calculated directly for a time series or a set of time series without reference to the generating process.

Statistical mechanics views the thermodynamic entropy as a measure of the microscopic information unspecified after stating a macrostate's parameters such as its energy. Here the thermodynamic entropy is determined by the multiplicity of the constant-$\mathbf{U}$ subsets:

$$S(\mathbf{U}, L) = \log_2 \|\mathcal{S}_{\mathbf{U}}^L\|$$

The entropy $S(\mathbf{U}, L)$ defines the typical states and fluctuations about them. We associate typicality with the standard statistical mechanical interpretation of entropy as the phase space volume associated with a parametrized set of macroscopic states. The parameter here, as

in statistical mechanics, is the energy. When we consider the entropy S as a macroscopic parameter, the equation

$$S = S(\mathbf{U}, L)$$

is a fundamental relation in the sense that it uniquely determines a macroscopic state via the U and L dependence of S. For a closed and isolated system within a specified energy interval $\Delta\mathbf{U}$

$$S(\mathbf{U}^*, L) = \sup_{\mathbf{U}\in\Delta\mathbf{U}} S(\mathbf{U}, L)$$

where $\mathbf{U}^*$ is the value of U in the interval $\Delta\mathbf{U}$ at which $S(\mathbf{U}, L)$ attains its maximum. The typical microstates of the system are members of the equilibrium macrostate ensemble. In statistical mechanics this is the microcanonical ensemble and in the context of inference is the situation in which one is given a complete set of data and there is no interaction of the system under study with the observer, the energy shell of width $\Delta\mathbf{U}$ is fixed, and the inference is essentially trivial in that all available microstates are weighted equally.

For a closed but not isolated system, for which only the average energy is specified, a variational principle determines the equilibrium macrostate as that having an energy $\mathbf{U}^*$ maximizing the entropy

$$S(\mathbf{U}^*, L) = \sup_{\mathbf{U}} S(\mathbf{U}, L)$$

Typical microstates, in the sense discussed in the introduction, are those sequences with the measured thermodynamic parameters $(\mathbf{U}^*, L, \mathbf{S}(\mathbf{U}^*, L))$. In statistical mechanical language we are now dealing with a canonical ensemble. Using the canonical ensemble and associated distribution over microstates in particular and closed but not isolated systems in general is a natural context in which to discuss inference; the observer determines the weights of the various microstates from the available data, i.e. from an interaction with the system under study. Note that the probability of a fluctuation about equilibrium, i.e. a measure of the typicality of a fluctuation, is by Einstein's formula [22]

$$Pr(\mathbf{U}) \propto e^{-[S(\mathbf{U}^*, L) - S(\mathbf{U}, L)]}$$

This gives a measure of the probability of observing a sequence from the class of sequences having energy U. In information theory this is the essential content of the Shannon-McMillan theorem.[3] In that context the Shannon-McMillan theorem allows for the construction of efficient codes by ignoring atypical sequences.

To complete the association of particular problems of inference with standard thermodynamic systems we consider open systems, i.e. thermodynamic systems that can exchange matter as well as energy with their environment. Our interpretation of the equivalent problem of inference is a situation in which an observer's model takes account of the possibility of accounting for more or less data, i.e. we allow for fluctuations in the number of modeled degrees of freedom.

As in standard equilibrium thermodynamics we can move into other representations that use intensive parameters via an appropriate Legendre transform $\mathcal{L} \circ S(\mathrm{U}, L)$ of the entropy. The Legendre transform $\mathcal{L} \circ S(\mathrm{U}, L)$ that replaces energy as the extensive parameter by β, an intensive parameter, yields the Helmholtz free energy $F(\beta, L)$ via

$$-\beta F(\beta, L) = S(\mathbf{U}, L) - \beta \mathbf{U}$$

in which

$$\beta = \frac{\partial S}{\partial \mathbf{U}}$$

Note that while β is an intensive parameter it should be distinguished from intensive parameters which are obtained by considering the thermodynamic limits of extensive quantities. For systems in equilibrium it is common to obtain the thermodynamic potentials, such as the free energy $F(\beta, L)$ as appropriate Legendre transforms of the energy function $\mathrm{U}(S, L)$ which defines a fundamental relation via the equation

$$\mathbf{U}=\mathrm{U}(S, L)$$

just as $S(\mathrm{U}, L)$ does. Although less common, taking the Legendre transform of the entropy function $S(\mathrm{U}, L)$, as is convenient in our case, is formally equivalent to considering

Legendre transforms of $\mathbf{U}(S, L)$. The related macroscopic quantities, which correspond to the thermodynamic potentials, are known as Massieu functions, e.g. $-\beta F(\beta, L)$ in the above case.[5]

We see that if the only information available about a particular process, is specified by $S(\mathbf{U}, L)$ then the principle of maximum entropy specifies the equilibrium state as that for which $\beta = 0$. The interpretation of this is in terms of a high temperature limit for which the ratio of bond strength (as a measure of the information contained in the transition probabilities) to ambient temperature is vanishingly small. If alternatively the effect of the the environment on the system is negligible this ratio diverges, $\beta = \infty$, and the equilibrium state is the only state, i.e. the state of lowest energy (highest probability) or ground state. For values of β between 0 and ∞ the equilibrium macrostate is determined via the Boltzmann distribution as a function of β. For the case $\beta = 1$ we note that the partition function, which will be discussed below, is

$$\begin{aligned} Z &= \sum_{s^L} 2^{\beta \log_2 Pr(s^L)} \\ &= \sum_{s^L} Pr(s^L) \\ &= 1 \end{aligned}$$

i.e. just the sum over the distribution on the microstates which is inferred from the data. Therefore, in this case, the free energy vanishes

$$\begin{aligned} \beta F &= -\log_2 Z \\ &= 0 \end{aligned}$$

As previously mentioned, the systems we discuss in this paper have a bounded spectrum and hence a discussion of β values between 0 and $-\infty$ make sense in terms of systems with population inversions, where the single equilibrium state at $\beta = -\infty$ corresponds to the state of highest energy (lowest probability)

In the context of inference β can be thought of as an averaging parameter. When $\beta = 1$ the Boltzmann factors $2^{\beta \log_2 Pr(s^L)}$ used to calculate averages reduce to the cylinder probabilities and the averages calculated are arithmetic means. With $\beta \neq 1$ the averages calculated are geometric means. For a nice discussion of "parametrized averaging" see [23].

Transtensive Thermodynamics

The extensive quantities were defined for cylinders of finite length L without regard for their scaling properties as a function of L. As expected, the extensive quantities diverge in the thermodynamic limit. For example consider the total energy of a sequence as given by equation (484). The probability of any particular sequence $Pr(s^L)$ vanishes as $L \to \infty$, therefore its total energy diverges. Is there a way to normalize the extensive quantities so that they and their fluctuations are comparable in magnitude across different lengths L? The answer is yes and the appropriate normalized quantities are the transtensive ones we now define.

To do this each cylinder s^L is viewed as an element of the support of uniform size $\{\epsilon_\omega = N^{-1}(L), \forall\, \omega \in \mathcal{S}^L\}$. The normalization we seek takes the change in the size $N(L)$ of the support into account. And so we define the transtensive energy to be

$$\mathcal{U}_{s^L} = \frac{\log_2 Pr(s^L)}{\log_2 N^{-1}(L)} \tag{501}$$

or

$$\mathcal{U}_{s^L} = \frac{\mathbf{U}_{s^L}}{-H_0(L)} \tag{502}$$

The total energy is then

$$\rho(\mathcal{U}, L) = \frac{\log_2 Pr(\mathcal{U}_{\mathcal{S}^L})}{\log_2 N^{-1}(L)} = \frac{R(\mathbf{U}, L)}{-H_0(L)}$$

with

$$\begin{aligned} Pr(\mathcal{U}_{\mathcal{S}^L}) &= \sum_{\omega \in \mathcal{S}^L_{\mathcal{U}}} Pr(\mathcal{U}_\omega) \\ &= Pr(\mathbf{U}_{\mathcal{S}^L}) \end{aligned}$$

The sum is taken over the probability level sets of $\mathcal{S}^L$ given by

$$\mathcal{S}^L_{\mathcal{U}} = \{\omega : \mathcal{U}_\omega = \mathcal{U}, \omega \in \mathcal{S}^L\}$$

As in the extensive case, by fixing a formal parameter $\mathcal{U}$ we focus on the constant-$\mathcal{U}$ subset $\mathcal{S}^L_{\mathcal{U}}$. Its informational size is measured with $\rho(\mathcal{U}, L)$. The relationship between the cylinder distribution and distribution of information parameters $\rho(\mathcal{U}, L)$ is shown schematically in figure 103.

We now define $s(\mathcal{U}, L)$, the transtensive entropy by counting cylinders with a given value of $\mathcal{U}$. Normalizing, we obtain

$$s(\mathcal{U}, L) = \frac{\log_2 \left\| \mathcal{S}_{\mathcal{U}}^L \right\|}{-\log_2 N^{-1}(L)} = \frac{S(\mathbf{U}, L)}{H_0(L)} \tag{506}$$

Using the natural estimator

$$Pr(\mathcal{U}_{\mathcal{S}^L}) = \frac{\left\| \mathcal{S}_{\mathcal{U}}^L \right\|}{N(L)}$$

we have

$$\rho(\mathcal{U}, L) = -\frac{\log_2 Pr(\mathcal{U}_{\mathcal{S}^L})}{\log_2 N^{-1}(L)} = s(\mathcal{U}, L) - 1 \tag{508}$$

We see that the quantities $\rho(\mathcal{U}, L)$ and $s(\mathcal{U}, L)$ are simply related. Due to the fact that we consider temporal sequences, the term on the RHS of equation (508) is unity. If we had assumed a particular embedding dimension d for the distribution, then that constant would have been d.

We also define a transtensive Helmholtz free energy $\mathcal{F}(\beta, L)$ via a Legendre transform of the transtensive entropy

$$\begin{aligned} -\beta \mathcal{F}(\beta, L) &= s(\mathcal{U}, L) - \mathcal{U}\beta \\ &= \frac{-\beta \mathbf{F}(\beta(\mathcal{U}), L)}{H_0(L)} \end{aligned} \tag{509}$$

where

$$\begin{aligned} \beta &= \frac{\partial s(\mathcal{U}, L)}{\partial \mathcal{U}} \\ &= \beta(\mathcal{U}, L) \end{aligned}$$

Intensive Thermodynamics

Intensive quantities are simply obtained from the finite, normalized transtensive quantities by taking the thermodynamic limit, i.e. letting $L \to \infty$. In this limit, and even for the simplest iterative stochastic processes such as the biased coin, the quantities

$$\rho(\mathcal{U}) = \lim_{L \to \infty} \rho(\mathcal{U}, L)$$

$$s(\mathcal{U}) = \lim_{L \to \infty} s(\mathcal{U}, L)$$

become continuous functions of the energy density

$$\mathcal{U} = \mathcal{U}_{\omega \in \mathbf{A}^\infty} = \lim_{L \to \infty} \mathcal{U}_{s^L}$$

.[20] Similarly, for the intensive free energy

$$\begin{aligned} -\beta f(\beta) &= \lim_{L \to \infty} -\beta \mathcal{F}(\beta, L) \\ &= \lim_{L \to \infty} [s(\mathcal{U}, L) - \mathcal{U}\beta] \\ -\beta f(\beta) &= s(\mathcal{U}) - \mathcal{U}\beta \end{aligned}$$

where $s(\mathcal{U})$ and $\mathcal{U}$ are as defined above and

$$\beta = \frac{\partial s}{\partial \mathcal{U}}$$

To clarify the distinction between the transtensive and intensive free energy we use the relation

$$H_0(L) \underset{L \to \infty}{\sim} C_0(L) + hL \tag{515}$$

derived in [12], where $C_0(L)$ is the topological complexity of a process and h is it's topological entropy. For finitary ϵ-machines, such as the examples discussed above, $C_0(L)$ converges to a finite limit as $L \to \infty$ and $H_0(L)$ diverges linearly with rate h, as discussed in [8]. Starting with the relation (509) and the transtensive quantities defined in equations (502) and (506)

$$\frac{\beta \mathbf{F}(\beta, L)}{H_0(L)} = -\frac{S(\mathbf{U}, L)}{H_0(L)} - \left(-\beta \frac{\mathbf{U}}{H_0(L)}\right)$$

We see that in the limit as $L \to \infty$

$$\lim_{L \to \infty} \frac{\beta \mathbf{F}(\beta, L)}{C_0(L) + h_0 L} = \lim_{L \to \infty} \left[-\frac{S(\mathbf{U}, L)}{C_0(L) + h_0 L} + \beta \frac{\mathbf{U}}{C_0(L) + h_0 L} \right]$$

Using relation (515)

$$\lim_{L \to \infty} \frac{\beta \mathbf{F}(\beta, L)}{L} = \lim_{L \to \infty} \left[-\frac{S(\mathbf{U}, L)}{L} + \beta \frac{\mathbf{U}}{L} \right]$$

and so we find the Legendre transform

$$-\beta f(\beta) = s(\mathcal{U}) - \mathcal{U}\beta$$

for finitary processes, i.e. for which $C_0(L)$ and h_0 are finite. How this limit is obtained depends on the convergence rate of $C_0(L)$ as a function of L. For finite data sets the distinction between the transtensive and intensive quantities may therefore be significant and it is for this reason that we wish to make the distinction. This is important in cases for which the asymptotic nature of a process is inferred from a finite data set. The degree to which intensive quantities, calculated from the inferred model, approximate transtensive quantities, obtained from subsequent finite data sets, is a measure of the statistical significance of the inference. This distinction is critical for high complexity process such as the process defined by sampling the time series of the logistic map at the period doubling accumulation, on a binary partition.[12] In this case $C_0(L)$ diverges linearly as a function of L.

Cylinder Fluctuation Spectra

We next consider two examples to illustrate the thermodynamic properties of finitary systems. The first case involves the analysis of a binary string produced by the simulation of a biased coin with $Pr(s^L = 1) = 0.6$ and $Pr(s^L = 0) = 0.4$. As has been shown for this case, [12], $C_0(L) = 0$ for all L and $h = 1$. The intensive quantities give arbitrarily good approximations of the transtensive quantities calculated at finite lengths. In figure 104 we show the transtensive function $L^{-1}S(\mathbf{U}, L)$ with $L = 10$, estimated from a simulated data set of length 10^7 as compared to the asymptotic function $s(\mathcal{U}, L)$ shown in figure 105, computed in a manner to be described below.

Our second example involves a data set generated by iterating the logistic map on a binary partition with Misiurewicz parameter value r_M. For this system $C_0(L) = 2$ for L equal to or greater than 10 and $h = 0.823171....$ In figure 106, we show $L^{-1}S(\mathbf{U}, L)$ with $L = 10$ and data set of length 10^7 for this case. We also show $s(\mathcal{U}, L)$ for this system in figure 107.

These examples clearly demonstrate the limitations of estimating intensive quantities directly from finite data sets. At $L = 10$ the convex hull, or upper envelope, of $L^{-1}S(\mathbf{U}, L)$ is clearly a

better approximation to $s(\mathcal{U}, L)$ in the case of the biased coin. In fact for high complexity and/or entropy processes the approximation of $s(\mathcal{U}, L)$ with $L^{-1}S(\mathrm{U}, L)$ is prohibitively resource intensive and more effective means of approximating $s(\mathcal{U}, L)$ are necessary.

In the following we will suggest an approach that infers the structure of a process from a finite data set. From this inferred structure we can approximate transtensive quantities for arbitrary L. We will demonstrate that, for finitary processes, this approach using what we refer to as reconstructed ϵ-machines yields a far more accurate measure of fluctuation spectra and is a far more efficient method for estimating intensive thermodynamic quantities than histogram estimation.

Reconstructed ϵ-Machines

In this section we review ϵ-machines: the model class that we will use to describe distributions over temporal sequences of measurement symbols. Elsewhere we have shown how an ϵ-machine can be reconstructed from a time series. Here we assume a machine is given and ask what properties of the underlying data source it captures. Equivalently, when using the machine as a generator of sequences, we ask for the properties of the output strings.

A finitary ϵ-machine $\mathrm{M} = \{\mathbf{V}, \mathbf{E}, \mathbf{A}, \mathrm{T}\}$ describes the structure of strings over symbols in some alphabet, $s \in \mathbf{A}$. The machine structure consists of a finite set of states, or vertices $\mathbf{V} = \{v_i : i = 0, 1, \ldots, V-1\}$. A set of labeled edges $\mathbf{E} = \left\{e : e = \left(v \underset{s}{\rightarrow} v'\right); v, v' \in \mathbf{V}, s \in \mathbf{A}\right\}$ gives the state to state transitions over a single time step. $E = \|\mathbf{E}\|$ is the total number of edges. There is a unique state $v_0 \in \mathbf{V}$ specified as the initial or start state. On top of the bare connectivity structure stochastic transition matrices give the conditional transition probabilities

$$\mathrm{T} = \left\{T^{(s)} : \left(T^{(s)}\right)_{ij;s} = t_{ij;s} \in [0,1], \;\; i, j = 0, \ldots, V-1, s \in \mathbf{A}\right\}$$

where $t_{ij;s} = p(v_j, s|v_i) = p(v_j|v_i, s)p(s|v_i)$ is the probability of the transition to state v_j on symbol s given the machine is in state v_i. We will assume, consistent with the general definition of an ϵ-machine as a causal model, that the underlying automaton is deterministic. This means that the symbol determines the successor state and so $p(v_j|v_i, s) = 1$ or, equivalently, $t_{ij;s} = p(s|v_i)$. Note that the transitions from each state are normalized

$$\sum_{\substack{v' \in \mathbf{V} \\ s \in \mathbf{A}}} p(v', s|v) = 1$$

Recall that the stochastic connection matrix is given by

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

By ignoring the edge labels, it describes a Markov chain over the machine states $\mathbf{V}$. It is interesting to note that the original machine M need not describe a finite state Markov

chain over the edge label alphabet A, or over finite blocks of symbols. Although M has a finite number of states, there can exist a countable or even an uncountable number of infinite sequences generated by M such that the probability of each does not satisfy the condition for the Markovian property of factoring into products of finite-conditioned conditional probabilities. One immediate consequence is that a strictly Markov representation of the sequences described by such a machine M requires an infinite number of states. Another consequence was seen in the comparison of the Golden Mean and Even systems. This mathematical fact is the basis of several important subtleties in the overall theory. Historically, it appears to have been the underlying cause for the introduction of Sofic systems into ergodic theory,[27] definite events in automata theory, [4]and the study of finite complement languages for the spatial patterns generated by cellular automata.[28, 19] We refer the reader to our review of this and related properties.[9, 10] As we will show later on in the following, this dichotomy has important consequences for fluctuation spectra. It also has important effects on an observer's ability to synchronize with a data source.[6]

The stationary probability of the machine states is given by the left eigenvector associated with the eigenvalue with magnitude unity of the stochastic connection matrix

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

With the state probabilities at hand, the measure-theoretic entropy h_μ for an ϵ-machine is directly computed via

$$h_\mu = -\sum_{v \in \mathbf{V}} p_v \sum_{\substack{v' \in \mathbf{V} \\ s \in \mathbf{A}}} p_{v \to v'} \log_2 p_{v \to v'}$$

It gives the asymptotic growth rate of Shannon information in sequences as a function of increasing length.

The measure-theoretic complexity gives the information size of the machine

$$C_\mu = -\sum_{v \in \mathbf{V}} p_v \log_2 p_v$$

It is the average amount of information in bits of observing a machine state.

A stochastic machine is a model for a data source in the sense that it describes first of all a unique distribution over the sequences it generates. Consider a particular sequence $s^L = s_0 s_1 \ldots s_{L-1}$. Its probability is given directly by the machine as follows

$$\begin{aligned} Pr(s^L) &= Pr(s_0 s_1 \ldots s_{L-1}) \\ &= Pr(s_0) Pr(s_1|s_0) Pr(s_2|s_0 s_1) \cdots Pr(s_{L-1}|s_0 \cdots s_{L-2}) \\ &= p(s_0|v_\lambda) p(s_1|v_{s_0}) p(s_2|v_{s_0 s_1}) \cdots p(s_{L-1}|v_{s_0 \cdots s_{L-2}}) \\ &= p(s_0|v_0) p(s_1|v_1) p(s_2|v_2) \cdots p(s_{L-1}|v_{L-1}) \end{aligned} \tag{526}$$

where the notation $v_{s_0 s_1 \cdots s_{k-1}}$ in the penultimate line refers to the state v_k to which the machine is brought upon following the sequence of transitions selected by the string $s_0 s_1 \cdots s_{k-1}$. λ is the null string. In this way, the machine allows for the factoring of the joint distribution $Pr(s^L)$ into a product of conditionally-independent transition probabilities. A machine is also a model in the sense that it gives an explicit picture of one mechanism by which the observed data could have been produced. It also generalizes from the observed data to unobserved sequences.

For completeness we note that there are several other uses to which an observer who has estimated a machine can put it. First, the machine describes the relaxation of the observer's knowledge of the source's state as successive measurements are collected. Second, the computational properties of the machine gives rise to a semantics of individual measurements. The meaning of a measurement is defined generally by the machine's structure and specifically by the state in which the machine is found at the time of the measurement.[6]

Finally we show the graphs associated with the ϵ-machines for the processes discussed in the above section on cylinder histograms. In figure 108 we show the single vertex process for the biased coin example.

In figure 109 we show the graph obtained by ϵ-machine reconstruction from the data set generated by iterating the logistic map on a binary partition with Misiurewicz parameter value r_M. The branching probabilities in the graph are estimated from the data stream.

In figures 110 and 111 we show the graphs for the golden mean and even processes defined above.

Lastly, in figure 112, we show the graph associated with a nonbinary process, called the trinomial process for obvious reasons. The trinomial process generates all strings defined on a three letter alphabet. This process proves useful in the general discussion of fluctuation spectra below.

Fluctuation Spectra via Machine Combinatorics

A previous section gave an explicit way to compute a process's fluctuation spectrum using subsequence histograms estimated from a representative time series. Now with the notion of a finitary machine, we show how the fluctuation spectrum can be obtained directly using the machine reconstructed from the same time series. The machine is used to reorganize the telescoping product of conditional transition probabilities, given in (526) above. Additionally, it operates as a generator of subsequences, many of which are not in the original time series. In this sense, the reconstructed machine allows one to generalize to unobserved behavior, the goal of any inductive modeling scheme. Finally, by studying the fluctuation spectrum in this particular context we obtain an explicit understanding of the added structure that is inferred, but **not** observed. This is striking in the (typical) case where the set of sequences with full measure, which we call "typical" sequences, is not the set of all sequences generated by the machine. In information theoretic terms, this occurs when the measure entropy is strictly less than the topological entropy. In this case, there are improbable, atypical sequences, possibly associated with a subprocess having positive topological entropy. From the observation of a typical sequence, it is possible to infer, within this Bayesian context, the existence of the atypical sequences. The basic analysis of this paper can be seen as an exploration of this phenomenon.

We develop a general procedure for calculating $S(\mathcal{U})$ for a given ϵ-machine, which is either inferred from data or can be simply taken to be the definition of a stochastic process. This is done in three steps. The first is the calculation of the energy, the second determines the entropy at the given energy, and the last step improves the method's efficiency. The method is illustrated in an analysis of various stochastic machines in the section that follows.

First we calculate $\mathcal{U}_{s^L}$ for a given cylinder s^L. Notice that $\mathcal{U}_{s^L}$ is completely determined by

the frequency with which each edge of the ϵ-machine is visited for the given cylinder

$$\mu(s^L) = \prod_{i=0}^{L-1} p(s_i|v_i) = p_1^{c_1^L} p_2^{c_2^L} \cdots p_E^{c_E^L}$$

where $\left\{ p_i \equiv p_{v \to v'} : i = 0, \ldots, E-1, v, v' \in \mathbf{V}, s \in \mathbf{A} \right\}$ is the set of conditional transition probabilities associated with the edges $\mathbf{E}$ and c_i^L is the number of times that the given sequence visits the i^{th} edge. We use this notation, that labels the counts with a single subscript denoting the appropriate edge, for simplicity. An equivalent notation that we will use later when we wish to emphasize the property of edges as labeled, ordered pairs of vertices, explicitly shows the vertex and label dependence of the particular edge. So for future reference we note that the following notation conventions are equivalent

$$c_i^L \equiv c_{v \to v'}^L \equiv c_{v,v',s}^L$$

where recall that

$$p_i \equiv p_{v \to v'} \equiv p_{v,v',s}$$

From the transtensive definition (501), we have

$$\mathcal{U}_{s^L} = -H_0^{-1} \log_2 \mu(s^L) = -H_0^{-1} \sum_{i=1}^{E} \ln p_i^{c_i^L}$$

Hence to compute $\mathcal{U}_{s^L}$ we only need know the edge visitation vector $\mathbf{c}^{s^L} = (c_1^L, c_2^L, \ldots, c_E^L)$. Having done so, we have also calculated $\mathcal{U}$ for all those other cylinders with the same visitation vector. This class will be denoted by its vector of edge counts $\mathbf{c}^L = (c_1^L, c_2^L, \ldots, c_E^L)$.

Define the transition information vector to be

$$\mathbf{I} \equiv -H_0^{-1}(\ln p_1 \ , \ \ln p_2 \ , \ \ldots \ , \ln p_E)$$

This expresses the variation in transition probabilities via their self-informations. Using it yields a compact matrix representation for a cylinder's energy

$$\mathcal{U}_{s^L} = \mathbf{c}^L \cdot \mathbf{I}$$

where $\mathbf{c}^L$ is the visitation frequency vector for s^L.

The transtensive entropy $S(\mathcal{U},L)$ is a measure of the multiplicity of energy states. Thus, to begin the second step in the development, we need to determine this combinatorial quantity. We define the $r \times E$ integer matrix

$$C^L = \begin{pmatrix} \mathbf{c}_1^L \\ \vdots \\ \mathbf{c}_r^L \end{pmatrix}$$

where each of the r rows is a distinct visitation vector $\mathbf{c}_i^L = (c_{i1}^L, c_{i2}^L, \ldots, c_{iE}^L), \quad i = 1, \ldots, r.$ Note that each row is normalized to the cylinder length

$$\sum_{j=1}^{E} c_{ij}^L = L$$

The number r of distinct visitation vectors can be calculated using Whittle's formula for Markov processes.[2] Unfortunately, this is no simpler than explicitly calculating the visitation vectors and enumerating them for a particular L via the procedure we give in the following. We do note, however, that the procedure is efficient and exact.

Next we define the $r \times E$ integer matrix

$$B^L = \begin{pmatrix} \mathbf{b}_1^L \\ \vdots \\ \mathbf{b}_r^L \end{pmatrix}$$

where the element $(B^L)_{ij}$ is the number of L-cylinders (i) whose visitation vector is $\mathbf{c}_i^L$ and (ii) the last transition along whose path is the j^{th} edge. The i^{th} row sum of B^L gives the number of L-cylinders having a value of $\mathcal{U}$ that is determined by $\mathbf{c}_i^L$. This defines the count

$$N\left(\mathcal{U}(\mathbf{c}_i^L), L\right) = \sum_{j=1}^{E} (B^L)_{ij}$$

A number of visitation vectors, though, may have the same value of $\mathcal{U}$. In that case we not only calculate the row sums as above but add together all the row sums associated with the visitation vectors $\mathbf{c}_i^L$ having the same $\mathcal{U}$

$$N(\mathcal{U}, L) = \sum_{i:\, \mathcal{U}(\mathbf{c}_i^L)=\mathcal{U}} \sum_{j=1}^{E} (B^L)_{ij}$$

This gives a unique value $N(\mathcal{U}, L)$ as a function of $\mathcal{U}$. Then from the transtensive definition (506) of the entropy we have the estimate

$$S(\mathcal{U}, L) = \frac{\log_2 N(\mathcal{U}, L)}{H_0}$$

To summarize, for a given length L, once C^L and B^L are computed, the calculation of $\mathcal{U}_L$ and $S(\mathcal{U}, L)$ is straightforward.

At this point we could stop and use the above to improve the previous estimation from cylinder histograms. This is, however, an unnecessary way station since we are assuming to have a machine or to have estimated one from the reconstruction method. The result is that we can go much further to develop a recursive algorithm to generate C^{L+1} and B^{L+1} from C^L and B^L. To do this, we systematically consider all possible visitation vectors with cylinder length $L+1$ by adding one to each component of each legal length L visitation vector. Then we test the new frequency vector for legality and whether or not it has already occurred in C^L; counting multiple occurrences along the way. For example, at some point in the procedure we construct

$$\mathbf{c}_{i'}^{L+1} = (c_{i_1}^L \quad \cdots \quad c_{ij}^L + 1 \quad \cdots \quad c_{iE}^L)$$

The we consider the quantity

$$B_{i'k}^{L+1} = \sum_{j=1}^{E} B_{ij}^L (T_0)_{jk}$$

where T_0 is the machine's connection matrix. If $B_{i'k}^{L+1}$ is zero then $\mathbf{c}_{i'}^{L+1}$ is not a legal vector. If it is not zero, then we test whether or not it is the same as any of the existing visitation vectors.

If it is, then $B^{L+1}_{i'k}$ is added to the k^{th} column of the row corresponding to the existing visitation vector in B^{L+1}. If it is not, then we form a new row in C^{L+1} consisting of $\mathbf{c}^{L+1}_{i'}$ and form a new row in B^{L+1} with $B^{L+1}_{i'k}$ in the k^{th} column and zero's in all other columns.

This completes our description of an iterative procedure for generating C^L and B^L for any length L. The result is, of course, a way of calculating $\mathcal{U}_L$ and $S(\mathcal{U}, L)$. We have shown in this section how the ϵ-machine efficiently organizes the enumeration of sequences and the estimation of their probabilities. It is used like a Perron-Frobenius operator to estimate the sequence measure at successively longer cylinder lengths. Most importantly, it gives estimates of improbable and even unobserved sequences. The result is a fluctuation spectrum at any cylinder length and any degree of accuracy. The main errors between this estimate and that of the underlying process are localized in this way to the estimation of the machine itself. The errors are not directly dependent on the observation of low probability sequences as in the direct histogram approach.

Finally, we noted above that Whittle's theorem gives an alternative approach to the probability estimates for sequences from a Markov chain.[2] First, it is important to note that the class of stochastic processes specified by ϵ-machines is more general than that of Markov chains. Second, a comparison shows that the algorithm specified by Whittle's formula is less efficient and slower; requiring as it does storage of a larger set of matrices, each of larger size, than those defined above and the calculation of products of large combinatorial factors. Whittle's formula, though, is the theoretical basis for asymptotic goodness-of-fit criteria for finitary machines.

Comparison of Histogram and Combinatorial Fluctuation Spectra

The figures in this section compare Renyi entropy spectra estimated from histograms and from machines. It is clear that the latter, as introduced in the previous section, lead to superior estimates. The direct histograms are very bad. (See figures 104 and 106.) These can be corrected by assuming Gaussian distributions about the combinatorial peaks. This is then integrated, the sum being plotted over the combinatorial peak. An approximation to the convex entropy function $S(U)$ is the convex hull of the point set obtained directly from the histogram.

We stress here the importance of estimating a model of the observed process before calculating derivative statistics like fluctuation spectra. By reconstructing machines we are able to extrapolate to large cylinder length and approximate asymptotic quantities, given this model, far more efficiently than trying to estimate these quantities directly from histograms of the data. As an example, for the typical types of system discussed in this section we are able to obtain approximations to fluctuation spectra for cylinders of up to approximately $L = 70$ in times on the order of minutes on a Sparcstation 1, and with moderate memory requirements (obviously far less than 2^{70h_0} histogram bins required for direct histogram estimation)

Another important point to note is the ability of the combinatorial enumeration method (but also to a limited extent the histogram method) to detect observationally degenerate ground states, somewhat analogous to the multiple ground states found in the case of frustrated systems. Restrictions, due to the structure of the process (and explicitly contained in the graph representing the process) often lead to the case of multiple paths in the graph having the same energy, in particular, cases for which this energy is the lowest energy for sequences of a given length.

In figures 113 and 114 we plot the combinatorial fluctuation spectra as described in the last section, i.e. obtained from the machine, for the the biased coin and Misiurewicz examples.

An important distinction between the direct histogram and combinatorial enumeration plots is that while they approximate the same quantity, namely the fundamental function $S(U)$, the way they do so is slightly different. For the histograms we simply sum over all cylinders with the same energy to obtain the entropy, thereby ignoring degeneracies. It should be noted that this is a general characteristic of any method that attempts to determine fluctuation spectra directly from measurements of state space distributions.[17] In combinatorial enumeration, however, we distinguish between different sets of cylinders having different edge visitation patterns but the same energy, hence $S(U)$ may not actually be a function, i.e. for each energy it may be multivalued. This fact is stressed in [20] for multinomials and is seen clearly for the trinomial and even system cases discussed below. For $S(U)$ to be considered as a fundamental relation for a system and hence describe its thermodynamic stability it must in fact be a convex function. The standard technique, using either Lagrange multipliers or Darwin-Fowler methods, for constructing an approximation to $S(U)$ from finite sets is to consider the convex hull of the point set obtained in the S, U plane as was mentioned above in the context of direct histogram estimation. For a heuristic description of such a procedure, in the context of dimension spectra, see [1] and for a rigorous justification, in the context of large deviation theory, see [15].

In figure 115 we show a normalized count distribution over the edge simplex for a trinomial process as shown in figure 112. The edge simplex in this case is the two dimensional set of possible edge visitations for the given sequence length L. The reason that the simplex is two dimensional is that the length constraint

$$\sum_{i=0}^{2} n_i = L$$

restricts the number of edge visitations for a given edge, say n_2, to a single value once the other two edge visitation numbers, n_0 and n_1, have been determined. The count distribution is simply the number of possible walks in the graph representing the process, that yield that particular set of edge visitation counts. For the case of the trinomial process, these numbers are simply determined by the trinomial distribution, and this plot is simply a plot of the trinomial

coefficients. Note that in this figure the level sets determined by two dimensional planes parallel to the $x - y$ plane all have the same value of $S(U)$ and would be summed over when approximating fluctuation spectra directly from histograms.

We also show $S(U)$ as a function of U in figure 116, for the trinomial process. Note that in this simple case the $S(U)$ plot can be viewed as a projection of the simplex plot.

We also show in figures 117 and 118, $S(U)$ as a function of U obtained by combinatorial enumeration for two other three edge processes, both having the same overall probability for the occurrence of a one, $p(1)$. The first is the golden mean system (figure 110) and the second is the even system (figure 111). Note that for the even system, the transient behavior implied by soficity, [10], leads directly to a far greater complexity in the combinatorial enumeration plot of the approximation to $S(U)$ than in the case of the golden mean system.

Thermodynamics of ϵ-Machines

To study the fluctuations in sequences, we focus on the variation in their probabilities. To this end, we introduce parametrized symbol transition matrices $\left\{T_\beta^{(s)} : s \in \mathbf{A}\right\}$ where

$$\left(T_\beta^{(s)}\right)_{ij} = e^{\beta \ln p_{v_i \to v_j}}$$

We think of each setting of the formal parameter β as emphasizing a different set of sequences. These can be associated with equiprobability level sets within the set of all sequences. This interpretation of β will become clearer as the discussion proceeds.

The associated connection matrix is defined as above $T_\beta = \sum_{s \in \mathbf{A}} T_\beta^{(s)}$. The maximum eigenvalue λ_β and associated left $\vec{l}_\beta$ and right $\vec{r}_\beta$ eigenvectors are determined by the linear equations

$$\vec{l}_\beta T_\beta = \lambda \vec{l}_\beta$$

$$T_\beta \vec{r}_\beta = \lambda \vec{r}_\beta$$

The eigenvectors are chosen so that their product is normalized to give an effective state probability $\vec{p}_\beta = \left\{p_v = \left(\vec{l}_\beta\right)_v (\vec{r}_\beta)_v : v \in \mathbf{V}\right\}$ where

$$\sum_{v \in \mathbf{V}} (\vec{p}_\beta)_v = 1$$

The machine Renyi entropy is

$$h_\beta = \frac{\log_2 \lambda_\beta}{1 - \beta}$$

This should not be confused with the entropy rate of the sequences generated by the machine with transition weights biased by a given β. That rate, the thermodynamic entropy, is given below. The thermodynamic interpretation of h_β is that it is a free energy, as will be shown below.

The machine Renyi complexity is

$$C_\beta \equiv -\sum_{v \in \mathbf{V}} (p_\beta)_v \log_2 (p_\beta)_v$$

The partition function for the machine is

$$Z_\beta \equiv \sum_{e \in \mathbf{E}} e^{\beta \ln p_e}$$

The machine free energy is then

$$F_\beta = -\beta^{-1} \ln Z_\beta$$

From Intensive to Machine Thermodynamics

The preceding development started first by considering the probabilistic structure in subsequences. The $S(\mathcal{U})$ function was developed to capture the spectrum of fluctuations in that distribution. We then showed that the reconstructed machine allows one to obtain, in a Bayesian modeling framework, arbitrarily accurate estimates of $S(\mathcal{U})$ via an iterative enumeration algorithm for computing the probability of all the machine's subsequences. In that discussion, we focused on the total information quantities, considering the infinite length limit only at the end. In this section, we take the final step which is to show that the infinite sequence length fluctuation spectrum can be directly computed from the reconstructed machine. This and the preceding results give a complete picture of fluctuation spectra for finitary processes, as far as the underlying idea itself goes. The additional benefit is that this lays the foundation for an analysis of the statistical inference problem of finitary machine. This will be the subject of a future paper. The main insight is that the machine itself is to be considered an intensive thermodynamic structure: the effective equations of motion. As will be seen, quantities directly estimated from it control the growth rates and convergence rates of the total informational quantities discussed above.

Assuming that we have reconstructed a machine, as in the previous section, we can write the sequence probabilities

$$\begin{aligned}\mu(s^{\mathbf{L}}) &= Pr(s^{\mathbf{L}}) \\ &= Pr(s_0 s_1 \ldots s_{L-1}) \\ &= p(s_0|v_0)p(s_1|v_1) \cdots p(s_{L-1}|v_{L-1}) \\ &= \prod_{i=0}^{L-1} p(s_i|v_i) \\ &= \prod_{l,m \ st\, p_{l,m,s}\, n.e.\, 0} p_{l,m,s}^{c^{\mathbf{L}}_{l,m,s}}\end{aligned}$$

where

$$p_{l,m,s} = p_{l \underset{s}{\rightarrow} m}$$

and $c^{\mathbf{L}}_{l,m,s}$ is the number of edge visits to the edge $l \underset{s}{\rightarrow} m$ for $s^{\mathbf{L}}$ and such that

$$\sum_{l,m,s} c^{\mathbf{L}}_{l,m,s} = L$$

Then

$$\mu^{\beta}(s^{\mathbf{L}}) = \prod_{l,m \ s.t.\, p_{l,m,s}\, n.e.\, 0} p_{l,m,s}^{\beta c^{\mathbf{L}}_{l,m,s}}$$

and we define the partition function

$$Z_{\beta} = \sum_{s^{\mathbf{L}}} \mu^{\beta}(s^{\mathbf{L}})$$

If we group the cylinders into sets which have the same initial and final vertices we can write

$$\begin{aligned}Z_{\beta} &= \sum_{s^{\mathbf{L}}} \mu^{\beta}(s^{\mathbf{L}}) \\ &= \sum_{i,j} \left(T_{\beta}^{L-1}\right)_{i,j}\end{aligned}$$

where recall that T_{β} is a matrix with elements $(T_{\beta})_{ij}$ that are greater than or equal to zero, hence by the Frobenius-Perron theorem T_{β} has a largest simple eigenvalue and corresponding left and

right eigenvectors $\vec{l}, \vec{r}$. We normalize $\vec{l}, \vec{r}$, in what follows so that $\vec{l} \cdot \vec{r} = 1$. Noting that we can diagonalize T_β (or in the general case at least reduce it to Jordan canonical form)

$$BT_\beta B^{-1} = D_\beta$$

hence we can write

$$T_\beta = B^{-1} D_\beta B$$

and

$$\begin{aligned} T_\beta^{L-1} &= \left(B^{-1} D_\beta B\right)_1 \left(B^{-1} D_\beta B\right)_2 \cdots \left(B^{-1} D_\beta B\right)_{L-1} \\ &= B^{-1} D_\beta^{L-1} B \end{aligned}$$

so

$$\left(T_\beta^{L-1}\right)_{ij} = \sum_k a_{ijk} \lambda_k^{L-1}$$

where λ_k is the k^{th} eigenvalue of T_β. Therefore for large L

$$\begin{aligned} Z_\beta(L) &= \sum_{i,j} \left(T_\beta^{L-1}\right)_{ij} \\ &\approx \lambda_\beta^{L-1} \sum d_{i,j,\beta} + \mathrm{O}\left(\lambda_2^{L-1}\right) \end{aligned}$$

where λ_β is the largest eigenvalue of T_β, λ_2 is the second largest eigenvalue and we note that $d_{ij\beta} = r_i l_j$. We now define the machine free energy, equivalent to the previously defined free energy of the cylinder set

$$\begin{aligned} f(\beta) &= \lim_{L \to \infty} \frac{1}{L} \log_2 Z_\beta(L) \\ &= \log_2 \lambda_\beta \end{aligned}$$

We also define $b_{ij} = -\beta \log_2 p_{ij}$ where $(T_\beta)_{ij} = e^{b_{ij}}$. For simple eigenvalue λ_β the quantities $\lambda_\beta, \vec{r}, \vec{l}$ are differentiable with respect to $(T_\beta)_{ij}$. We proceed to take derivatives. Starting with

$$T_\beta \vec{r} = \lambda_\beta \vec{r}$$

we make a smooth perturbation (in virtue of the differentiability of λ_β and $\vec{r}$)

$$\left(T_\beta + dT_\beta\right)\left(\vec{r} + \vec{dr}\right) = \left(\lambda_\beta + d\lambda_\beta\right)\left(\vec{r} + \vec{dr}\right)$$

$$T_\beta \vec{r} + T_\beta \vec{dr} + (dT_\beta)\vec{r} \approx \lambda_\beta \vec{r} + \lambda_\beta \vec{dr} + (d\lambda_\beta)\vec{r}$$

$$T_\beta \vec{r} + (dT_\beta)\vec{r} \approx \lambda_\beta \vec{dr} + (d\lambda_\beta)\vec{r}$$

Now multiply both sides on the left by $\vec{l}$

$$\lambda_\beta \vec{l} \cdot d\vec{r} + \vec{l}(dT_\beta)\vec{r} \approx \lambda_\beta \vec{l} \cdot d\vec{r} + d\lambda_\beta \left(\vec{l} \cdot \vec{r} \right)$$

and

$$d\lambda_\beta = \sum_{i,j} l_i (dT_\beta)_{ij} r_j$$

and therefore

$$\frac{d\lambda_\beta}{d(T_\beta)_{ij}} = l_i r_j$$

Also, assuming that $f(\beta)$ is differentiable in the thermodynamic limit

$$\begin{aligned} df(\beta) &= d(\log_2 \lambda_\beta) \\ &= \lambda_\beta^{-1} d\lambda_\beta \\ &= \lambda_\beta^{-1} \sum_{i,j} l_i (dT_\beta)_{ij} r_j \\ &= \lambda_\beta^{-1} \sum_{i,j} l_i (T_\beta)_{ij} db_{ij} r_j \end{aligned}$$

where we have used

$$\begin{aligned} (T_\beta)_{ij} &= e^{-b_{ij}} \\ d(T_\beta)_{ij} &= -db_{ij} e^{-b_{ij}} \\ &= -(T_\beta)_{ij} db_{ij} \end{aligned}$$

so

$$\frac{\partial f(\beta)}{\partial b_{ij}} = -\lambda_\beta^{-1} l_i (T_\beta)_{ij} r_j$$

Thermodynamically the energy associated with each edge is determined by the asymptotic edge frequency for a given cylinder

$$\begin{aligned} c_{i,j} &= -\frac{\partial f(\beta)}{\partial b_{ij}} \\ &= -\lambda_\beta^{-1} l_i (T_\beta)_{ij} r_j \end{aligned}$$

where

$$c_{i,j} = \lim_{L\to\infty} \frac{c_{i,j}^{\mathbf{L}}}{L}$$

and where $c_{i,j,s}^{\mathbf{L}}$ was defined and discussed in the above section on fluctuation spectra via machine combinatorics. We have dropped the symbol subscript here because we are assuming that there is a single directed edge from vertex i to vertex j. This type of machine can always be obtained for any finitary process, [10], so we can drop the symbol subscript here without loss of generality. We define the stochasticized version, S_β, of T_β

$$(S_\beta)_{ij} = \lambda_\beta^{-1} r_i^{-1} (T_\beta)_{ij} r_j$$

The left eigenvector associated with the largest eigenvalue ($\lambda_{S_\beta} = 1$) is given by $\pi_i = l_i r_i$. We note that S_β obeys the standard properties of a stochastic matrix

$$(S_\beta)_{ij} \geq 0 \, , \sum_j (S_\beta)_{ij} = 1 \;\; \forall i \, , \;\; \sum_i \pi_i = 1$$

$$\sum_i \pi_i (S_\beta)_{ij} = \pi_j \;\; , \;\; c_{i,j,s} = \pi_i (S_\beta)_{ij}$$

where in the expression for $c_{i,j,s}$ we are not summing on the repeated index.

Finally we calculate the thermodynamic entropy of the system with inverse temperature β. Instead of summing over the energies of the cylinders to obtain the average energy we take the equivalent (in the thermodynamic limit) sum over edges in the machine

$$u_{i,j,s} = -\beta c_{i,j,s} \log_2 p_{i,j,s}$$

and we wish to consider

$$S(\mathcal{U}) = \inf_b \left(f(\beta) + \beta \mathcal{U} \right)$$

Taking the inf over b entails that

$$u_{i,j,s} = -\frac{\partial f(\beta)}{\partial b_{ij}} = -c_{i,j,s} \log_2 p_{i,j,s}$$

Hence we obtain

$$\begin{aligned} S(\mathcal{U}(\beta)) &= \log_2 \lambda_\beta + \sum_{i,j} c_{i,j,\beta}(-\beta \log_2 p_{i,j,s}) \\ &= \log_2 \lambda_\beta - \sum_{i,j} \pi_i (S_\beta)_{ij} \log_2 p_{i,j,s}^\beta \\ &= -\sum_{i,j} \pi_i (S_\beta)_{ij} \log_2 \left(\frac{p_{i,j,s}^\beta}{\lambda_\beta} \right) \end{aligned}$$

Rewriting this in terms of T_β and then the stochasticized matrix S_β we obtain

$$\begin{aligned} S(\mathcal{U}(\beta)) &= -\sum_{i,j} \pi_i (S_\beta)_{ij} \log_2 \left(\frac{(T_\beta)_{ij}}{\lambda_\beta} \right) \\ &= -\sum_{i,j} \pi_i (S_\beta)_{ij} \log_2 \left[r_i (S_\beta)_{ij} r_j^{-1} \right] \\ &= -\sum_{i,j} \pi_i (S_\beta)_{ij} \log_2 (S_\beta)_{ij} \\ &\quad + \sum_{i,j} \pi_i (S_\beta)_{ij} (\log_2 r_i - \log_2 r_j) \end{aligned}$$

Rewriting the last sum

$$\begin{aligned} S(\mathcal{U}(\beta)) &= -\sum_{i,j} \pi_i (S_\beta)_{ij} \log_2 (S_\beta)_{ij} \\ &\quad + \left(\sum_i \pi_i \log_2 r_i - \sum_j \pi_j \log_2 r_j \right) \end{aligned}$$

and noting that the two terms in parentheses are equal we finally obtain

$$\begin{aligned} S(\mathcal{U}(\beta)) &= -\sum_{i,j} \pi_i (S_\beta)_{ij} \log_2 (S_\beta)_{ij} \\ S(\mathcal{U}(\beta)) &= h_\mu(S_\beta) \end{aligned}$$

and we see that the thermodynamic entropy for the machine at parameter value β is just the K-S or metric entropy of the maximal, stochasticized machine S_β.

The main point of this section has been to show that the functions $\mathcal{U}(\beta)$ and $S(\mathcal{U}(\beta))$ are given directly in terms of the principle eigenvalue and eigenvectors of S_β and T_β.

The heuristic interpretation of this derivation is as follows. First, setting the formal β specifies a particular weighting on the machine. This can be thought of as changing the

estimated transition counts from which the machine was estimated. Then S_β is the stochastic machine with that alternative process's probabilistic structure.

Recall the relationship between $\mathcal{U}(\beta)$ and $S(\mathcal{U}(\beta))$

$$\mathcal{U}(\beta) = \beta^{-1}(h_\mu(S_\beta) - \log_2 \lambda_\beta)$$

$\mathcal{U}$ is a measure of the mismatch between the maximum entropy stochastic process and the raw partition function growth rate (pressure). Note the similarity to the thermodynamic relation

$$U = TS + F$$

where we have the Helmholtz free energy $F = -\beta \log \lambda_\beta$

With these expressions $\mathcal{U}(\beta)$ and $S(\mathcal{U}(\beta))$ in terms of the machine eigenvalues and eigenvectors, we can directly compute the infinite sequence fluctuation spectrum given a stochastic machine.

Fluctuation Extremes

We notice from the definition of the energy $\mathcal{U}$ that the fluctuation extremes are determined by minimum and maximum probability cycles γ in the machine. Specifically $\mathcal{U}_{\min} = -\frac{\log_2 Pr(\gamma_{\max})}{|\gamma_{\max}|}$ and $\mathcal{U}_{\max} = -\frac{\log_2 Pr(\gamma_{\min})}{|\gamma_{\min}|}$. If there is more than one maximum cycle, $|\gamma_{\max}| > 1$, then the ground state $\mathcal{U}_{\min}$ is degenerate and characterized by the fact that $S_{\min}(\mathcal{U}_{\min}) > 0$. Since, in this paper, we are dealing with systems having bounded spectra, equivalent arguments hold for the fluctuation maximum, i.e. $|\gamma_{\min}| > 1 \Rightarrow S_{\min}(\mathcal{U}_{\max}) > 0$.

We also mention that, as expected $S_{\max} = S_{\beta=\mu} = \log_2 \lambda_{\beta=0}$ and $S_{\beta=1} = h_\mu$.

Complexity Spectra

Preceding sections were concerned with the asymptotic growth of the total Renyi entropy and the convergence of the subsequence distribution in the "thermodynamic", large cylinder length limit. The reconstructed machines allow for a much more detailed investigation of the

nonasymptotic properties, as shown in [12]. We will show now how the complexity spectrum, introduced in [11], fits in by considering finite length cylinders. Recall the total Renyi entropy and its relationship to the complexity and the Renyi entropy rate

$$H_\alpha(L) = C_\alpha + h_\alpha L$$

Following the standard definitions, but applying them to the total Renyi entropy, there is an associated fluctuation spectrum $\left(\hat{S}(L), \hat{\mathcal{U}}(L)\right)$ at each finite length L. The interpretation of this is that at finite L all we assume to govern the process is the length L-cylinder distribution. Under this, there is a process generating infinite sequences satisfying those and only those constraints. And we can ask for the fluctuation of this "extended" process. It is easy to derive new functions

$$\hat{\mathcal{U}}(L) = \mathcal{U}^{\mathbf{V}} + LU$$

$$\hat{S}(L) = S^{\mathbf{V}} + LS$$

where $(S, \mathcal{U})$ are the same as above and $(S^{\mathbf{V}}, \mathcal{U}^{\mathbf{V}})$ are the fluctuation spectra determined by the state distribution $\vec{p}_v(S_\beta)$ via a Legendre transform from (C_β, β)

$$\mathcal{U}^{\mathbf{V}}(\beta) = \frac{\partial}{\partial \beta}(\beta - 1)C_\beta$$

$$S^{\mathbf{V}}(\beta) = -(\beta - 1)C_\beta + \beta \mathcal{U}^{\mathbf{V}}$$

The latter $S^{\mathbf{V}}(\beta)$ is the complexity spectrum. The interpretation of this is straightforward. The complexity spectrum describes the fluctuations of the state sequence distribution, but sample not according to the Markov process determined by the stochastic transition matrix, but instead via IID sampling using only the state probabilities. That is, sequences over the state alphabet, **not** over the measurement alphabet. This should be contrasted with what we have been concentrating on so far, the probabilistic structure of the sequences over the measurement alphabet, that label the transitions, which come ultimately from an instrument. The state alphabet is an inferred grouping of measurement subsequences such that the joint distribution of measurements factors into conditionally-independent distributions. This factoring greatly reduces the volume over which one must estimate probabilities from observed counts, for example. It also can be interpreted as the identification of a symmetry.

Through this, we see that the length L fluctuation spectrum linearly separates into two spectra: complexity of states and entropy of transitions. Asymptotically, the specific S and $\mathcal{U}$ are dominated by the conventional $(S,\mathcal{U})$, if we assume that the complexity spectrum C_β and $\frac{\partial C_\beta}{\partial \beta}$ are bounded in L. At a phase transition with vanishing metric entropy, such as period-doubling and frequency-locking where $C_\beta(L) \propto \log L$,[12] the former assumption does not hold and the complexity spectrum adds logarithmic corrections as $L \to \infty$.

This discussion also shows how the machine spectra are related to the sequence spectra. Thus, a similar conclusion holds as last section: that estimating one's assumed model and computing its properties is better than estimating those properties from raw data.

Machine Fluctuation Spectra

Examples

In figures 105 and 107 discussed in previous sections and figures 119, 120, and 121 below we show the asymptotic fluctuation spectra calculated directly from the ϵ-machines. In these cases we note that the asymptotic fluctuation spectra are determined strictly by the structure of the underlying Markov process, and hence for example are the same for the golden mean and even systems. This demonstrates that although fluctuation spectra are useful for the ways discussed in this paper, they are by no means complete invariants for any particular process as would be expected from the above discussion of complexity spectra. To determine important structural characteristics associated with the process, e.g. strict soficity, means other than fluctuation spectra are necessary.

As the extreme energy values are easily calculated directly from the graph as well as by combinatorial enumeration we note with no surprise that that the extreme values for the combinatorial enumeration plots agree well with the machine fluctuation spectra plots.

Representation and Utility: Some Final Comments

In the preceding we have assumed *a priori* that fluctuation spectra are the appropriate way to characterize a process. We note that there are limitations in the representational power in that they are not complete invariants over the space of finitary machines. Given this assumption, nonetheless, we would argue that it is tantamount in a large fraction of cases to restricting oneself to a model class of finitary machines. Given this, in turn, it is clear that one gains a tremendous amount of accuracy, clarity of purpose, and utility, by first estimating the model implicitly assumed and not the physical invariant. In a very practical sense, which we have attempted to demonstrate by examples, given the assumed model class more information can be extracted from a data stream.

Analysis goes back to statistical methods for Markov chains developed in the '50s and '60's.[2] The historical irony is that contemporary discussion of the fluctuation spectra fall far short of that early work's generality. The present paper has attempted to integrate this earlier work with the contemporary view and by considering machines that include stochastic Sofic systems, to get beyond pure Markov theory. The further historical irony is that we end up back in the framework of communication channels that Shannon investigated in the '40's. It was he who solved the stochasticization problem that was used to find the maximal machine.

As a final comment, we would like to throw the preceding work into question. And along with this to present a critique of the philosophical, although **not** the mathematical, basis of contemporary discussions of the application of dynamical systems. We have in mind especially the "multifractal" analysis for nonlinear modeling of experimental time series. As we have shown above given a reconstructed ϵ-machine specified by (say) 32 estimated transition probabilities the fluctuation spectrum for the process can be estimated. To be generous in the following we take these parameters to be real numbers. Consider the alternative approach, that we have also discussed, of estimating the fluctuation spectra directly from the time series. Independent of

how the spectrum is obtained, we end up with two representations of the underlying process: the machine and the fluctuation spectrum. What type of representation is the latter? How are we to specify it? The answers are simple. The spectrum is a function of a real-valued parameter and so the representation is an infinite number of real numbers. Rather more than 32 real numbers. The first point is just this straightforward: the estimated machine is a vastly more compact representation. Or conversely, the fluctuation spectrum is an exceedingly redundant representation. (Just as an infinite number of harmonic coefficients are required for a Fourier representation of a simple square wave.)

This presents us with a conundrum. Why use the fluctuation spectrum? What does it tell us? It can tell us no more that the reconstructed machine since the spectrum is (even efficiently) computable from it, with proportionate additional effort. It tells us something about an idealized and particular view of the statistical structure of the infinite sequences. But is it more useful in (say) prediction and control applications implemented in either digital or analog hardware than the machine itself? The conclusion seems to be that the machine in all practical circumstances, such as system identification, to take an example, is the best representation. This is in fact the natural conclusion and motivation of modeling as distinct from prediction. The plots we obtain from the machine enumeration do give a very intriguing and informative view of the explosive combinatorics of a process's (or a machine's) path algebra. For low complexity processes this graphical representation seems quite instructive. It also highlight the weakness of using only the supremum "$f(\alpha)$" curve.

Two further critique's of the ideas are: (1) From the view of machine reconstruction, dimension is a subsidiary and derivative concept of entropy; and (2) It is clear from the mathematical development that fluctuation spectra are only very coarse summaries of a finitary process's properties. It is less clear, but nonetheless true, that fluctuation spectra fail to capture crucial computational structures. In the case of the golden mean and even system analyses this was a type of infinite memory called parity.

Acknowledgements

The authors thank John Ashley and Brian Marcus for useful discussions and Professor Carson Jeffries and Jeff Scargle for their support during this project. JPC wishes to thank the Santa Fe Institute for the warm hospitality during this work's completion; which was supported by a Robert Maxwell Foundation Visiting Professorship. The research conducted at UC Berkeley was funded in part by the AFOSR.

Figures

Figure Captions

Figure 99 $\log_2 Pr(s^L)$ vs. $s^L \in [0,1]$ for a binary data stream of length $k = 10^7$ generated by simulation of a biased coin with $Pr(s = 0) = .4$ and $Pr(s = 1) = .6$ and for cylinder lengths L from 1 to 9.

Figure 100 $\log_2 Pr(s^L)$ vs. $s^L \in [0,1]$ for a binary data stream of length $k = 10^7$ generated by simulation of the logistic map on a binary partition at the parameter value $r_M \approx 3.927737$ and for cylinder lengths L from 1 to 9.

Figure 101 $\log_2 Pr(s^L)$ vs. $s^L \in [0,1]$ for a binary data stream of length $k = 10^7$ generated by simulation of the golden mean system for cylinder lengths L from 1 to 9.

Figure 102 $\log_2 Pr(s^L)$ vs. $s^L \in [0,1]$ for a binary data stream of length $k = 10^7$ generated by simulation of the even system for cylinder lengths L for 1 to 9.

Figure 103 Schematic representation of scaled histogram $(\mathcal{U}_{sL}, \rho(\mathcal{U}, L))$ indices.

Figure 104 Transtensive Entropy vs. Energy for cylinder length 10 and data set of length 10^7 for the biased coin with branching probabilities $Pr(s = 1) = .6$ and $Pr(s = 0) = .4$.

Figure 105 Intensive Entropy vs. Energy for the biased coin with branching probabilities $Pr(s = 1) = .6$ and $Pr(s = 0) = .4$.

Figure 106 Transtensive Entropy vs. Energy for cylinder length 10 and data set of length 10^7 for the logistic map at r_M.

Figure 107 Intensive Entropy vs. Energy for the logistic map at r_M.

Figure 108 Machine for the biased coin with edges labeled by symbols and corresponding branching probabilities, $Pr(s = 0) = .4$, $Pr(s = 1) = .6$.

Figure 109 Machine for the logistic map on a binary partition with Misiurewicz parameter value r_M and with edges labeled by symbols and corresponding branching probabilities estimated from the data stream.

Figure 110 Machine for the golden mean system with edges labeled by symbols and corresponding branching probabilities.

Figure 111 Machine for the even system with edges labeled by symbols and corresponding branching probabilities.

Figure 112 Machine for the trinomial process with edges labeled by symbols and corresponding branching probabilities, $Pr(s = 0) = .15$, $Pr(s = 1) = .6$, $Pr(s = 2) = .25$.

Figure 113 Fluctuation spectrum obtained by combinatorial enumeration for the biased coin process, with cylinders of length L= 300.

Figure 114 Fluctuation spectrum obtained from combinatorial enumeration for the Misiurewicz logistic process, with cylinders of length L= 39.

Figure 115 Normalized count distribution over edge simplex for the trinomial process with L = 50 .

Figure 116 Fluctuation spectrum obtained from combinatorial analysis of machine for the trinomial process, with cylinders of length L= 50.

Figure 117 Fluctuation spectrum obtained from combinatorial analysis of machine for the golden mean process, with cylinders of length L= 300.

Figure 118 Fluctuation spectrum obtained from combinatorial analysis of machine for the even system, with cylinders of length L= 70.

Figure 119 Machine fluctuation spectrum for the trinomial system with $Pr(s = 0) = .15$, $Pr(s = 1) = .6$, $Pr(s = 2) = .25$.

Figure 120 Machine fluctuation spectrum for the golden mean system with $Pr(s = 1) = .6$.

Figure 121 Machine fluctuation spectrum for the even system with $Pr(s = 1) = .6$.

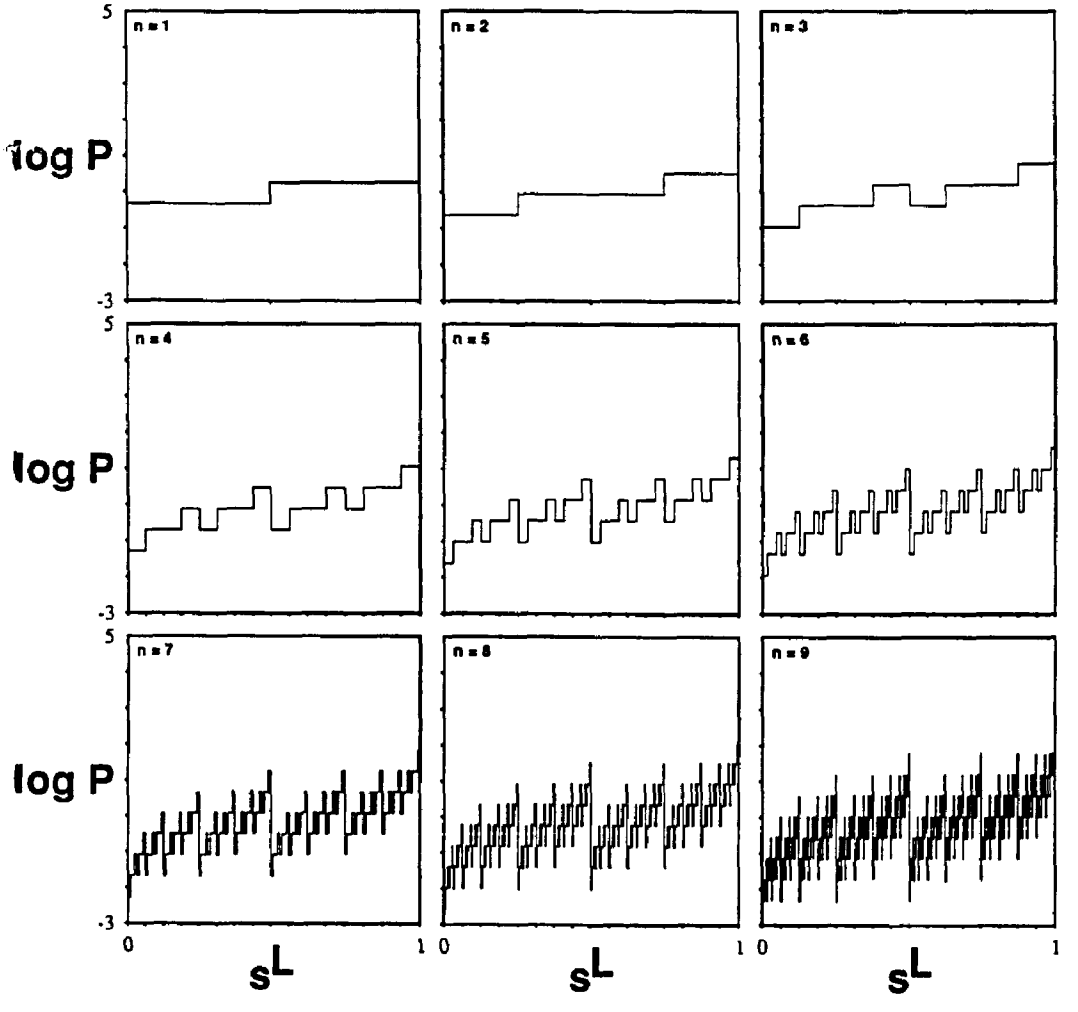


Figure 99

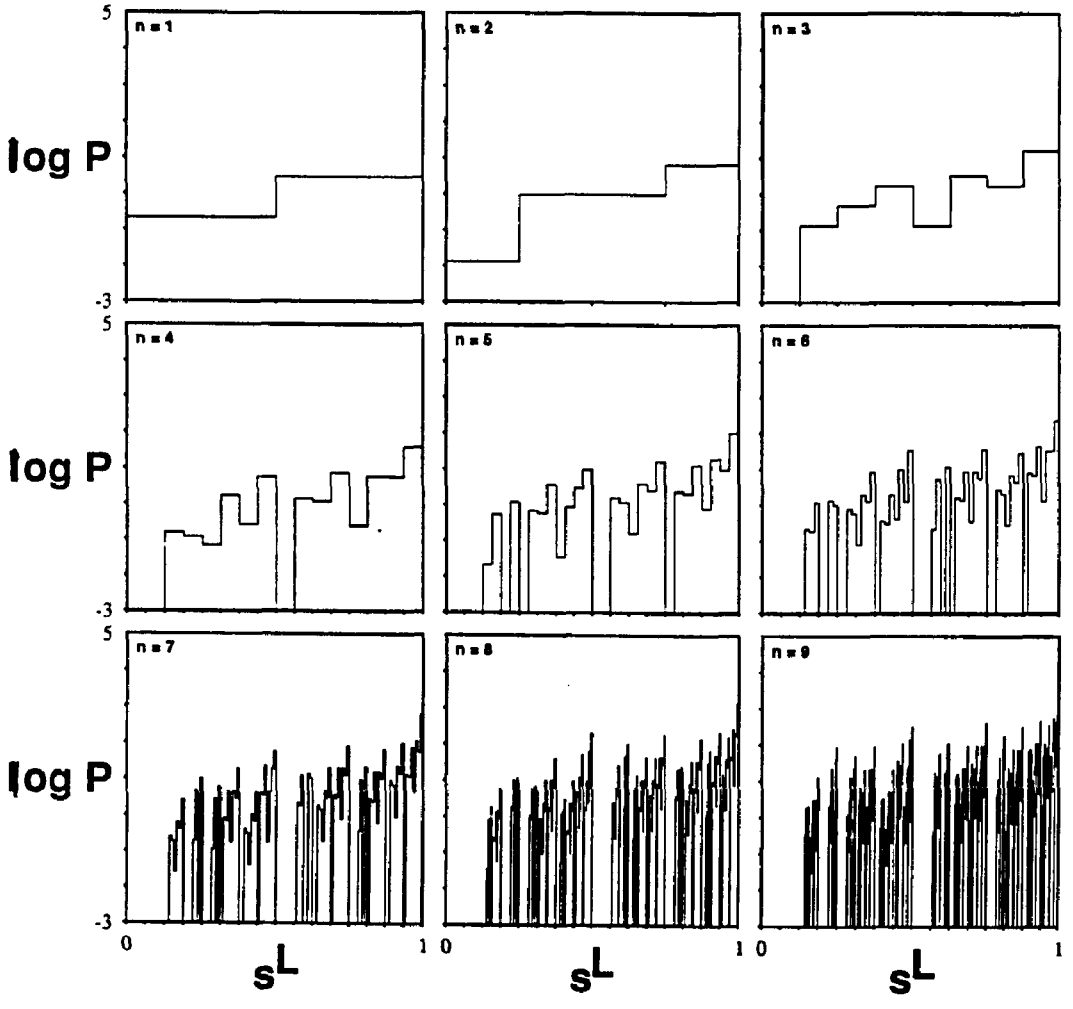


Figure 100

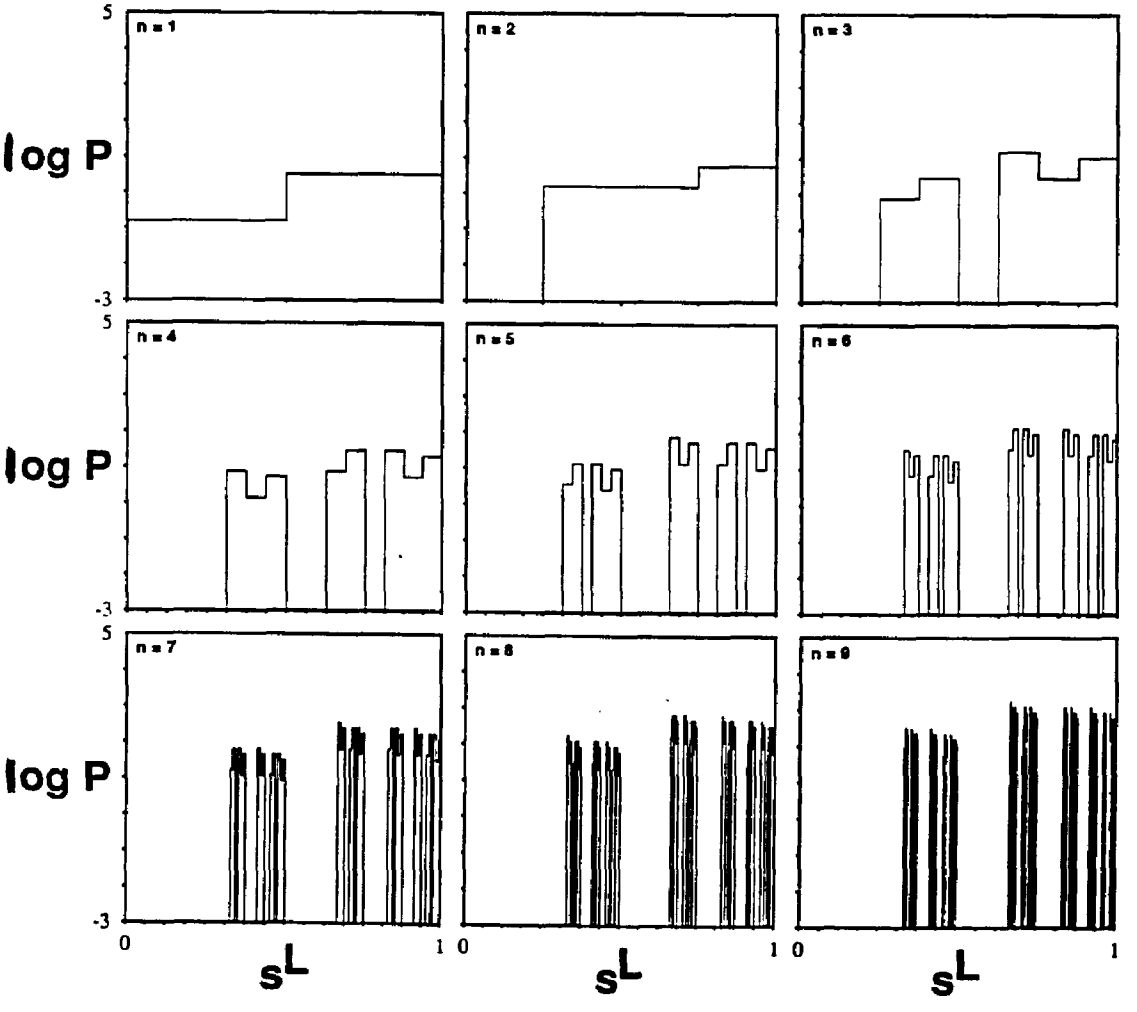


Figure 101

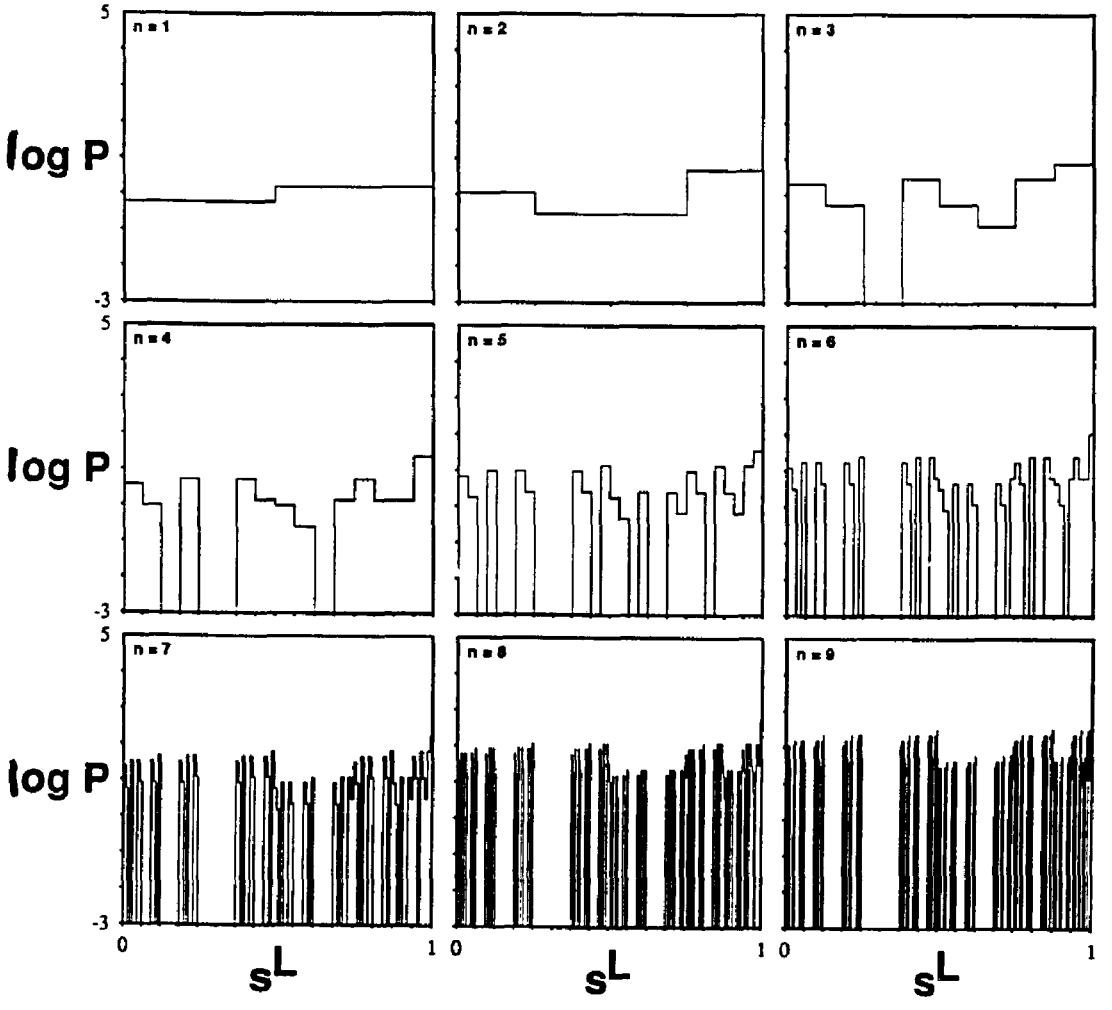


Figure 102

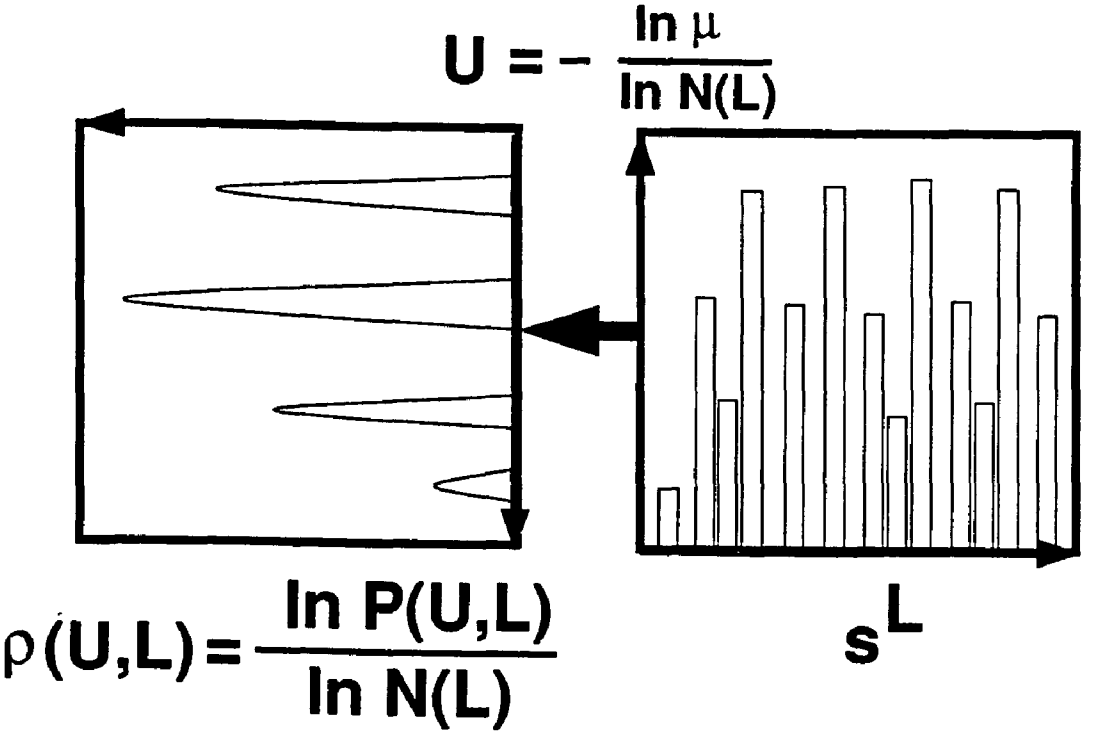


Figure 103

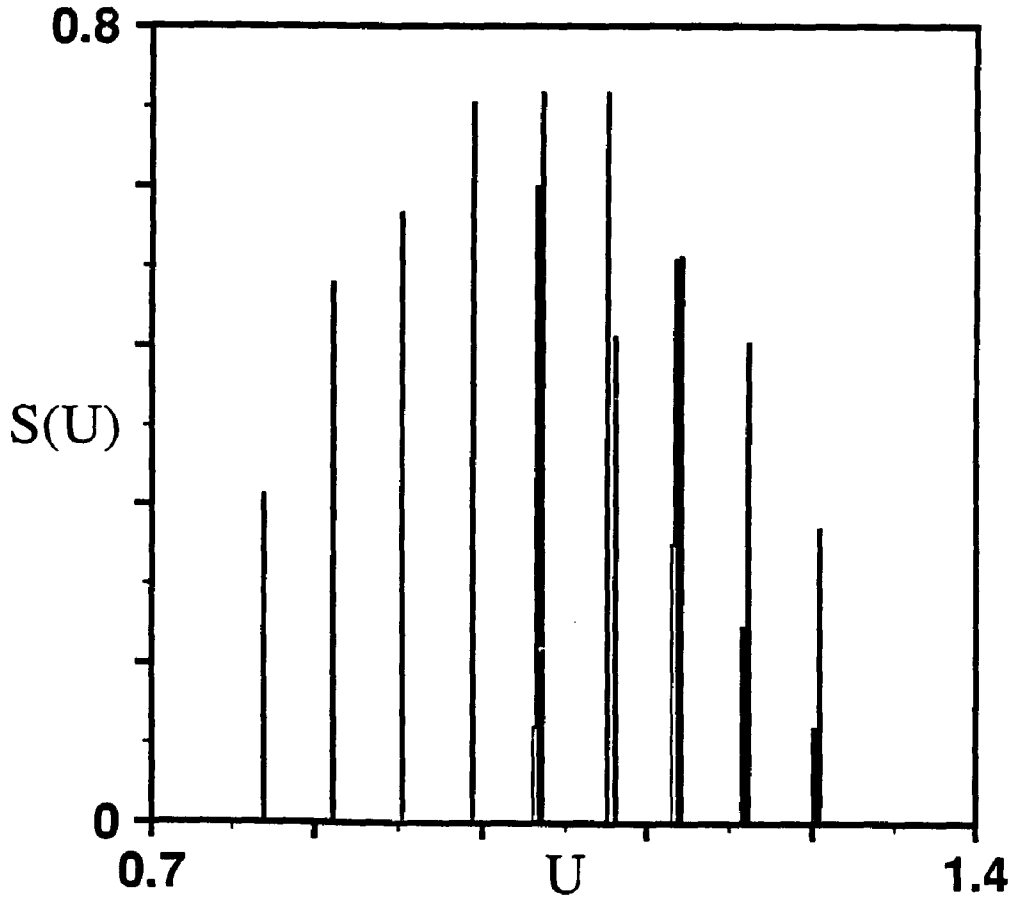


Figure 104

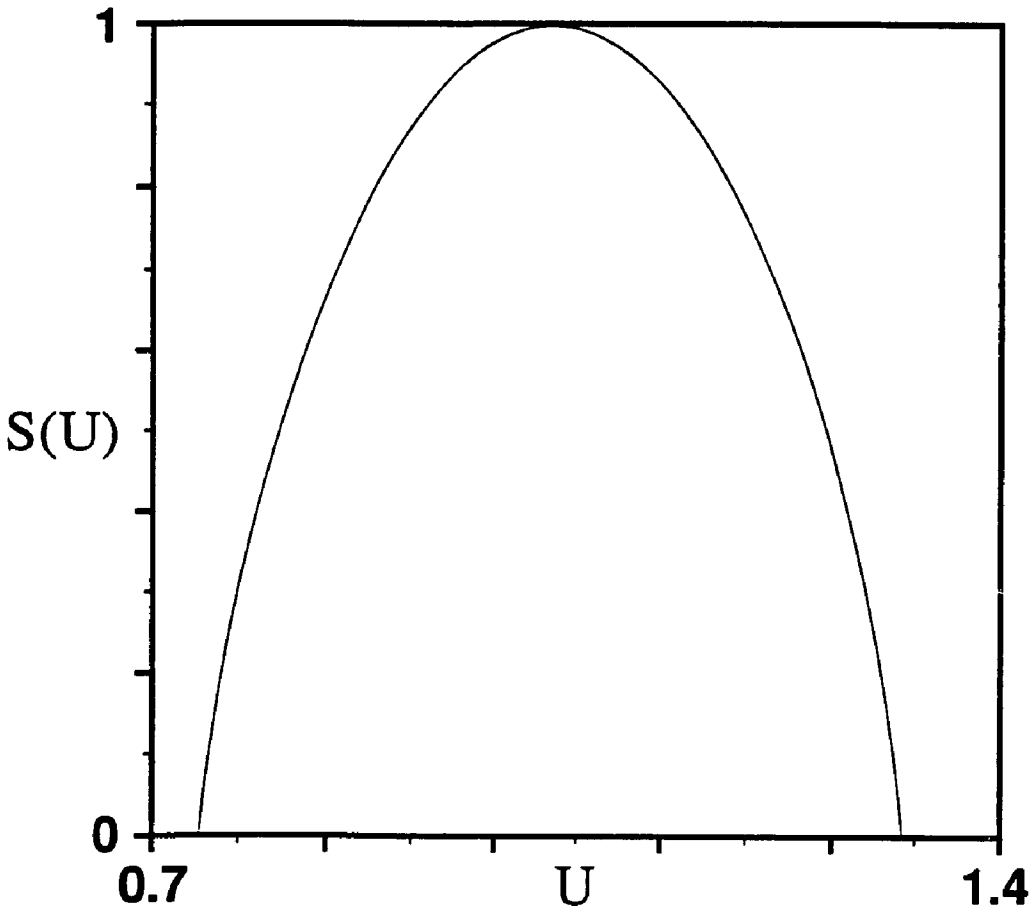


Figure 105

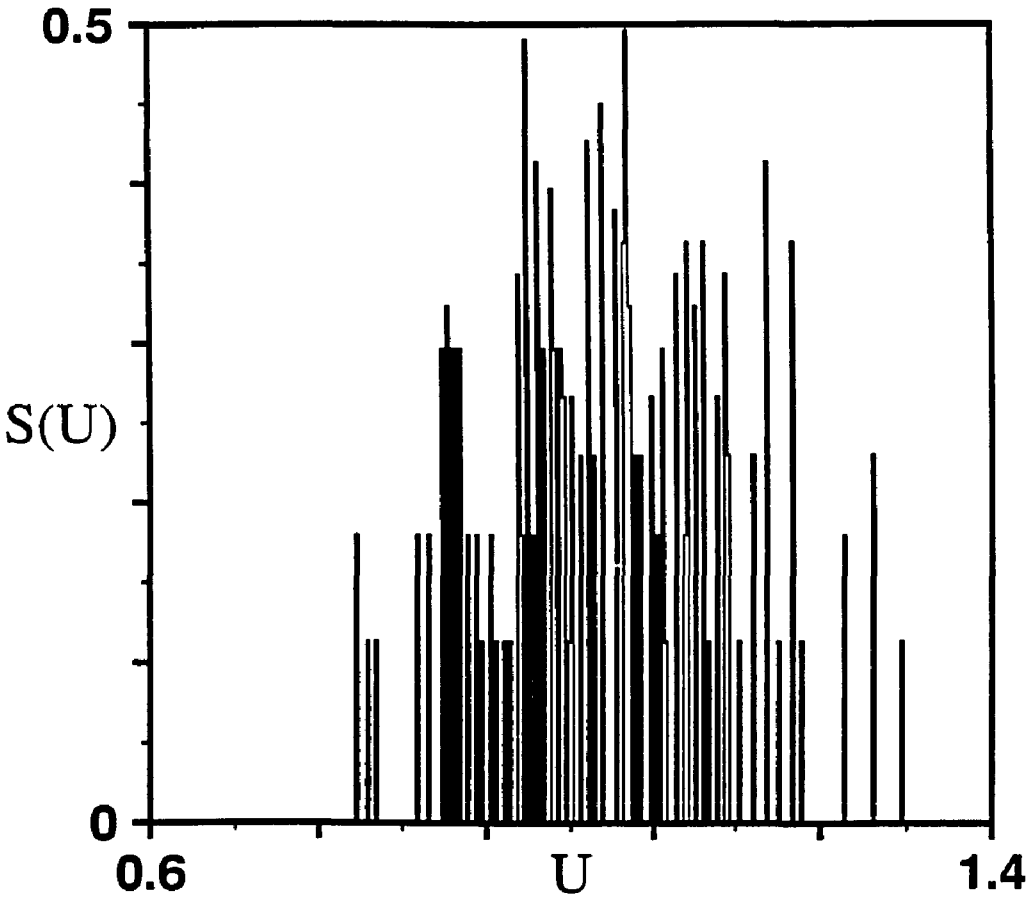


Figure 106

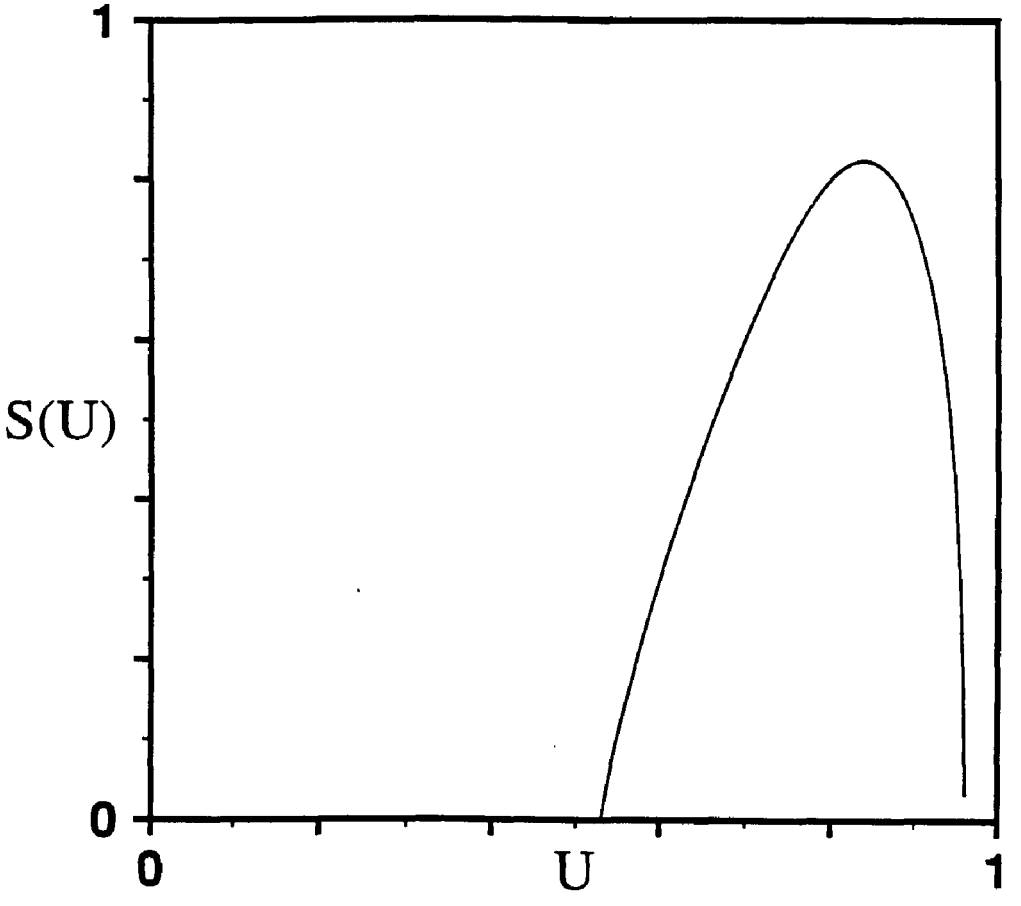


Figure 107

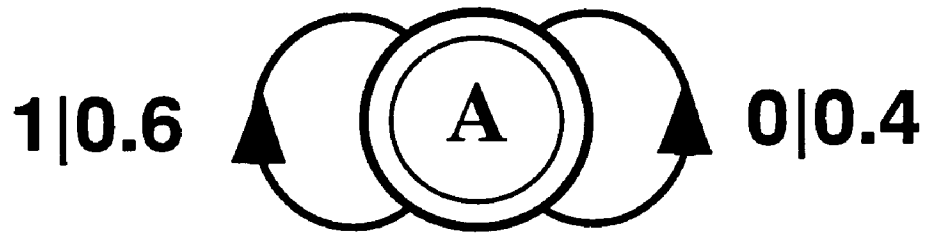


Figure 108

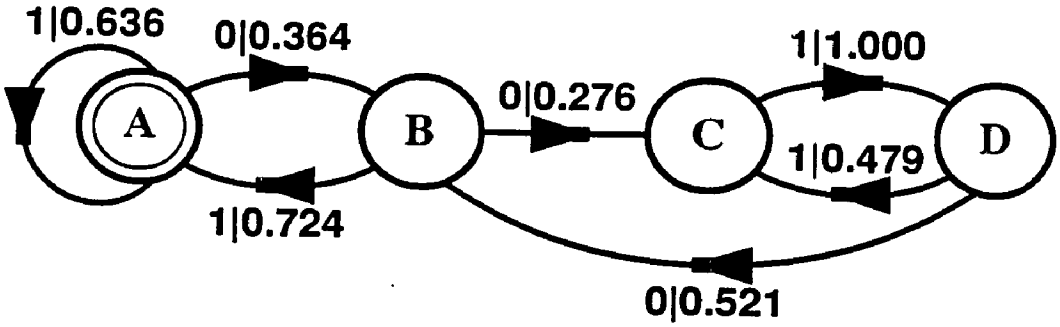


Figure 109

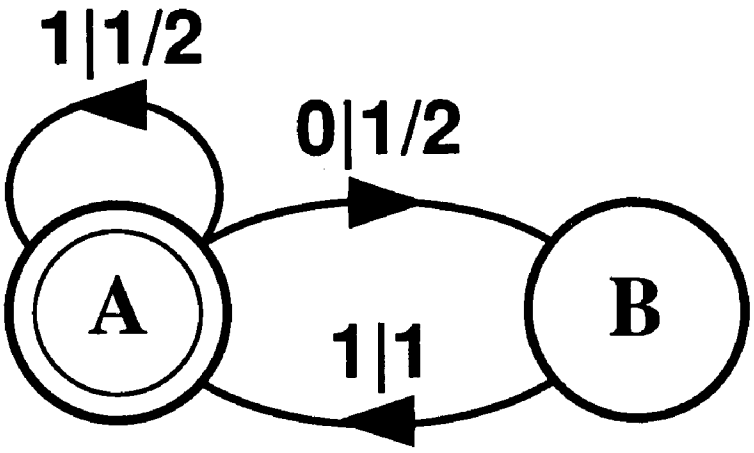


Figure 110

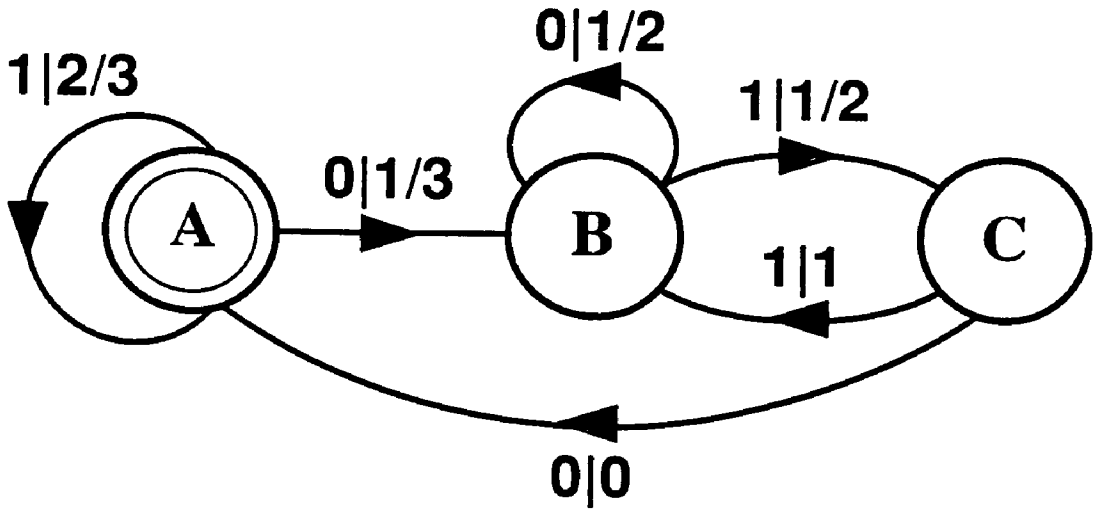


Figure 111

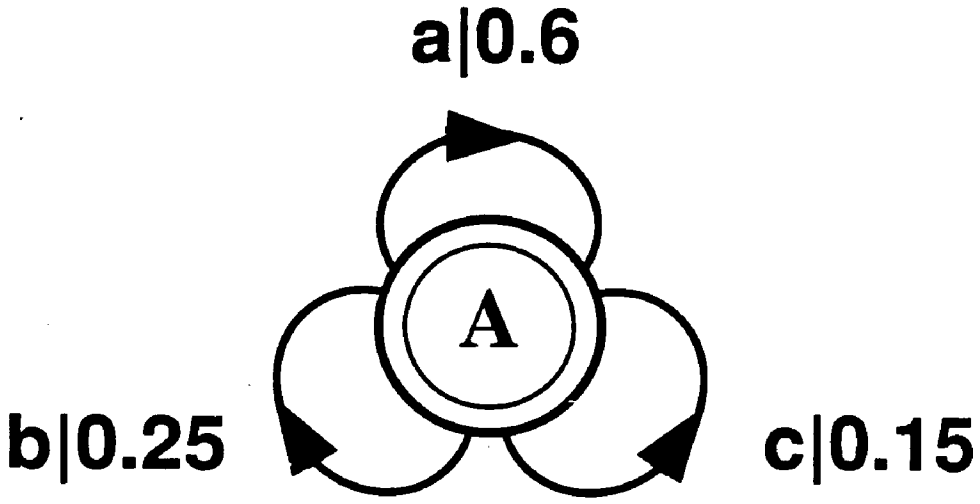


Figure 112

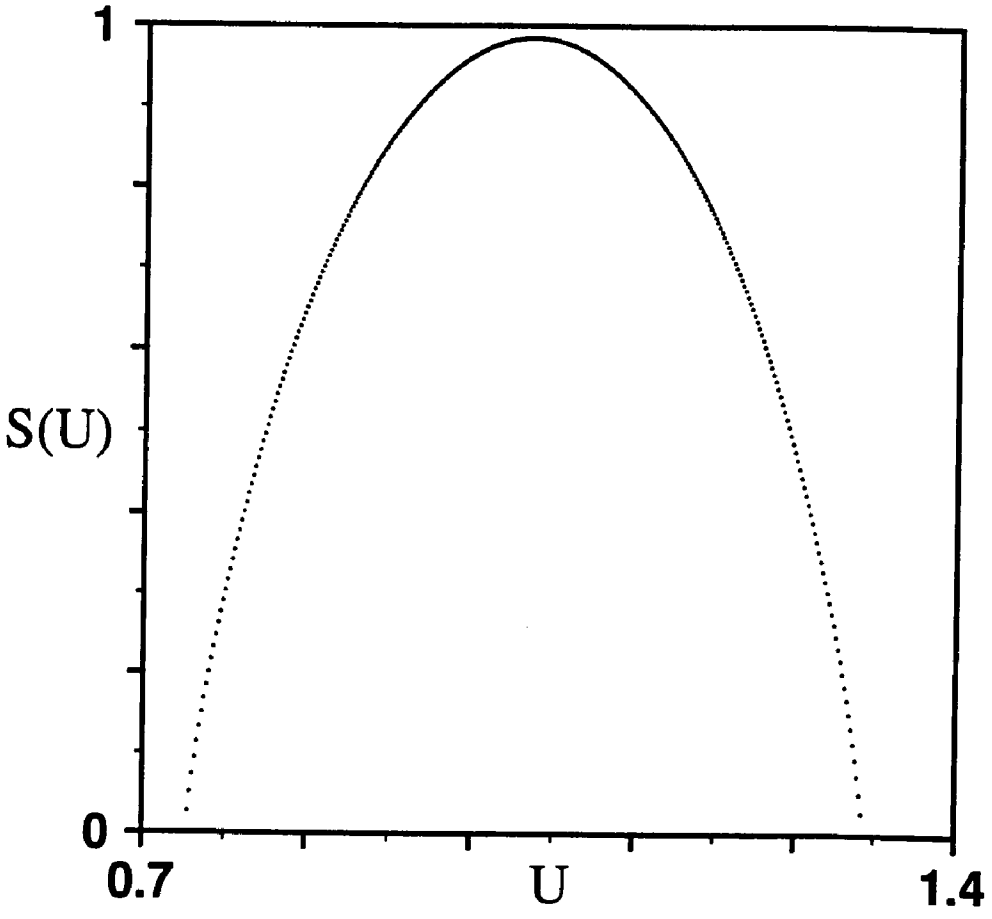


Figure 113

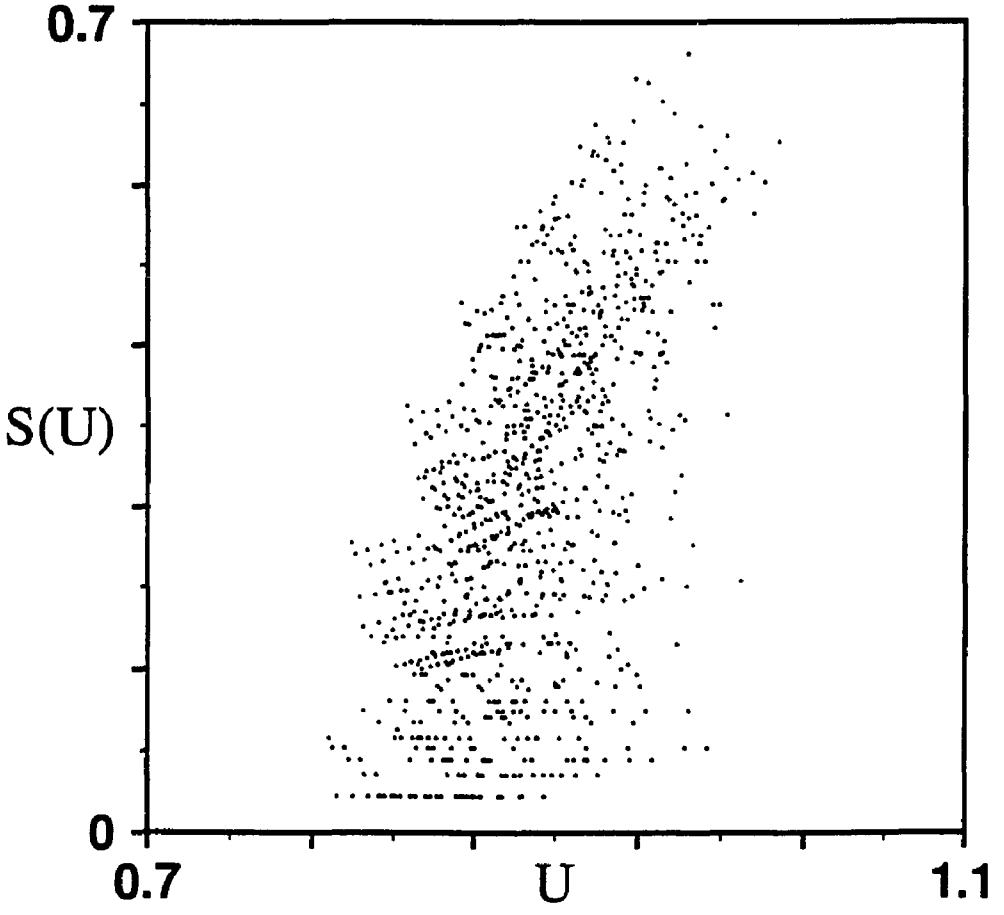


Figure 114

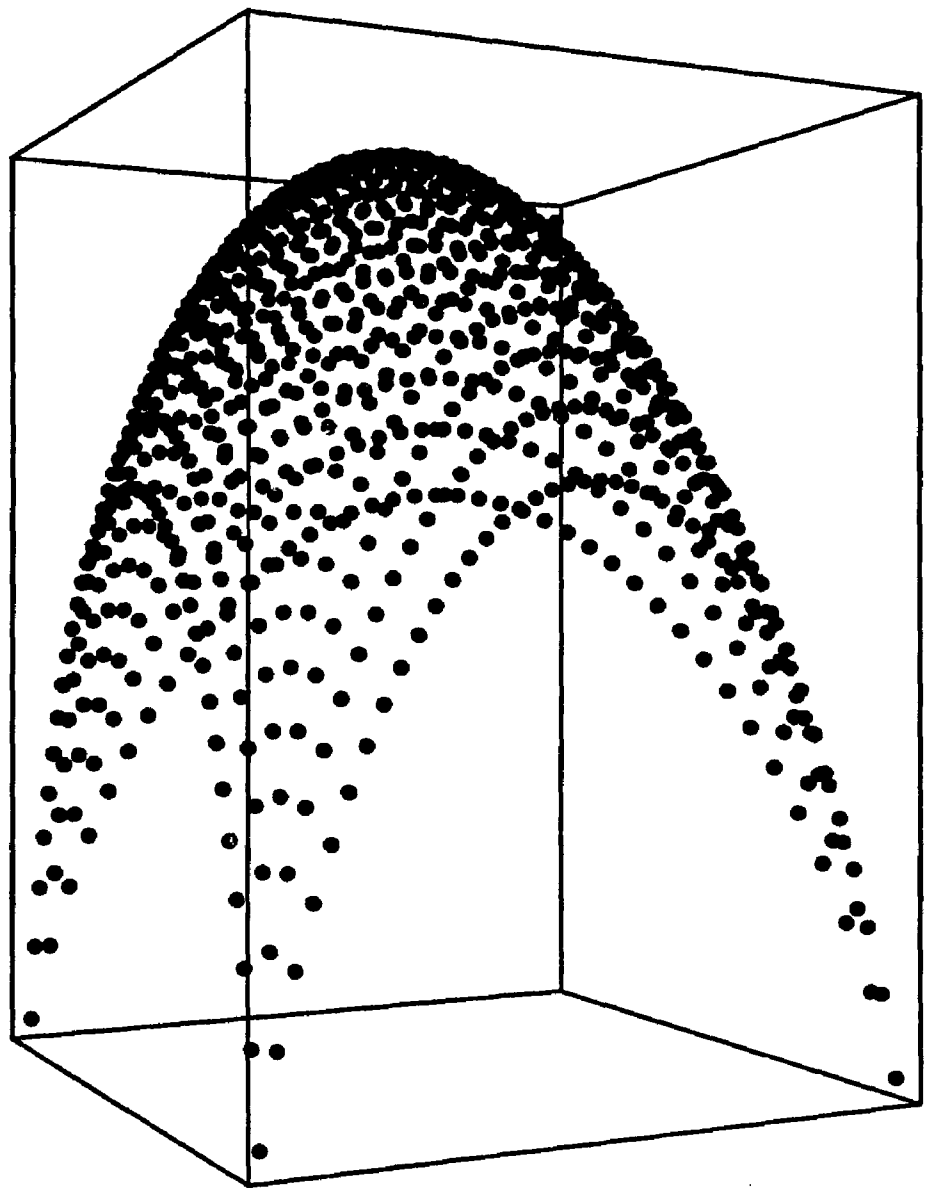


Figure 115

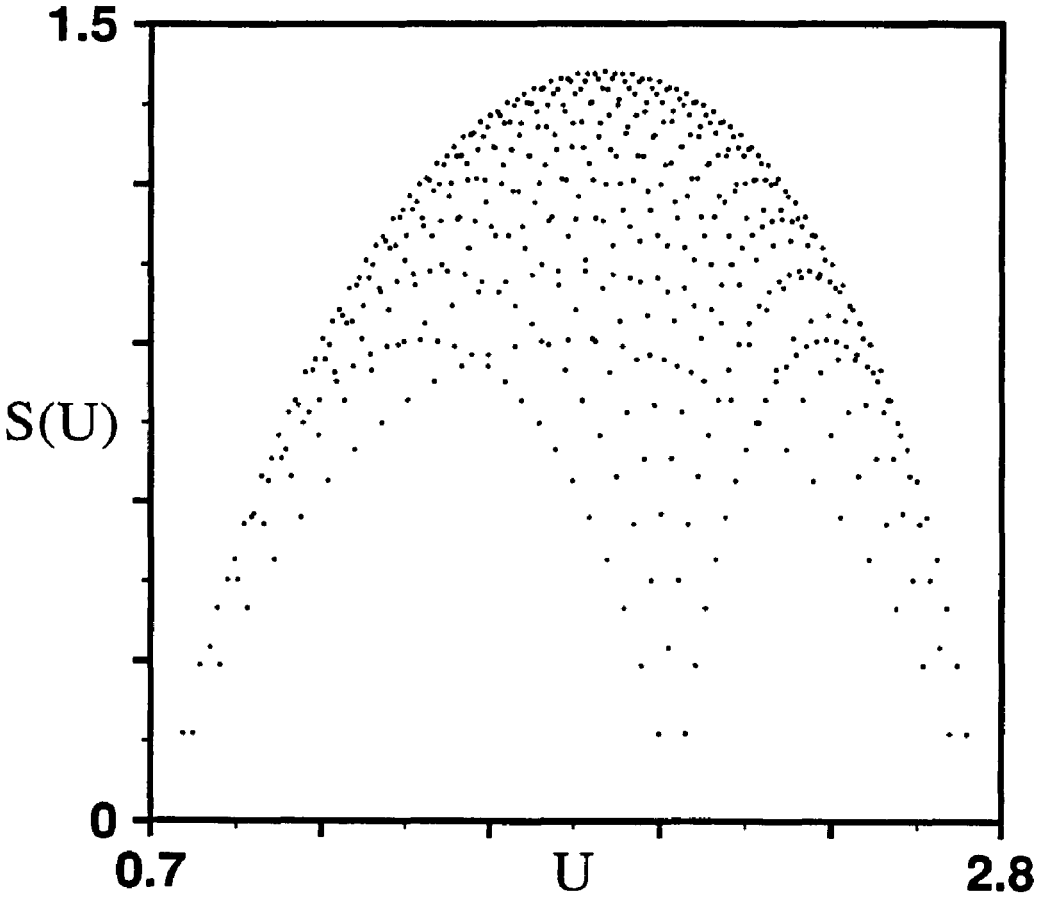


Figure 116

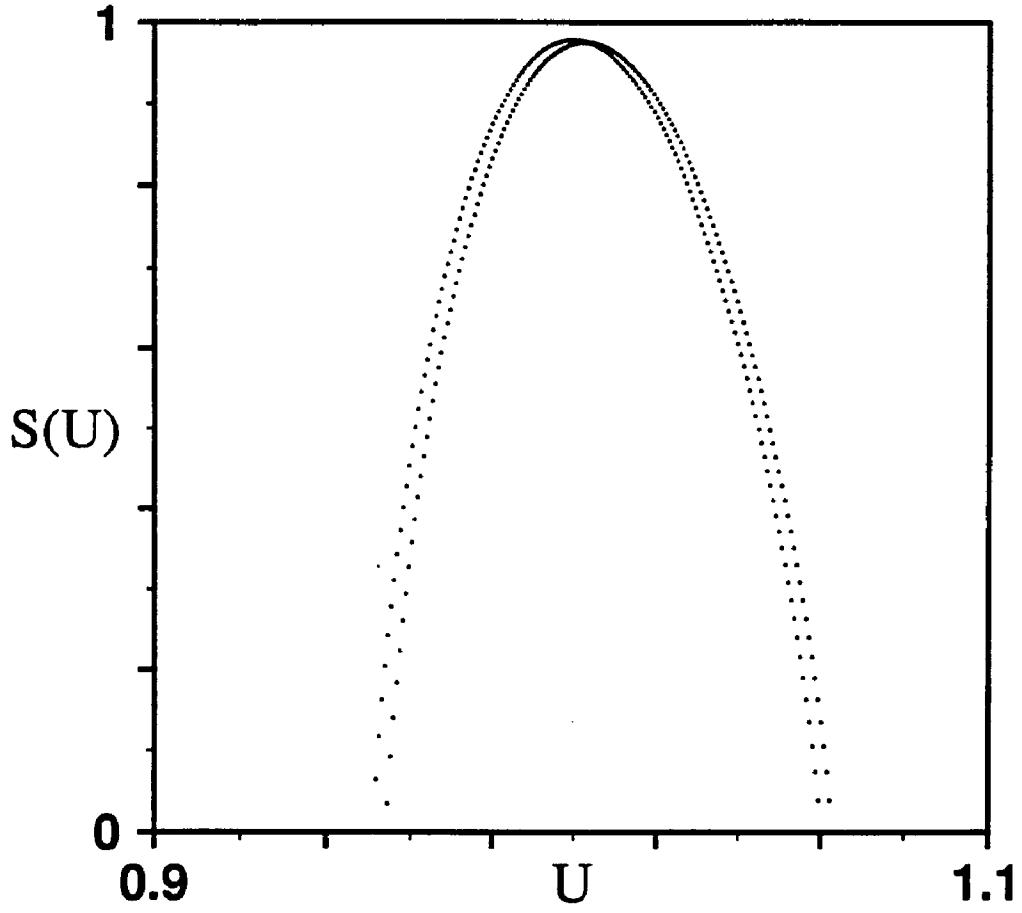


Figure 117

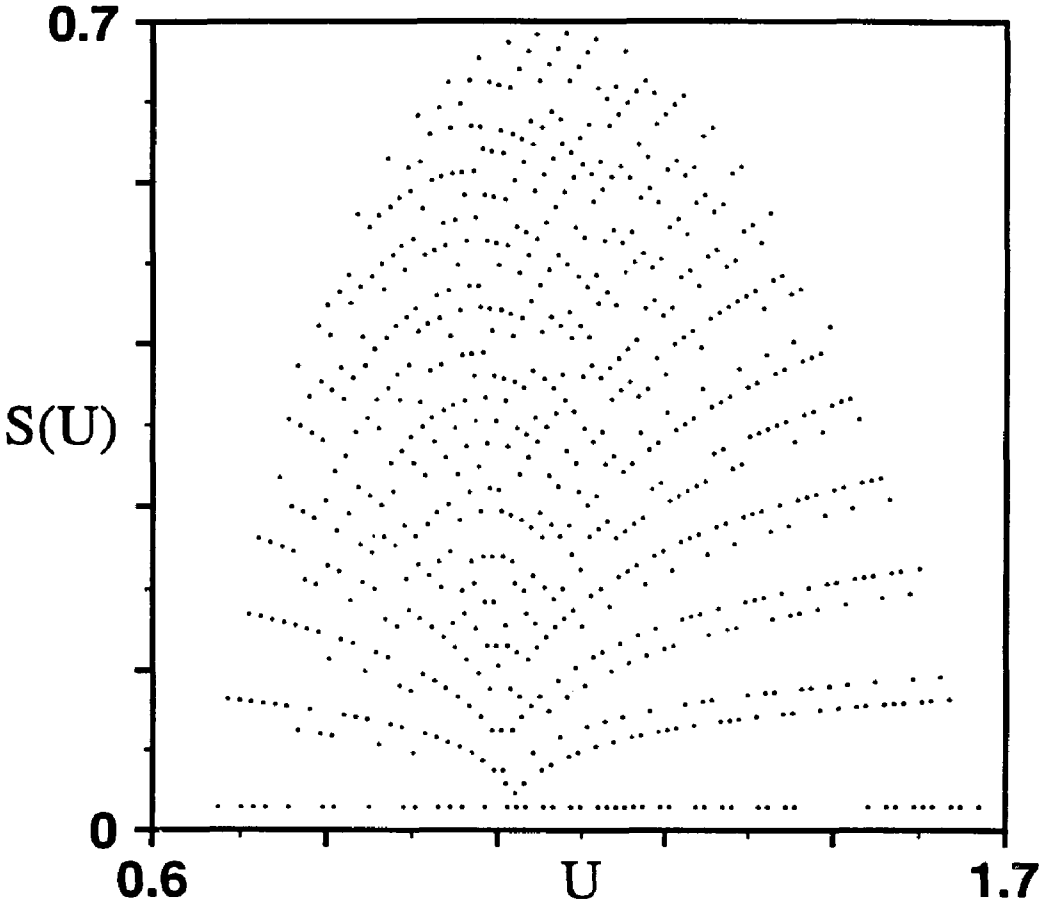


Figure 118

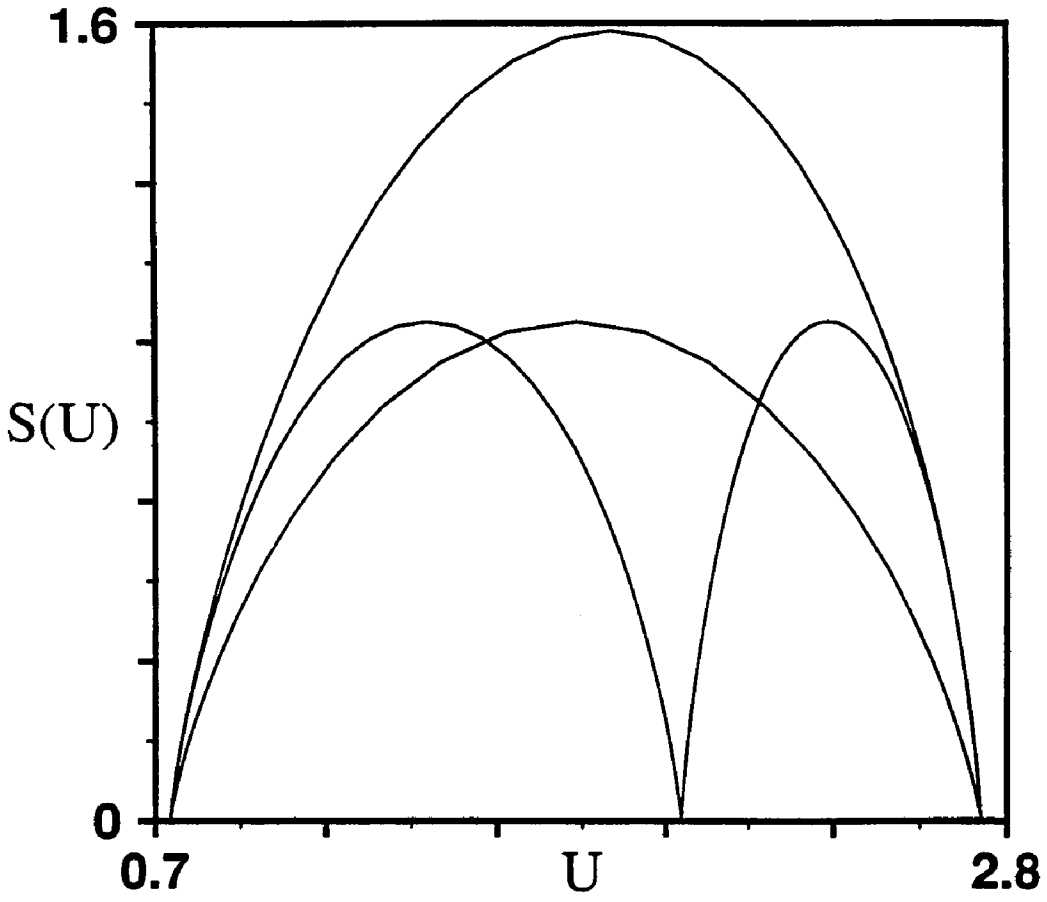


Figure 119

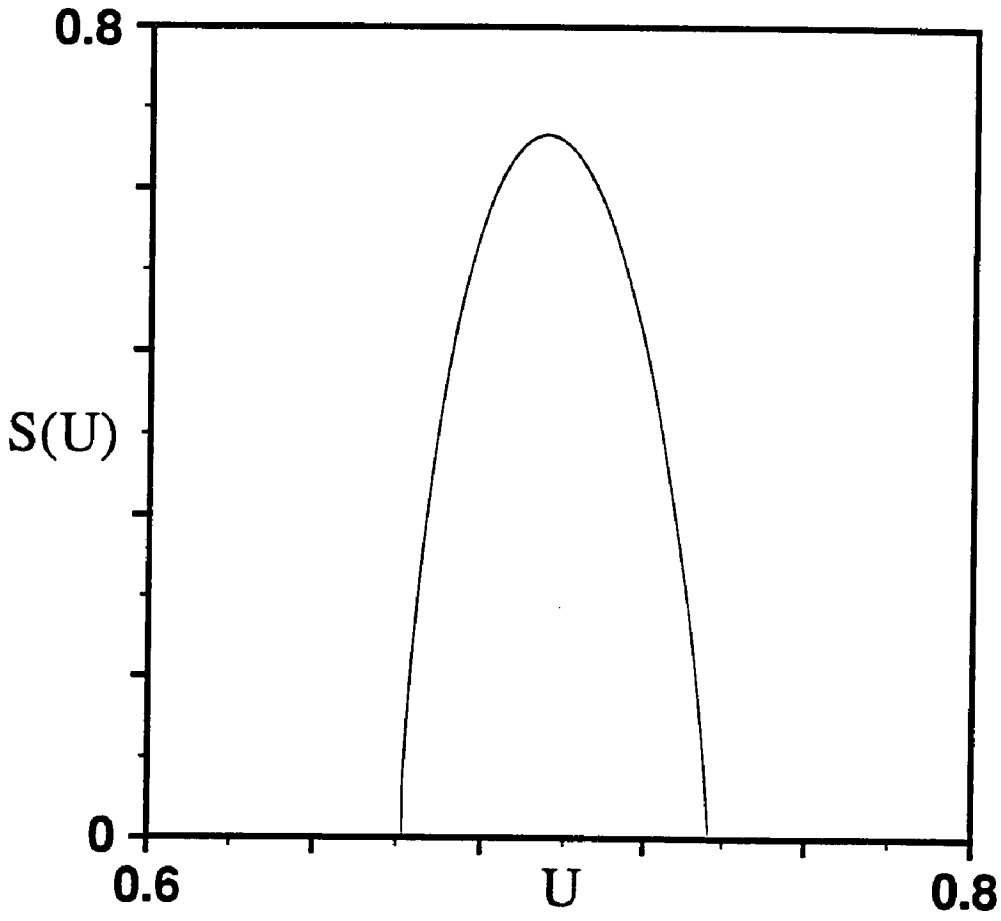


Figure 120

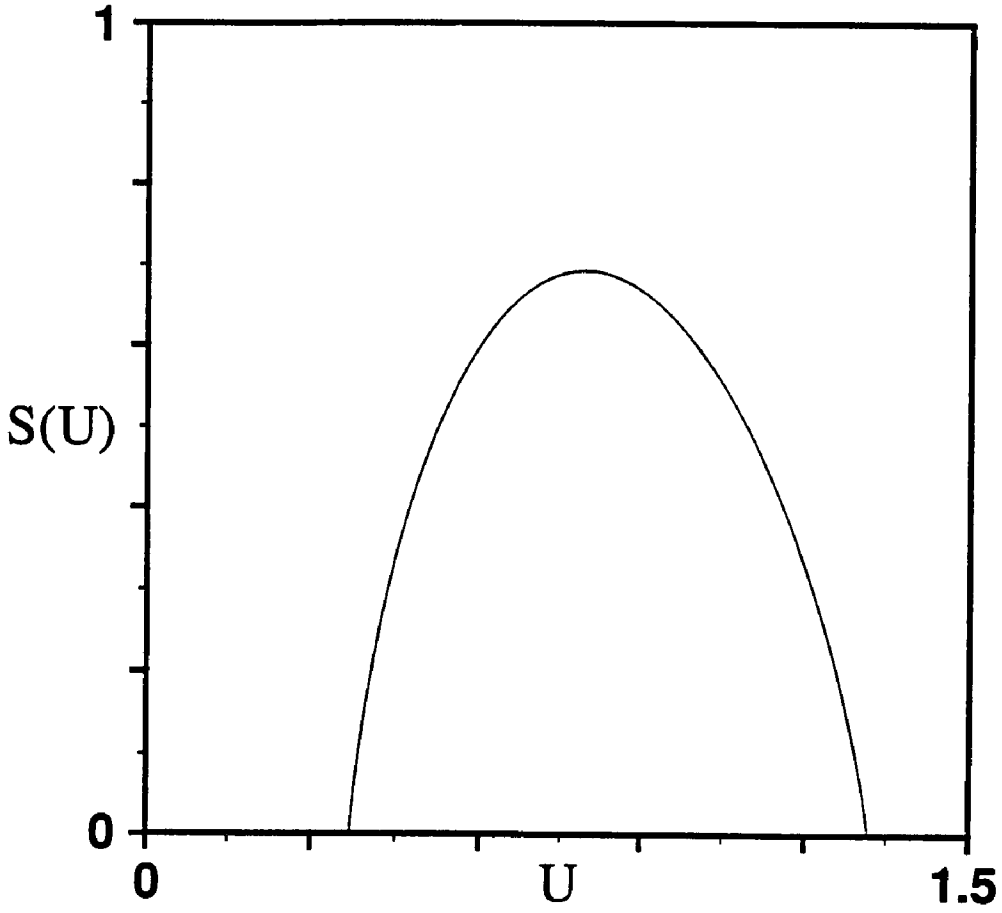


Figure 121

Bibliography

[1] H. Bai-Lin. *Elementary Symbolic Dynamics and Chaos in Dissipative Systems.* World Scientific, 1989.

[2] P. Billingsley. Statistical methods in markov chains. *Ann. Math. Stat.*, 32:12, 1961.

[3] R. E. Blahut. *Principles and Practice of Information Theory.* Addision-Wesley, 1987.

[4] J. A. Brzozowski. A survey of regular expressions and their applications. *IEEE Trans. on Electronic Computers*, 11:324, 1962.

[5] H.B. Callen. *Thermodynamics and an Introduction to Thermostatics.* John Wiley And Sons, New York, 1985.

[6] J. P. Crutchfield. Knowledge and meaning ... chaos and complexity. In L. Lam and H. C. Morris, editors, *Modeling Complex Systems*, Berlin, 1991. Springer-Verlag.

[7] J. P. Crutchfield and N. H. Packard. Symbolic dynamics of one-dimensional maps: Entropies, finite precision, and noise. *Intl. J. Theo. Phys.*, 21:433, 1982.

[8] J. P. Crutchfield and N. H. Packard. Symbolic dynamics of noisy chaos. *Physica*, 7D:201, 1983.

[9] J. P. Crutchfield and K. Young. Finitary ϵ-machines. in preparation, 1989.

[10] J. P. Crutchfield and K. Young. Finitary ϵ-machines: A sofic primer. in preparation, 1989.

[11] J. P. Crutchfield and K. Young. Inferring statistical complexity. *Phys. Rev. Let.*, 63:105, 1989.

[12] 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.

[13] D. C. Dennett. *The Intentional Stance.* MIT Press, Cambridge, Massachusetts, 1987.

[14] J.-P. Eckmann and D. Ruelle. Ergodic theory of chaos and strange attractors. *Rev. Mod. Phys.*, 57:617, 1985.

[15] R. S. Ellis. *Entropy, Large Deviations, and Statistical Mechanics*, volume 271 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, New York, 1985.

[16] J. W. Gibbs. *Elementary Principles of Statistical Mechanics*. Dover, New York, 1960.

[17] T.C. Halsey, M.H. Jensen, L.P. Kadanoff, I. Procaccia, and B.I. Shraiman. Fractal measures and their singularities: the characterization of strange sets. *Phys. Rev. A*, 33:1141, 1986.

[18] C. Kittel and H. Kroemer. *Thermal Physics*. W.H. Freeman and Company, San Francisco, 1980.

[19] D. A. Lind. Applications of ergodic theory and sofic systems to cellular automata. *Physica*, 10D:36, 1984.

[20] B. B. Mandelbrot. Fractals in geophysics. *Pure and App. Geop.*, 131, 1989.

[21] B. McMillan. The basic theorems of information theory. *Ann. Math. Stat.*, 24:196, 1953.

[22] L.E. Reichl. *A Modern Course in Statistical Physics*. University of Texas Press, Austin, 1988.

[23] A. Renyi. Some fundamental questions of information theory. In *Selected Papers of Alfred Renyi, Vol. 2*, page 526. Akademii Kiado, Budapest, 1976.

[24] D. Ruelle. *Thermodynamic Formalism*. Addision Wesley, Reading, MA, 1978.

[25] C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, Champaign-Urbana, 1962.

[26] P. Walters. *An Introduction to Ergodic Theory*. Springer-Verlag, New York, 1982.

[27] B. Weiss. Subshifts of finite type and sofic systems. *Monastsh. Math.*, 77:462, 1973.

[28] S. Wolfram. Computation theory of cellular automata. *Comm. Math. Phys.*, 96:15, 1984.