David Polant Feldman
September 1998
Physics

Computational Mechanics of Classical Spin Systems

Abstract

How does nature self-organize and how can scientists discover such organization? Is there an objective notion of pattern, or is the discovery of patterns a purely subjective process? And what mathematical vocabulary is appropriate for describing and quantifying pattern, structure, and organization? This dissertation compares and contrasts the way in which statistical mechanics, information theory, and computational mechanics address these questions.

After an in-depth review of the statistical mechanical, information theoretic, and computational mechanical approaches to structure and pattern, I present exact analytic results for the excess entropy and ϵ-machines for one-dimensional, finite-range discrete classical spin systems. The excess entropy, a form of mutual information, is an information theoretic measure of apparent spatial memory. The ϵ-machine—the central object of computational mechanics—is defined as the minimal model capable of statistically reproducing a given configuration, where the model is chosen to belong to the least powerful model class(es) in a stochastic generalization of the discrete computation hierarchy.

These results for one-dimensional spin systems demonstrate that the measures of pattern from information theory and computational mechanics differ from known thermodynamic and statistical mechanical functions. Moreover, they capture important structural features that are otherwise missed. In particular, the excess entropy serves to detect ordered, low entropy density patterns. It is superior in many respects to other functions used to probe the structure of a distribution, such as structure factors and the specific heat. More generally, ϵ-machines are seen to be the most direct approach to revealing the group and semigroup symmetries possessed by the spatial patterns and to estimating the minimum amount of memory required to reproduce the configuration ensemble, a quantity known as the *statistical complexity*. It is shown that the information theoretic and computational mechanical analyses of spatial patterns capture the intrinsic computational capabilities embedded in spin systems—how they store, transmit, and manipulate configurational information to produce spatial structure.

Finally, several approaches to generalizing the excess entropy and ϵ-machines to apply to multi-dimensional configurations are put forth. These measures are then calculated for a few simple, two-dimensional patterns.

Approved:

Ling-Lie Chau, Chair

Computational Mechanics of Classical Spin Systems

BY

DAVID POLANT FELDMAN
B.A. (Carleton College) 1991

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Physics

in the

OFFICE OF GRADUATE STUDIES
of the
UNIVERSITY OF CALIFORNIA
DAVIS

Approved:

Ling-Lie Chau, Professor of Physics, University of California at Davis, Chair

James P. Crutchfield, Research Professor, Santa Fe Institute

Steven Carlip, Professor of Physics, University of California at Davis

Committee in Charge

1998

Computational Mechanics of Classical Spin Systems

Abstract

How does nature self-organize and how can scientists discover such organization? Is there an objective notion of pattern, or is the discovery of patterns a purely subjective process? And what mathematical vocabulary is appropriate for describing and quantifying pattern, structure, and organization? This dissertation compares and contrasts the way in which statistical mechanics, information theory, and computational mechanics address these questions.

After an in-depth review of the statistical mechanical, information theoretic, and computational mechanical approaches to structure and pattern, I present exact analytic results for the excess entropy and ϵ-machines for one-dimensional, finite-range discrete classical spin systems. The excess entropy, a form of mutual information, is an information theoretic measure of apparent spatial memory. The ϵ-machine—the central object of computational mechanics—is defined as the minimal model capable of statistically reproducing a given configuration, where the model is chosen to belong to the least powerful model class(es) in a stochastic generalization of the discrete computation hierarchy.

These results for one-dimensional spin systems demonstrate that the measures of pattern from information theory and computational mechanics differ from known thermodynamic and statistical mechanical functions. Moreover, they capture important structural features that are otherwise missed. In particular, the excess entropy serves to detect ordered, low entropy density patterns. It is superior in many respects to other functions used to probe the structure of a distribution, such as structure factors and the specific heat. More generally, ϵ-machines are seen to be the most direct approach to revealing the group and semigroup symmetries possessed by the spatial patterns and to estimating the minimum amount of memory required to reproduce the configuration ensemble, a quantity known as the *statistical complexity.* It is shown that the information theoretic and computational mechanical analyses of spatial patterns capture the intrinsic computational capabilities embedded in spin systems—how they store, transmit, and manipulate configurational information to produce spatial structure.

Finally, several approaches to generalizing the excess entropy and ϵ-machines to apply to multi-dimensional configurations are put forth. These measures are then calculated for a few simple, two-dimensional patterns.

Computational Mechanics of Classical Spin Systems

For my Family:

Mom, Dad, Mark, Doreen, Monster[1], and Weasel[2]

[1] The larger of my two cats.

[2] The smaller of my two cats.

Contents

List of Figures

List of Tables

Acknowledgements

Without the assistance, support, and friendship of a great many people this dissertation would not have been possible. First, I extend thanks to my advisor, Jim Crutchfield. The work presented in here is the result of three and a half years of collaboration. Working with Jim has been a pleasure; I thank him for his guidance, insight, patience, and generosity.

I am grateful to Ling-Lie Chau for her support, showing confidence in my abilities, and encouraging me to find my own research path. I have benefitted greatly from her wisdom, her unbounded energy, and from her infectious enthusiasm for all areas of mathematical physics. I thank Steve Carlip for a careful reading of this manuscript and thoughtful advice at several stages of my graduate career.

Over the course of my research I have benefitted from interaction with a number of people: Lisa Borland, Daniel Cox, Deirdre Des Jardins, Rajiv R. P. Singh, and Gergely Zimányi. I have particularly benefitted from many conversations with Karl Young. Special thanks also are extended to Richard T. Scalettar: friend, mentor, and fountain of knowledge, statistical mechanical and otherwise. I am also grateful for the camaraderie and friendship of fellow graduate students: Brian Chiang, Matt Enjalran, Carey Huscroft, and Steve Lidia.

Much of the work in this dissertation was completed while at the Santa Fe Institute; I thank the institute for support and hospitality during my many visits.

While at Santa Fe, I benefitted from discussions with Erica Jen, Kristian Lindgren, Melanie Mitchell, Cris Moore, Mark Newman, Mats Nordahl, Cosma Shalizi, and Erik van Nimwegen. I especially thank Kristian Lindgren for a careful, critical reading of chapter 14.

Much of this dissertation was produced using free software and I thank software developers who are committed to ensuring that their programs and source code are freely distributed. In particular, I thank the Free Software Foundation and the writers of: the numerical algebra program Octave; the emacs editor; the LaTeX typesetting program; the gnuplot plotter; the newmat c++ matrix class; the gnu c++ complier gcc; and the Linux operating system.

Finally, I thank my family for unwavering support and love, and especially my wife, Doreen Stabinsky, for her companionship, understanding, and kindness.

This work was supported at UC Berkeley by ONR grant N00014-95-1-0524, at the Santa Fe Institute by ONR grant N00014-95-1-0975, and at UC Davis by a President's Dissertation Year Fellowship and a Chancellors Teaching Fellowship.

Part I

Introduction and Background

Chapter 1

Introduction

        the news to my left over the dunes and
reeds and bayberry clumps was
        fall: thousands of tree swallows
        gathering for flight:
        an order held
        in constant change: a congregation
rich with entropy: nevertheless, separable, noticeable
    as one event,
            not chaos: preparations for
flight from winter,
cheet, cheet, cheet, cheet, wings rifling the green clumps,
beaks
at the bayberries
    a perception full of wink, flight, curve,
    sound:
    the possibility of rule as the sum of rulelessness:
the "field" of action
with moving, incalculable center:

A. R. Ammons

in *Corson's Inlet* [1]

1.1 A Congregation Rich with Entropy

A. R. Ammons' poem *Corson's Inlet* [1] poses the central questions that this thesis aims to address. In the poem, the poet shares his thoughts as he walks along the beach by Corson's Inlet, about 30 miles south of Atlantic City, on the coast of New Jersey. Above, Ammons observes a flock of swallows moving from bayberry bush to bayberry bush, gathering energy for their fall migration southward.

The flock of swallows, Ammons notes, is both ordered and disordered at the same time. The flock is not static—it is an "order held in constant change." Within the loose boundaries of the flock the birds move about almost at random. The flock is "a congregation rich with entropy," and yet, not "chaos;" the flock is a noticeable event, a single entity. The poet observes the flock, but has a little difficulty compactly describing it—it is a mixture of opposites: order and chaos.

This passage is not the only place in the poem where Ammons observes the mixing of opposites, orders "held in constant change." Looking at the sand dunes, he notes that "manifold events of sand / change the dune's shape that will not be the same shape / tomorrow ... ." Earlier in the poem he writes that

```
in nature there are few sharp lines:  there are areas of
primrose
        more or less dispersed;
disorderly orders of bayberry: ...
```

To Ammons, nature is a mixture of order and disorder. Ample evidence presents itself throughout his walk, but one gets the sense that Ammons is groping, somewhat, as he tries to put these observations into words. His searching for words is particularly apparent when Ammons considers the boundaries between the dunes and creek:

```
by transitions the land falls from grassy dunes to creek
to undercreek:  but there are no lines, though
     change in that transition is clear
     as any sharpness:  but ‘‘sharpness’’ spread out ... .
```

Returning to the passage above, note that Ammons tries to account for the disorderly order of the flock by suggesting "the possibility of rule as the sum of rulelessness." There is significant randomness in the behavior of the individual birds, but nevertheless these individuals combine to produce a distinct flock. It's not the case, he presumes, that the birds are all following some "chief" bird. Rather, they're each obeying more or less their own commands, while trying to stay somewhat close to their neighbors.

By "rule as the sum of rulelessness," Ammons suggests that the disorder is not fighting against the pattern, but rather that the disorder plays an essential role in creating the pattern. The randomness is not noise, but is a part of the pattern. The flock of swallows is not a perfectly ordered pattern perturbed by a small amount of noise. The flock is not a "noisy" version of an army marching in lock step.

Put another way, there is a great deal of randomness responsible for patterns Ammons sees as he takes his walk. For example, in the ecology of Corson's Inlet Ammons sees:

```
                              ... the working in and out, together
and against, of millions of events:
...
                            not predictably (seeing me gain
the top of a dune,
the swallows
could take flight --- some other fields of bayberry
     could enter fall
     berryless) ...
```

Despite the clear transition between dune and creek, and the easily recognizable flock of swallows, Ammons recognizes that there is randomness to the natural orders that he sees, and that the forms he sees arise not despite of this randomness, but *because of* it. The processes that produce the patterns contain elements of randomness. More importantly, the patterns themselves contain disorder as well as order.

In summary, Ammons perceives many forms and patterns as he walks along Corson's Inlet. Yet these shapes are not the orderly shapes of Euclidean, or even fractal geometry; rather, they are "rich with entropy." Moreover, this disorder is not noise on top of the patterns he sees—the entropy is a consequence of the pattern itself. There is nothing atypical about what Ammons sees in Corson's Inlet. Nature is replete with intricate mixtures of order and disorder of the sort Ammons describes.

1.2 Central Questions

Ammons's observations on natural patterns, and his searching for words to describe them, leads naturally to the questions I consider in this thesis. These questions fall into three broad areas. First, what is a *pattern* [35]? An initial response might be that a pattern is some observed regularity or repeated tendency. But can any general method for discovering these regularities be given? Or is the discovery of patterns a fundamentally subjective process? For some time now there has been considerable interest in "pattern-forming" systems (see, e.g., Ref. [159] and references therein). But what exactly does "pattern" mean in this setting? Who is to say what patterns are and who determines which systems have generated patterns and which have not? Can we meaningfully compare two different patterns? Moreover, as Ammons reminds us, many natural patterns are only approximate. So how do we manage to separate pattern from mere noise? Presumably, we also have to consider the possibility that noise is part of the pattern. Is there some way to formalize what a noisy pattern is?

A second area of questions concern *organization*: what does it mean to say that a system is organized? In statistical mechanics, *order* is often associated with a broken symmetry. For example, the Ising model orders by acquiring a net magnetization when the spin-flip symmetry is broken. Can we define a similarly general notion of organization? Can we distinguish between different types of organization? There has been much effort expended recently to study "self-organizing" systems (see, e.g., Refs.[8, 67, 113] and references therein). But who is to say which systems are organized and which are not? The organized-nonorganized distinction is very crude. Can't there be degrees of organization? How would one say that one system is more organized than another?

A third set of questions revolves around information processing: how can we detect the computation being performed by a physical process or by other natural systems, such as the immune system or the visual cortex, in which pattern recognition, decision-making, and the like are (ostensibly) the central functions of the underlying dynamical behavior? In a condensed matter system, for example, how must spatial information be stored and shared so that it can reach a critical state? How much historical memory is required to produce a given configuration? How do raw dynamical degrees of freedom support computation—the storage, transmission, and manipulation of information?

Are pattern, organization, and computation related in any way? The central thesis of this work is that they are intimately related and that inquiring about a system's computational capabilities is a concrete way to address questions of pattern and organization [33]. Computation, pattern, and organization are related in that they are all statements about the relationships within and between a system's components and behavior. Restated in a more direct way, the hypothesis is simply that analyzing how a process "computes"—stores historical information, transmits it between internal degrees of freedom, and uses it to produce future behavior and system configurations—reveals how it is organized and what types of patterns it generates.

Establishing this hypothesis requires adopting a particular stance. Throughout this work the central issue is *discovering*, as opposed to *verifying*, pattern. The verification that a configuration displays one of a certain, *a priori* selected set of symmetries is not at issue here though it is, admittedly, an important concern (see, e.g., Refs. [66] and [129] and references therein). Rather, the goal is to determine what organization and which *kind* of patterns are intrinsic to a process: What in a process's configurations and temporal behavior indicates how it is organized?

Faced with analyzing a system of many interacting components, it is usually necessary to resort to a statistical description of some sort. A statistical approach also becomes necessary when considering the trajectories followed by a chaotic dynamical system. To capture essential aspects of such systems, statistical analyses typically entail calculating some average property: temperature, compressibility, Lyapunov exponents, escape rates, and so on. However, these are certainly not the only quantities about which one can ask. In this work I consider some of the more detailed, yet still statistical quantities that one can measure in a many-body setting and that indicate a process's degree of organization.

Statistical mechanics has a very limited set of tools for discovering and quantifying structure, pattern, information processing, and memory in physical systems. It is my contention that to satisfactorily address the questions posed above some tools must be added to the statistical physicist's tool-box. This dissertation reviews and adapts techniques and concepts from information and computation theories that will enable us to address questions of memory, structure, organization, and pattern. It applies these techniques to simple statistical mechanical systems to show that a richer set of tools is available for discovering pattern and describing organization in many-body systems.

1.3 Some Historical Context

Historically, the issues of pattern and organization have been the province of spatially extended many-body systems, as analyzed by the theory of critical phenomena, to mention one approach. More recently, though, many of the same questions have arisen in the conundrum of deterministic chaotic dynamical systems: simple, but nonlinear processes produce unpredictable, seemingly random behavior. Physics has long possessed a measure of the uncertainty of a probabilistic system—namely, the Shannon entropy [30, 142] of the underlying distribution. The Shannon entropy, introduced over 100 years ago by Boltzmann, was adapted in the '50's by Kolmogorov [85] and Sinai [147] to the study of dynamical systems. This, in turn, formed the foundation for the statistical analysis of deterministic sources of apparent randomness in the late '60's through the early '80's. These efforts to describe the randomness of a dynamical system have been rather successful. The metric entropy, Lyapunov exponents, and fractal dimensions form a widely applicable set of tools for detecting and quantifying unpredictable behavior; see, e.g., Refs. [12, 121].

Since this time, however, it has become more broadly understood that a system's randomness and unpredictability fail to capture its patterns and correlational structure. This realization has led to a considerable effort to develop a general measure or set of measures that quantify the structure of a system and the patterns it generates [5, 13, 33, 43, 44, 52, 60, 61, 64, 74, 96, 104, 128, 132, 144, 152, 158, 155, 168]. These quantities are often referred to as "complexity" measures. More properly, they should be called "structural complexity" or "statistical complexity" measures to distinguish them from Kolmogorov-Chaitin complexity [94], a measure of randomness, and computational complexity [123], a measure of resource (run time or storage) requirements in the theory of algorithms.

In this thesis I shall consider two approaches to measuring structure. First, we'll see how information theory provides a measure of the memory stored in a system's configurations. To date, this information theoretic measure of memory and related quantities have been estimated for the symbolic dynamics of chaotic dynamical systems [43, 46, 58, 144, 152], cellular automata [64, 104], stochastic automata [96], spin systems [3, 38, 54], and hidden Markov models [33, 154].

Second, I'll examine how the architectural analysis of information processing provided by computation theory can be used to describe structure more completely than by using information theory or, for that matter, statistical mechanics. By using a hierarchical approach that begins with the least computationally powerful model classes, it is possible to infer the computation being performed by the system. This approach, an extension of statistical mechanics that includes elements of statistical inference and computation theory, is known as *computational mechanics.*

For a more detailed discussion of the motivations and central issues that underlie computational mechanics, the reader is referred to Refs. [32, 33, 35]. Computational mechanics has been applied to the period-doubling and quasiperiodic routes to chaos [44, 45], the dripping faucet [62], one-dimensional cellular automata [70, 71], globally coupled maps [48], recurrent hidden Markov models [33, 154], and stochastic resonance [165]. Computational mechanics has also been proposed [131] as a useful tool with which to re-examine the learning paradox of developmental psychology that concerns the discovery of new patterns, not seen before [35].

1.4 Possible Applications

There are three overlapping areas of application of the tools for discovering and quantifying pattern, computation, and organization developed here. First, the methods should be of benefit when considering small-scale physical systems as the basis of useful information-processing devices [50]. Along the same line, the information theoretic approach to memory might help clarify issues surrounding the "memory" observed in systems with charge density waves [29] or with glassy dynamics [57].

Second, it is likely that the structures that emerge in the canonical models of many-body systems (e.g., the Ising, XY, and Heisenberg models) can be analyzed more thoroughly through the use of computational mechanics and information theory. These model systems have formed the basis for much of our understanding of critical phenomena. Thus, it seems natural to reexamine these models by applying the computational and information theoretic apparatus discussed here.

Third, statistical mechanical techniques are now being applied to a wide range of "nontraditional" systems, such as self-organized criticality [8], genetic algorithms [156, 157], traffic flow [115, 116], and learning dynamics in neural networks [139, 162]. Extant quantities in statistical mechanics have been influenced by the observability constraints of physical experiments. For the most part, only directly measurable quantities such as the pressure, conductivity, or net magnetization have been thoroughly developed. However, for some of these more "exotic" systems such measurability constraints may not be limitations, since the microstates themselves can be directly observed. In these cases one need not carry forward the traditional constraints, especially when new structural questions require different quantities to be estimated.

Information theory and computational mechanics provide a richer set of tools for studying these sorts of systems. Of course, the most revealing and meaningful quantities will always depend on the specific features of the system under study. It is not my intention to argue for *one* particular way to measure organization or pattern. Rather, I suggest that to fully capture patterns and organization in a wide range of many-body systems, the probes offered by statistical mechanics fall short; concepts and methods from information and computation theories become necessary.

1.5 Overview

The presentation is organized as follows. Part I contains an introduction to the central questions posed, and reviews the approaches to pattern and organization discussed here. Specifically, Chapter 1 (the current chapter), poses the central questions this thesis attempts to answer. In Chapter 2 I review the basic statistical mechanical approaches to detecting and quantifying features in many-body systems. I also use this section to fix the notation and context that I will assume for the rest of the development. Chapters 3 and 4 review information theory. Chapter 3 defines and motivates the key quantities of information theory: the Shannon entropy, joint and conditional entropies, mutual information, and information gain. Chapter 4 applies these information theoretic quantities to an infinite chain of random variables. This leads to the entropy density, a measure of the irreducible randomness of the system, and the excess entropy, a form of mutual information that provides a measure of the apparent spatial memory of the process. Chapter 5 then gives a concise, but self-contained, review of computational mechanics. When reviewing each of these three different approaches to pattern and structure, I begin by considering the central questions that motivate each approach. We shall see how the quantities introduced arise as natural answers to these questions. Awareness of the different motivating issues is crucial to understanding the differences and similarities between the three views of organization and pattern.

In Part II I report the results of applying the different measures of structure to finite-range one-dimensional spin systems. Chapter 6 contains a brief review of the statistical mechanics of one-dimensional spin systems. In Chapter 7 I present methods for calculating the excess entropy and ϵ-machines for one-dimensional, finite-range spin systems. I then embark on a comparison of information theoretic, statistical mechanical, and computational mechanical approaches to pattern and structure. In Chapter 8 I compare the excess entropy with the structure factors of statistical mechanics. We will see that the excess entropy is capable of detecting periodic structure of any periodicity and thus may be viewed as an "all-purpose order parameter" for periodic patterns. In Chapter 9, I show that ϵ-machines are necessary to describe the structure of entropic patterns that do not have a strong periodic component. In so doing, I illustrate how an ϵ-machine provides an "irreducible representation" of an approximate symmetry. In Chapter 10 I directly compare the excess entropy with a number of commonly used measures in statistical mechanics: the correlation length, specific heat, ferromagnetic structure factors, and the nearest-neighbor correlation function. I argue that while there are qualitative similarities between all these functions, none can be viewed as a measure of memory in the sense that the excess entropy can be. Furthermore, I find that all these functions are maximized at different parameter values, indicating that they are not trivially related and that the statistical mechanical functions cannot be used to determine the parameter values at which a system's spatial memory is maximized.

In Part III I present a critical analysis of a statistical complexity measure introduced in Ref. [106]. Specifically, I show that this statistical complexity measure vanishes in the thermodynamic limit for all finite-memory ergodic Markov chains, regardless of whatever structural features these systems possess. This class of systems includes all finite-range one-dimensional discrete spin systems.

In Part IV I present some preliminary work on extending the information and computation theoretic approaches to more than one spatial dimension. In Chapter 12 I introduce the some of the questions surrounding pattern and organization in two dimensions. In this chapter I also review the statistical mechanics of the two-dimensional Ising model, and briefly discuss critical phenomena in general, including critical exponents and universality. In Chapter 13 I present exact results for the nearest-neighbor two-spin mutual information for the two-dimensional Ising model. In chapter 14, I discuss several proposals for extending information theory and computation mechanics to apply to two-dimensional configurations.

I conclude in Part V by summarizing the main results of the dissertation and discussing several open questions.

Some of the text presented in this dissertation has appeared elsewhere. In particular, chapters 2, 4, 5, and 7-10 have been published (with some slight modifications) as Ref. [54]. Portions of Ref. [54] have also been incorporated into chapters 1, 3, and 15. Much of Chapter 4 was taken from a set of (unpublished) notes I distributed as part of a series of lectures at the Santa Fe Institute in July, 1997. Chapter 11 has been published as Ref. [55].

Chapter 2

Review of Statistical Mechanics

We begin this chapter by reviewing equilibrium statistical mechanics. Statistical mechanics is, of course, a very rich and deep field, and thus we cannot review it thoroughly here. Refs. [27, 130, 136, 170] are good texts that provide an up-to-date account of statistical mechanics. In Sec. 2.2 we'll fix the notation used to describe spin systems and will define some of the central quantities of the statistical mechanics of spin systems: The thermodynamic entropy, the partition function, the Helmholtz free energy, and the magnetization. In Sec. 2.3 we introduce some of the functions often used to detect structure in spin systems: the correlation function and correlation length, the susceptibility and structure factors, and the specific heat. Finally, in Sec. 2.4 we discuss how, as a practical matter, it is necessary to look for structure in the Boltzmann distribution in order to perform the necessary sums to evaluate the partition function.

We shall also review some basic statistical mechanics in Chapter 6, in which we discuss the one-dimensional Ising model; in Sec. 12.3 the two-dimensional Ising model is review. Statistical mechanics is also treated in Secs. 12.4 and 12.5 where the modern theory of critical phenomena and the notion of universality are very briefly reviewed.

2.1 The Boltzmann Distribution

A central concern of equilibrium statistical mechanics is determining how physically observable, bulk quantities can be explained from the behavior of the system's constituents. For example, how are the conductivity, heat capacity, and compressibility of a metal determined by the interactions between the electrons and nuclei that make up that metal?

The starting point for such calculations is a knowledge of the microphysics—typically, the Hamiltonian for the system expressed as a sum or an integral over the system's internal degrees of freedom. The connection between the energy determined by the Hamiltonian and the joint probability over the internal degrees of freedom is given by

$$\Pr(\mathcal{C}) \propto e^{-\beta \mathcal{H}(\mathcal{C})} , \tag{2.1}$$

where $\mathcal{C}$ is a configuration of the system and $\mathcal{H}$ is the system's Hamiltonian. The quantity $\beta = 1/(k_B T)$ is the inverse temperature and k_B is Boltzmann's constant.

For the remainder we set k_B equal to one.

In principle, given a Hamiltonian one can use Eq. (2.1) to calculate macroscopically observable average quantities. However, performing the necessary sums is usually prohibitively difficult, a consideration we shall return to at the end of this chapter. Before we begin examining how statistical mechanics goes about discovering and quantifying structure, we pause to establish some notation and set the context for the following development.

2.2 Spin Systems: Notation and Definitions

The main object of our attention will be a one-dimensional chain $\overset{\leftrightarrow}{S} \equiv \ldots S_{-2} S_{-1} S_0 S_1 \ldots$ of spins (random variables) S_i's that range over a finite set $\mathcal{A}$. For a spin-K system, $|\mathcal{A}| = 2K + 1$. Alternatively, one may also consider the chain as being a stationary time series of discrete measurements or the symbolic dynamics arising from the partitioning of a dynamical system. In Part IV we discuss extending our analysis—including the information and computation theoretic measures of structure—to more than one spatial dimension. For the time being, however, we shall restrict ourselves to one-dimensional distributions that are time independent. That is, we consider only equilibrium distributions. If we imposed some time dependence—say a Glauber dynamics or an update rule for a one-dimensional cellular automaton—then we would need to include a time index on all the spin variables.

We divide the chain into two semi-infinite halves by choosing a site i as the dividing point. Denote the left half by

$$\overleftarrow{S}_i \equiv \ldots S_{i-3} S_{i-2} S_{i-1} \tag{2.2}$$

and the right half by

$$\overrightarrow{S}_i \equiv S_i S_{i+1} S_{i+2} S_{i+3} \ldots \ . \tag{2.3}$$

We will assume that a spin system is described by a spatial shift-invariant measure μ on bi-infinite configurations $\cdots s_{-2} s_{-1} s_0 s_1 s_2 \cdots ; s_i \in \mathcal{A}$. The measure μ induces a family of distributions that will be of primary interest. Let $\Pr(s_i)$ denote the probability that the i^{th} random variable S_i takes on the particular value $s_i \in \mathcal{A}$ and $\Pr(s_{i+1}, \ldots, s_{i+L})$ the joint probability over blocks of L consecutive spins. We assume spatial translation symmetry; $\Pr(S_{i+1} = s_i, \ldots, S_{i+L} = s_L) = \Pr(S_1 = s_1, \ldots, S_L = s_L)$. Let a block of L consecutive spin variables be denoted by $S^L \equiv S_1 \ldots S_L$. We shall follow the convention that a capital letter refers to a random variable, while a lower case letter denotes a particular value of that variable. Thus, s^L denotes a particular spin-block configuration of length L. Finally, in the following we shall use the term *process* to refer to the joint distribution over the bi-infinite chain of variables. For a more rigorous definition of the term process, see Ref. [154].

We now define the Hamiltonians we shall use to generate equilibrium distributions of our spin chain. A general Hamiltonian for a one-dimensional chain of N

spins that interact in pairs is given by

$$\mathcal{H}(s^N) = -\sum_{i,j=1}^{N} J_{ij} s_i s_j - B \sum_{i=1}^{N} s_i \ , \tag{2.4}$$

where, as usual, the J_{ij}'s are parameters determining the strength of coupling between spins, B represents an external field, and for a spin 1/2 system, $s_i \in \{+1, -1\}$ or $s_i \in \{\uparrow, \downarrow\}$. Below, we shall consider only interactions within a finite range $R < \infty$; that is,

$$J_{ij} = 0, \ |i - j| > R \ . \tag{2.5}$$

We shall also consider only coupling constants that are translationally invariant; i. e., those depending only on $|i - j|$ and not i and j individually. Hence, we define:

$$J_r = J_{ij} \ , \tag{2.6}$$

where $r = |i - j|$. Despite these restrictions on J_{ij}, the quantities discussed below are perfectly general and apply to any lattice system.

The canonical partition function for these spin systems is defined by

$$Z_N = \sum_{\{s^N\}} e^{-\beta \mathcal{H}(s^N)} \ . \tag{2.7}$$

The sum is understood to extend over all $|\mathcal{A}|^N$ possible configurations of length N.

The average internal energy U is simply the expectation value of the Hamiltonian and can be expressed as:

$$U = -\frac{1}{N}\frac{\partial \log Z_N}{\partial \beta} \ . \tag{2.8}$$

The free energy F per site is given by

$$F = -\frac{T}{N}\log Z_N \ . \tag{2.9}$$

In the thermodynamic limit, in which the system size N goes to infinity, Z_N typically diverges exponentially so F remains finite.

The thermodynamic entropy is defined as the logarithm of the number of microstates accessible at a given energy. In the canonical ensemble, the entropy per site $\mathbf{S}$ is related to the free energy per site F via:

$$\mathbf{S} = -\frac{\partial F}{\partial T} \ . \tag{2.10}$$

Finally, the magnetization m per site is defined as the average:

$$m \equiv \frac{1}{N}\langle \sum_{i=1}^{N} s_i \rangle \ , \tag{2.11}$$

where here and below angular brackets indicate thermal expectation value;

$$\langle \bullet \rangle \equiv \frac{1}{Z_N} \sum_{\{s^N\}} \bullet \, e^{-\beta \mathcal{H}(s^N)} \, . \tag{2.12}$$

For spin systems with a Hamiltonian of the form of Eq. (2.4),

$$m = -\frac{\partial F}{\partial B} \, . \tag{2.13}$$

2.3 Statistical Mechanical Measures of Structure

2.3.1 Correlation function and correlation length

With the above notational preliminaries out of the way, we consider our first statistical mechanical measure of "structure": the *two-spin correlation function* Γ_{ij}, defined in the usual way as

$$\Gamma_{ij} \equiv \langle (s_i - \langle s_i \rangle)(s_j - \langle s_j \rangle) \rangle \, . \tag{2.14}$$

This quantity is sometimes called the truncated or connected correlation function to distinguish it from $\langle s_i s_j \rangle$. It follows from translation invariance that $\langle s_i \rangle = \langle s_{i+k} \rangle, k = 1, 2, \ldots$. This enables us to write the correlation function as

$$\Gamma_{ij} = \langle s_i s_j \rangle - \langle s \rangle^2 \, , \tag{2.15}$$

where $\langle s\rangle \equiv \langle s_j\rangle$. Thus, Γ_{ij} measures the tendency of the fluctuations (about the mean value) of spins at site i and at site j to be correlated with one another.

Again from translation invariance it follows that $\langle s_i s_j\rangle = \langle s_{i+k} s_{j+k}\rangle$, $k = 1, 2, \ldots$. And so, the correlation function depends only on $r = |i - j|$ and not on i and j individually. This leads one to define:

$$\Gamma(r) \equiv \langle s_0 s_r\rangle - \langle s\rangle^2 \ . \tag{2.16}$$

Except at a critical point, the correlations die exponentially with increasing r; that is,

$$\Gamma(r) \sim e^{-r/\xi} \ \text{ as } \ r \to \infty \ . \tag{2.17}$$

The quantity ξ is called the *correlation length*. Simply stated, it measures the range of influence of a single spin. Equivalently, ξ gives the size of a typical ordered cluster of spins. An infinite correlation length typically indicates that the correlation function dies algebraically, rather than exponentially. This occurs at the critical points of continuous (second or higher order) phase transitions.

2.3.2 Susceptibility and structure factors

The *magnetic susceptibility* χ per site is defined as a measure of the system's linear change dm in magnetization per site due to the application of a small external field

dB. That is,

$$dm = \chi dB \; . \tag{2.18}$$

Thus,

$$\chi = \frac{\partial m}{\partial B} = -\frac{\partial^2 F}{(\partial B)^2} \; . \tag{2.19}$$

As is always the case with linear response functions [12, 17], χ can be written as a sum of correlation functions;

$$\chi = \lim_{N\to\infty} \frac{\beta}{N} \sum_{i,j=1}^{N} \Gamma_{ij} \; . \tag{2.20}$$

We can exploit the translation invariance of Γ_{ij} to perform one of the sums above. We then obtain:

$$\chi = \beta \left[2 \sum_{r=0}^{\infty} \Gamma(r) - \Gamma(0) \right] \; . \tag{2.21}$$

This expression for χ can be reconciled with its definition, Eq. (2.18), by realizing that, roughly speaking, the magnetization is more changeable with a variation in field dB the greater the correlations between spin pairs.

Eq. (2.20) tells us that χ is a sum over correlation functions and as such might serve as a global measure of structure. In particular, consider the term $\sum_{r=0}^{\infty} \Gamma(r)$, the sum over all possible two-spin correlation functions, from Eq. (2.21). At first blush, this seems to be an ideal quantity to use as an indicator of structure. By summing over all two-spin correlation functions χ appears to provide a measure of the total correlations across the lattice.

However, this turns out not to be the case. To see this, consider a system near an antiferromagnetic-paramagnetic transition. Clusters of ordered spins appear at all length scales, but the type of order within a cluster is antiferromagnetic—alternating up and down spins. Thus, the correlation functions $\Gamma(r)$ for such a system will alternate in sign with r and will tend to cancel each other out, resulting in a small quantity despite the presence of a strong antiferromagnetic ordering. To compensate for this, one could choose, for example, to multiply each term in the sum by $(-1)^r$. But this is a somewhat arbitrary adaptation to a particular set of spin couplings that derives ultimately from our own appreciation of the underlying order.

Instead, we can take the Fourier transform of spin configurations. The result is a function that is usually called the *structure factor*. It is given by

$$S(q) \equiv \sum_{r=0}^{\infty} e^{irq}\Gamma(r) \; . \tag{2.22}$$

The structure factor provides a measure of the correlation with a particular spatial periodicity, as measured by the wavenumber q. As an observable, $S(q)$ is important for both simulation and laboratory experiments. In a simulation it is often $S(q)$ that is calculated to look for a phase transition: an $S(q)$ that diverges as a function of system size is a clear indication of critical behavior. In the laboratory, order in a magnetic system is often probed by means of neutron scattering. Assuming dipole interactions and fixed target spins, the probability for scattering to occur with a momentum transfer q is proportional to $S(q)$; see, e.g., [17]. Neutron scattering is used, for example, to distinguish between a paramagnet and an antiferromagnet. Both types of materials have zero magnetization, but their magnetic structural properties are distinct.

Any transform (integral or discrete) carries with it representational restrictions that are implicit in its choice of function basis. For example, an interpretation of $S(q)$, as with all Fourier analysis, carries an assumption that the underlying order is a linear superposition of periodic configurations. Hence, as we shall see, $S(q)$ is not suited to detect aperiodicity. Moreover, it is sometimes the case that a particular choice of function basis results in an unnecessarily "large" description; for example, a Fourier decomposition of a square wave yields an infinite number of nonzero amplitudes.

Unfortunately, there is no universally accepted way to define a structure factor. One alternative is to define

$$\tilde{S}_1(q) \equiv \beta S(q) . \tag{2.23}$$

so that the susceptibility is more closely related to the structure factor: $\chi = 2\tilde{S}_1(0) - \beta\Gamma(0)$. Another alternative is to argue that if the structure factor is to measure correlation between spins, the "self-correlation" term $\Gamma(0)$ should be excluded from the sum; yielding

$$\tilde{S}_2(q) \equiv \sum_{r=1}^{\infty} e^{iqr}\Gamma(r) . \tag{2.24}$$

Both modifications of the structure factor do not significantly alter the features of its behavior reported in chapters 8-10. As such, we shall focus our attention on $S(q)$ as defined in Eq. (2.22).

2.3.3 Specific heat

We conclude this brief review of statistical mechanical measures of structure by commenting on the specific heat. The specific heat C is a linear response function defined by

$$dU = CdT \ , \tag{2.25}$$

where U is the internal energy. Like χ, C can be related to fluctuations—in this case, energy fluctuations:

$$C = \beta^2 \langle (U - \langle U \rangle)^2 \rangle \ . \tag{2.26}$$

As a result, C measures fluctuations in energy, not in correlations between spins. To see this, consider a paramagnet, a spin system in which there are no couplings between the spins and so the spin variables are independently distributed. The specific heat for such a system is nonzero, reaching a maximum in the $T \approx B$ region. That C is nonzero for this system—a clearly correlationless paramagnet—indicates that $C > 0$ is at best a misleading measure of spatial structure.

2.4 Searching for Structure is Implicit in the Practice of Statistical Mechanics

In the previous section we reviewed some basic quantities often used in statistical mechanics to detect and measure the presence of correlational structure. But there are other, more subtle ways in which the search for structure enters into statistical mechanics than in the use of its typical observables.

A calculation of (say) the partition function by explicitly considering all allowed configurations is infeasible for all but the smallest of systems. It is quite often the case, however, that the probability distribution to be summed over has symmetries or internal structure that render large portions of the sum in Eq. (2.7) redundant. Thus, one central challenge of statistical mechanics is to find these symmetries and figure out how to best exploit them.

As a simple example of the discovery and exploitation of symmetries consider again the paramagnet. Since the spins do not interact, the energy of the system depends only on how many spins are up, say. Equivalently, the probability distribution of a single spin is independent of the others. Due to this particularly simple symmetry in the joint probability distribution over spin configurations, thermodynamic averages may be calculated by using the binomial theorem, rather than a brute force enumeration of all possible configurations.

A less trivial example of the "covert" role of structure in statistical mechanics is found in the technique of transfer matrices. For one-dimensional systems with finite-range interactions, such as the one-dimensional Ising models considered here, the partition function can be re-expressed in terms of the dominant eigenvalue of this finite-dimensional matrix. Moreover, the joint probabilities over spin configurations follow from the dominant left and right eigenvectors. Hence, all thermodynamic averages can be determined given knowledge of the transfer matrix. Loosely speaking, the transfer matrix encodes all of the information about the system. In chapters 6-7 we shall discuss transfer matrix methods in more detail.

Unfortunately, the transfer matrix method does not always work, often failing for systems with disorder or long-range interactions. It is only successful for systems whose joint probability distribution over configurations factors in a certain way: namely, the distribution over the spin chain must decompose into independent distributions over contiguous spin blocks of finite size. Said another way, the stationary stochastic process generating the chain must be a finite-memory Markov process.

When the transfer matrix method fails, sometimes it is possible to use an infinite dimensional matrix, i. e., an operator [140]. Another approach is the diagrammatic perturbation expansions of statistical field theory where one or several fundamental interactions are identified and their contributions to the thermodynamic quantities in question are summed up by considering more and more complicated interactions [17, 124].

Yet another approach to finding and utilizing structure in the joint probability distributions over configurations relies on cycle expansion methods [47, 108]. Here one systematically approximates the partition function by considering the contributions from fundamental periodic configurations of successively longer periods. A particularly effective and elegant application of the cycle expansion technique is the calculation of the Lyapunov exponent of a product of random matrices [107].

The vantage point afforded by this brief overview suggests classifying statistical mechanical systems by considering the type of mathematical entity—contiguous blocks of variables, operators, fundamental interactions, cycles—needed to most efficiently "encode" their configurations so that calculations of thermal averages can be performed. In chapter 5 we shall see that the ϵ-machines of computational mechanics provide a formalization of this idea.

Chapter 3

Information Theory I: Shannon Entropy and its Variants

In this chapter and the following, we explore how information theory can be used to measure a process's unpredictability and memory capacity. We begin with a brief historical review of the questions that motivated the development of information theory. Then, the remainder of this chapter will focus on some of the key quantities of information theory as applied to one or two random variables. Subsequently, in chapter 4, we examine how these information theoretic functions are applied to stationary sequences of random variables.

3.1 Historical Introduction

To appreciate the interpretation and use of information theoretic concepts in the comparisons that we develop in the following, an historical review is helpful. This will be, of necessity, brief. The interested reader is strongly advised to read basic reference works such as Refs. [30] [142].

In the late 1940's Shannon founded the field of communication theory [142], motivated in part by his work in cryptography during World War II. This led to a study of how signals could be compressed and transmitted efficiently and error free. His basic conception was that of a *communication channel* consisting of an *information source* that produces messages which are encoded and passed through the channel. A receiver then decodes the channel's output in order to recover the original messages. An essential component of his analysis was the definition of the source's rate of information production, called the *source entropy rate*, and the maximum carrying capacity, called the *channel capacity*, of the (possibly noisy and error-prone) channel.

Much earlier, Hartley had proposed to measure the amount of information from a source via the logarithm of the number of possible source messages [72]. Shannon's definition of the source entropy adapted Hartley's measure to account for probabilistic structure in the source: some messages being more or less likely than others. He interpreted the negative logarithm of a message's probability as a measure of surprise: the more unlikely a message the more informative it was when it appeared. This surprise, averaged over a source's messages, is the source's entropy rate. The functional form of Shannon's entropy, as he realized, had already been developed by Boltzmann in late 1800's as a measure of disorder of thermodynamic systems [19]. In the following we will refer to this and related quantities as Shannon entropy, however, since it will be used in the sense intended by information theory.

It is important to emphasize that the core of information theory concerns not so much the various definitions of information and entropy, but rather the relationship between the source entropy rates that can be sustained through channels and those channels' capacities. These connections are what makes the similarity between Boltzmann's notion of thermodynamic entropy and Shannon's entropy rate so notable. Boltzmann clearly did not anticipate Shannon's use of entropy.

The primary results on which information theory is built and with which it finds its technological applications are Shannon's two central coding theorems. This first theorem says that information cannot be transmitted error-free through a channel at a rate higher the channel's capacity. The second theorem says that as long as the source's rate respects this limit then there exists an encoding and decoding scheme for the source's messages such that error-free transmission is possible and can occur at rates arbitrarily close to the channel capacity. These theorems will be restated more technically in Sec. 4.5.

The mathematical foundations of Shannon's communication theory followed quickly [111, 81], as did a number of applications and important extensions. For example, Jaynes re-introduced portions of information theory back into statistical mechanics, reformulating ensembles in terms of a maximum (Shannon) entropy assumption under various constraints [76]. This was partly motivated as an attempt to understand the role of probability in statistical mechanics and the similarities between statistical mechanics and statistical inference [31]. For a readable introduction to this approach to statistical mechanics see Ref. [7]; a more thorough account can be found in [63].

The basic quantities used in information theory are various forms of Shannon entropy: the entropy H of a distribution, the information gain $\mathcal{D}$ of one distribution with respect to another, and the mutual information I between two distributions. When adapted and applied to different communication problems, these are the quantities in which the results of the theory are expressed. It is noteworthy that many uses of information theory in statistical physics and in nonlinear dynamics mostly employ its basic quantities and do not use the more characteristic and central aspects of coding. The remainder of this chapter is concerned with motivating, defining, and discussing these three quantities: H, $\mathcal{D}$, and I.

3.2 Shannon Entropy

Throughout, we shall use capital letters to indicate a discrete random variable, and lowercase letters to indicate a particular value of that variable. For example, let X be a random variable. The variable X may take on the values $x \in \mathcal{X}$. Here $\mathcal{X}$ is the (finite) set of all possible values for X and is referred to as the *alphabet* of X.

The probability that X takes on the particular value x is written $\Pr(X = x)$, or just $\Pr(x)$. We may also form joint and conditional probabilities. Let Y be another random variable with $Y = y \in \mathcal{Y}$. The probability that $X = x$ and $Y = y$ is written $\Pr(X = x, Y = y)$, or $\Pr(x, y)$ and is referred to as a joint probability. The conditional probability that $X = x$ given $Y = y$ is written $\Pr(X = x|Y = y)$ or simply $\Pr(x|y)$.

3.2.1 Entropy as Uncertainty

The use of probabilities to describe a situation implies some uncertainty. If I toss a fair coin, I don't know what the outcome will be. I can, however, describe the situation with a probability distribution: $\{\Pr(\text{Coin} = \text{Heads}) = 1/2, \Pr(\text{Coin} = \text{Tails}) = 1/2\}$. If the coin is biased, there is a different distribution: e.g., $\{\Pr(\text{BiasedCoin} = \text{Heads}) = 0.9, \Pr(\text{BiasedCoin} = \text{Tails}) = 0.1\}$.

All probability distributions are not created equal. Some distributions indicate more uncertainty than others; it is clear that we are more in doubt about the outcome of the fair coin than the biased coin. The question before us now is: can we make this notion of uncertainty or doubt quantitative? That is, can we come up with some function that takes a probability distribution and returns a number that can be interpreted as a measure of the uncertainty associated with events governed by that distribution.

We proceed by considering what features such a measure of uncertainty should have, a line of reasoning first put forth by Shannon in his 1948 paper [142] that gave birth to information theory. For definiteness, we call this measure $H[X]$; H takes the probability distribution of X—$X = \{\, \Pr(x_1), \Pr(x_2), \cdots \Pr(x_N)\, \}, x_i \in \mathcal{X}$—and returns a real number. The picture here is that there are N possible values X can assume, and $\Pr(x_i)$ is the probability that X equals the i^{th} possible value, denoted x_i.

First, we surely want H to be maximized by a uniform distribution. After all, a uniform distribution corresponds to complete uncertainty. Everything is equally likely to occur — you can't get much more uncertain than that.

Second, it seems reasonable to ask that H is a continuous function of the probabilities. An arbitrarily small change in the probabilities should lead to an arbitrarily small change in H.

Third, we know that we can group probabilities in different ways. For example, consider a variable X with the following distribution

$$X = \{\, \Pr(X = a) = .5, \Pr(X = b) = .2, \Pr(X = c) = .3\, \} \; . \tag{3.1}$$

One way to view this distribution is that outcome C *or* B occurs half of the time. When it does occur, outcome B occurs with probability .4. That is:

$$X = \{\, \Pr(X\!=\!a) = .5, \Pr(X\!=\!Y) = .5, \} \quad Y = \{\Pr(Y\!=\!b) = .4, \Pr(Y\!=\!c) = .6\,\}\ . \tag{3.2}$$

We would like the uncertainty measure H not to depend on whatever grouping decisions we make. In other words, we want H to be a function of the distribution itself and not a function of how we group events within that distribution.

Remarkably, the above three requirements are enough to determine the form of $H[X]$ *uniquely* up to a multiplicative constant

3.2.2 Axiomatic Definition

Let's state the above three requirements more carefully and generally. Let $H[X]$ be a real-valued function of $\Pr(x_1), \Pr(x_2), \cdots, \Pr(x_N)$. We require that the following three statements hold:

1. $H[X]$ reaches a maximum when the distribution over X is uniform; $\Pr(x_i) = 1/N\ i = 1, 2, \cdots, N$.

2. $H[X]$ is a continuous function of the $\Pr(x_i)$'s.

3. The last requirement is awkward to write mathematically, but no less intuitive than the first two. As mentioned above, the idea is that we want $H[X]$ to

be independent of how we group the probabilities of individual events x_i into subsets. We follow the notation of Robertson [136]. Let the N probabilities be grouped into k subsets, y_k:

$$y_1 = \sum_{i=1}^{n_1} p_i \; ; \; y_2 = \sum_{i=n_1+1}^{n_2} p_i \; ; \; \ldots \tag{3.3}$$

Then, we require

$$H[X] = H[Y] + \sum_{j=1}^{k} y_j H[\{p_i/y_j\}_j] \; , \tag{3.4}$$

where the notation $\{p_i/y_j\}_j$ indicates that the sum extends over those p_i's that make up a particular y_j.

The above three requirements constrain the function H to be of the following form:

$$H[X] = -k \sum_{x \in \mathcal{X}} \Pr(x) \log \Pr(x) \; , \tag{3.5}$$

where k is an arbitrary positive constant [30, 136, 143]. The choice of constant amounts to a choice of units; we shall use base 2 logarithms and fix k at 1. The units of $H[X]$ for this choice of constant are called *bits* and are discussed below in Sec. 3.2.5.

Thus, we define the Shannon entropy of a random variable X by:

$$H[X] \; \equiv \; -\sum_{x \in \mathcal{X}} \Pr(x) \log_2 \Pr(x) \; . \tag{3.6}$$

The notation $H[X]$ can be misleading. $H[X]$ is *not* a function of X. It is a function of the *probability distribution* of the random variable X. The value of $H[X]$ does not depend on the values X assumes.

Note that the entropy is nonnegative. One can easily prove that

$$H[X] \geq 0 \ . \tag{3.7}$$

Also note that $H[X] = 0$ if and only if X is known with certainty: i. e., the probability of one outcome is 1 and the probability of all other outcomes is 0. (To show this one uses $\lim_{x \to 0} x \log_2 x = 0$.)

The axiomatic definition of H given above justifies the following statement: $H(p)$ is *the* quantitative measure of the amount of uncertainty associated with a probability distribution p. But the story does not end here. There are many other ways we can view the Shannon entropy, and indication of the quantity's fundamental role in computer science and physics. This multiplicity of interpretations is at times a curse, however, since the different interpretations can lead to confusion. In the following several sections, we explore some of these additional interpretations, hoping to give the reader a sense of the range of ways in which the entropy can be used.

3.2.3 Shannon Entropy as Thermodynamic Entropy

It is not hard to see that the Shannon entropy is equivalent to the usual thermodynamic entropy,

$$\mathbf{S}(E) = \log N(E) \tag{3.8}$$

where $N(E)$ is the number of accessible microstates as a function of energy E. Equilibrium statistical mechanics postulates that microstates of equal energy are assumed to be equally likely; the probability of the i^{th} microstate x_i occurring is

$$\Pr(x_i) = \frac{1}{N(E)}, \quad \forall\, i \;. \tag{3.9}$$

Plugging Eq. (3.9) into Eq. (3.6), we see immediately that the thermodynamic entropy, Eq. (3.8) results.

This connection with thermodynamics that led Shannon to call his uncertainty measure "entropy" [142]. Legend has it that he was encouraged to do so by John von Neumann, who said that since "no one really understands what entropy is," calling his new measure "entropy" would give Shannon "a big edge in the debates" [56, p. 123].

3.2.4 Shannon Entropy as Average Surprise

Here is another way to view Eq. (3.6): The quantity $-\log_2 \Pr(i)$ is sometimes referred to as the *surprise* associated with the outcome i. If $\Pr(i)$ is small, we would be quite surprised if the outcome actually was i. Accordingly, $-\log_2 \Pr(i)$ is large for small $\Pr(i)$. And if $\Pr(i)$ is large, we see that the surprise is small.

Thus, we may view Eq. (3.6) as saying that $H[X]$ is the expectation value of the surprise;

$$H[X] = \sum_{x \in \mathcal{X}} \{-\log_2 \Pr(x)\} \Pr(x) \; = \; \langle -\log_2 \Pr(x) \rangle \, , \tag{3.10}$$

where the angular brackets indicate expectation value;

$$\langle \bullet \rangle \; \equiv \; \sum_{x \in \mathcal{X}} \bullet \Pr(x) \; . \tag{3.11}$$

The entropy tells us, on average, how surprised we will be if we learn the value of the variable X. This observation strengthens the assertion that $H(p)$ is a measure of the uncertainty associated with the probability distribution p. The more uncertain we are about an outcome, the more surprised we will be (on average) when we learn of the actual outcome.

We can also use this line of reasoning to see why H is referred to as a measure of information. Let us return to the example of a coin toss. Suppose I told you the outcome of the toss of a fair coin. This piece of information would be quite interesting to you, since before I told you the outcome you were completely in the dark. On the other hand, if it is the biased coin with a 90% probability of heads that is thrown, telling you the outcome of the toss is not as useful. “Big deal” you might say. “I was already pretty sure it was heads anyway; you really haven’t given me much information.” It is in this sense that $H[X]$ provides a measure of information. The greater $H[X]$, the more informative, on average, a measurement of X is.

3.2.5 Entropy and Yes-No Questions

Entropy is also related to how difficult it is to guess the value of a random variable. This is discussed rather thoroughly and clearly in chapter 5 of Ref. [30]. Here, I’ll just explain the general ideas qualitatively.

We begin with an example. Consider the random variable X with following distribution:

$$\left\{ \begin{array}{cc} \Pr(X = A) = 1/2, & \Pr(X = B) = 1/4, \\ \Pr(X = C) = 1/8, & \Pr(X = D) = 1/8 \end{array} \right\} . \qquad (3.12)$$

On average, how many yes-no questions are required to figure out the value of X? A reasonable first guess is $X = A$. This guess is right half of the time, and hence, half of the time only one question is needed to determine the outcome of X. If this initial guess is incorrect, one would then guess that $X = B$. Again, this question will yield a “yes” half of the time. So, half of the time one will need to make the $X = B$ guess and half of the time that guess will be correct. As a result, 1/4 of the time it will take two guesses to determine X.

If the $X = B$ guess was incorrect, one more guess is needed, say, $X = C$.

Regardless of the outcome of this guess, the value of X is now known, since if $X \neq C$, it must be that $X = D$. So, half of the time the $X = B$ guess will be required, and half of the time that guess will be wrong, necessitating the $X = C$ guess. Hence, 1/4 of the time 3 guesses are needed. Adding this up, we have:

$$\text{Average \# of Guesses} \;=\; \frac{1}{2}(1) \,+\, \frac{1}{4}(2) \,+\, \frac{1}{4}(3) \;=\; 1.75\;. \tag{3.13}$$

It turns out that the entropy of the distribution given in Eq. (3.12) is exactly equal to 1.75.

This is not a coincidence. One can show that [30]

$$H[X] \;\leq\; \text{Average \# of Yes} - \text{No Questions to Determine X} \;\leq\; H[X]+1\;. \tag{3.14}$$

This result assumes that the guesser is making optimal guesses. That is, roughly speaking at every guess, he or she tries to “divide the probability in half.” This is exactly the strategy we employed in the above example.

Eq. (3.14) might appear a little mysterious as first. As a slightly less mysterious example, consider another distribution:

$$\left\{\begin{array}{c} \Pr(Y=\alpha)=1/4, \quad \Pr(Y=\beta)=1/4, \\ \Pr(Y=\gamma)=1/4, \quad \Pr(Y=\delta)=1/4 \end{array}\right\}. \tag{3.15}$$

Clearly it will take an average of 2 guesses to determine the value of Y. The variable X is easier to guess because a lot of the probability is concentrated on $X = A$ and $X = B$, and we can exploit this in our guessing.

This idea of entropy as the average number of yes-no guesses is consonant with our earlier interpretation of entropy as a measure of uncertainty. The more uncertain we are about an event, the harder it is to guess the outcome.

3.2.6 Entropy and Coding

The Shannon entropy $H[X]$ is also related to the length of the minimal code for X. Speaking somewhat loosely, a code is a mapping between sets of symbols. For example, the Morse code maps the English alphabet—letters, punctuation and spaces—to dots, dashes and spaces. We are interested in the average size of the minimal code: the shortest coding scheme that is still uniquely decipherable. Put another way, we are interested in compressing an information source, not encrypting it. Encryption entails devising a mapping between symbols that is difficult to guess. Here, we assume that both the sender and the receiver are privy to the particular mapping used to encode the variables.

How does one devise an efficient code? The idea is to choose short code words for objects that occur most frequently. However, we also must make sure that the code is uniquely decodable, i.e., that we don't compress the original signal so much that it can't be recovered. As an example, consider again the distribution of Eq. (3.12) from the previous section. Here, it makes sense to use the shortest possible code for the event $X = A$ since it has the highest probability of occurring.

This business of choosing more probable events should be familiar—it's identical to the strategy we employed when we were trying to guess what X was; the process of yes-no guessing specifies a binary code [30]. Returning to the example of Eq. (3.12), recall our procedure for guessing the outcome of X and consider the sequence of questions that led up to our determining a particular value. To make our code, for each "yes" answer we'll use a 1 and for each "no" answer we'll use a 0. The result is the following code:

$$\begin{aligned} A &\longrightarrow 1 \\ B &\longrightarrow 01 \\ C &\longrightarrow 001 \\ D &\longrightarrow 000 \end{aligned} \tag{3.16}$$

For example, if we discovered that $X = B$ we would have gotten a "no" to our first questions and a "yes" to our second, corresponding to 01. Note that this encoding scheme is uniquely decodable. For example, if we observed 1100001, we know without any ambiguity that the original sequence was $AADB$.

Given this correspondence between yes-no questions and binary coding, we see that Eq. (3.14) implies that:

$$H[X] \leq \text{Average Length of Binary Code for X} \leq H[X] + 1 \; . \tag{3.17}$$

Before concluding this section, we state a slightly more technical result. Suppose one is encoding N identically distributed random variables X with a binary code. Then, in the $N \to \infty$ limit [30]:

$$\frac{1}{N}(\text{Average Length of Optimal Binary Code for X}) = H[X] \; . \tag{3.18}$$

This is a form of the famous Shannon source coding theorem, to be discussed again in Sec. 4.5.

Each digit in a binary code corresponds to one *bit*, a flip-flop memory device that can be in one of two positions. Thus, Eq. (3.18) tells us that $H[X]$ is the average number of bits needed to store the value of the random variable X.

3.2.7 Summary

In summary, the Shannon entropy $H[X]$ of a random variable X is the answer to many, seemingly different questions. In Sec. 3.2.5 we have seen that the average number of yes-no questions needed to guess the outcome of X lies between $H[X]$ and $H[X] + 1$. Equivalently, we have seen in Sec. 3.2.6 that $NH[X]$ is the average length of the binary code for N successive outcomes of X. Hence, $H[X]$ is the average amount of memory, in units of bits, needed to store the value of X.

In Sec. 3.2.2 I discussed how the Shannon entropy $H[X]$ can be shown to be the *unique* measure of the uncertainty associated with the outcome of X. This axiomatic treatment puts H on very firm footing; there is nothing arbitrary about the definition of the Shannon entropy. Furthermore, we've seen that Shannon's entropy function is identical to the thermodynamic entropy, an (indirectly) experimentally measurable state function of a statistical mechanical system.

3.3 Joint and Conditional Entropy

We continue this brief introduction to several key quantities of information theory by defining some variants of the entropy discussed above. I'll also state some of the properties of and relationships between these quantities.

First, the *joint entropy* of two random variables, X and Y, is defined in the natural way:

$$H[X,Y] \equiv -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \Pr(x,y)\log_2 \Pr(x,y) \ . \tag{3.19}$$

The joint entropy is a measure of the uncertainty associated with a joint distribution.

Next, we define the *conditional entropy*:

$$H[X|Y] \equiv - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \Pr(x,y) \log_2 \Pr(x|y) \ . \tag{3.20}$$

The conditional entropy measures the uncertainty associated with a conditional probability; $H[X|Y]$ is the expectation value of the conditional surprise, $-\log_2 \Pr(x|y)$ where the average is weighted by the *joint* distribution.

By writing $\Pr(x,y) = \Pr(x)\Pr(y|x)$ and taking the expectation value of the logarithm of both sides of this equation, we see that the joint entropy obeys the following, pleasing chain rule:

$$H[X,Y] \ = \ H[X] + H[Y|X] \ . \tag{3.21}$$

There are two noteworthy consequences of this observation. First, we may write

$$H[Y|X] = H[X,Y] - H[X] \ . \tag{3.22}$$

As $H[X] \geq 0$, we obtain the sensible result that conditioning reduces entropy. That is, knowledge of one variable can never increase our uncertainty about other variables. Second, Eq. (3.21) makes it quite clear that

$$H[Y|X] \ \neq \ H[X|Y] \ . \tag{3.23}$$

3.4 Mutual Information

We now turn our attention to mutual information. We define the *mutual information* $I[X;Y]$ of two random variables X and Y via:

$$I[X;Y] \equiv \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \Pr(x,y)\log_2 \frac{\Pr(x,y)}{\Pr(x)\Pr(y)} \ . \tag{3.24}$$

Some straightforward manipulations show us that

$$\begin{aligned} I[X;Y] &= H[X] - H[X|Y] & (3.25)\\ &= H[Y] - H[Y|X] & (3.26)\\ &= H[Y] + H[X] - H[X,Y] \ . & (3.27) \end{aligned}$$

The above shows quite clearly that $I[X;Y] = I[Y;X]$.

Eq. (3.25) shows us why I is called the mutual information; we see that mutual information between two variables is the reduction in uncertainty of one variable due to knowledge of another. If knowledge of Y reduces our uncertainty of X, then we say that Y carries information about X.

Looking at Eq. (3.24), it's not hard to see that $I[X;Y]$ vanishes if X and Y are independently distributed; $\Pr(x,y) = \Pr(x)\Pr(y)$. Also, we see that the mutual information between two variables vanishes if both variables have zero entropy. The mutual information is positive in between these two extremes.

3.5 Information Gain

The Shannon entropy of source X measures the average uncertainty of observing outcomes x if we expect the outcomes to occur with probability $\Pr(x)$. But what if, despite the actual events occurring according to $\Pr(X)$, we have prior knowledge that leads us to expect the outcomes are distributed with probability $Q(x)$? The relative information obtained in observing X is then given by the *information gain* $\mathcal{D}(P|Q)$:

$$\mathcal{D}(P|Q) \equiv \sum_{x \in \mathcal{A}: Q(x) > 0} P(x) \log_2 \frac{P(x)}{Q(x)} \ , \tag{3.28}$$

where we assume that if $Q(x) = 0$, then $P(x) = 0$.

The quantity $\mathcal{D}$ is often referred to as a distance, but it is neither symmetric in P and Q nor does it obey a triangle inequality. It is, however, nonnegative, and is zero only when the two distributions are equal. $\mathcal{D}(P|Q)$ is, in a sense, the number of bits it takes to change distribution P into Q. Note that $\mathcal{D}(\mathrm{P}(x)|U) = \log_2 |\mathcal{A}| - H[X]$, where U is the uniform distribution.

Chapter 4

Information Theory II: Entropy Density and Entropy Convergence

We now shift the emphasis back to analyzing spin configurations, or, more generally, a stationary chain of random variables. In this chapter we apply the information theoretic quantities introduced in the previous chapter to the spin chain. By considering the manner in which the entropy of the spin chain grows as more spins are considered, we'll be led to two important quantities: the entropy density and the excess entropy. The latter is especially noteworthy since it can be interpreted as a measure of the process's memory.

4.1 Entropy Growth

Consider again the bi-infinite sequence $\ldots S_{-2}S_{-1}S_0S_1S_2\ldots$. The average uncertainty of observing an L-spin block S^L is given by the Shannon entropy of the joint distribution $\Pr(s^L)$ [30]:

$$H(L) \equiv - \sum_{s^L \in \mathcal{A}^L} \Pr(s^L) \log_2 \Pr(s^L) \ . \tag{4.1}$$

We define $H(0) \equiv 0$ and, for later use, $H(L) = 0, L < 0$. The block entropy is nonnegative, $H(L) \geq 0$, and monotonic in L; $H(L) \leq H(L+1)$. That is, adding an additional random variable cannot reduce uncertainty [30]. A schematic plot of $H(L)$ vs. L is shown in Fig. (4.1) for a typical information source.

4.2 Entropy Density and Entropy Density Convergence

The spatial density of the Shannon entropy of the spin configurations is defined by

$$h_\mu \equiv \lim_{L \to \infty} \frac{H(L)}{L} \ , \tag{4.2}$$

where μ denotes the measure over bi-infinite configurations that induces the L-block joint distribution $\Pr(S^L)$. The quantity h_μ measures the irreducible randomness in spatial configurations: the randomness that remains after the correlations and

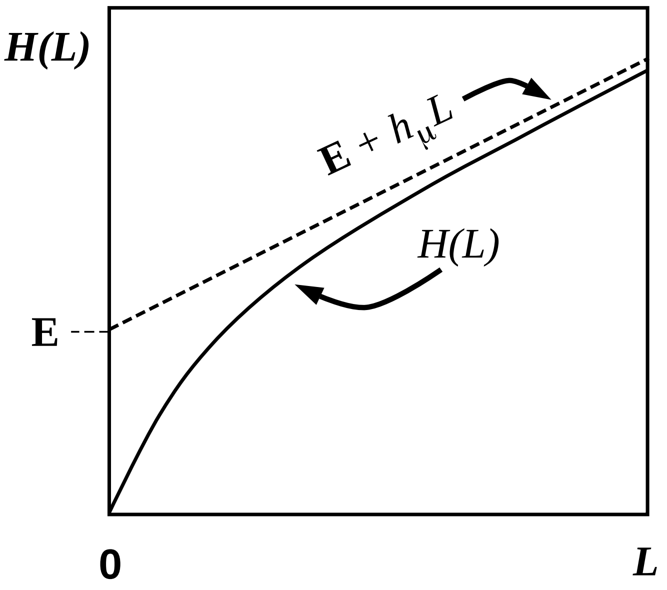


Figure 4.1: Total Shannon entropy growth for a typical information source: a schematic plot of $H(L)$ versus L. $H(L)$ increases monotonically and asymptotes to the line $\mathbf{E} + h_\mu L$, where $\mathbf{E}$ is the excess entropy and h_μ is the source entropy rate.

structures in larger and larger spin blocks are taken into account. For physical systems h_μ is equivalent to thermodynamic entropy density—$\mathbf{S}$ in Eq. (2.10)—in units where $k_B / \log_e 2 = 1$. The entropy density is also known as the entropy rate or the metric entropy, depending on the application context.

The entropy density h_μ can be re-expressed as:

$$h_\mu = \lim_{L \to \infty} \left[H(L+1) - H(L) \right] \ . \tag{4.3}$$

Thus, we see that the curve's slope as $L \to \infty$ in Fig. (4.1) corresponds to the entropy density h_μ.

Eq. (4.3) can also be rewritten by using the conditional entropy as defined in Eq. (3.20) [30]:

$$\begin{aligned} h_\mu &= \lim_{L\to\infty} H[S^L|S^{L-1}] \\ &= \lim_{L\to\infty} H[S_L, S_{L-1}, \cdots, S_1|S_{L-1}, S_{L-2}, \cdots, S_1] \\ &= \lim_{L\to\infty} H[S_L|S_1 \ldots S_{L-1}] \ . \end{aligned} \tag{4.4}$$

Thus, h_μ is the uncertainty of the next spin value s_L conditioned on the first $(L-1)$ spins in the L-block, as $L \to \infty$. This reinforces the interpretation of h_μ as the irreducible randomness associated with the system. Eq. (4.4) indicates that h_μ measures our uncertainty about the variable S_L given knowledge of *all* the spins that preceded it. In this sense h_μ measures, in units of bits per site, the per-spin unpredictability of the infinite string. Note that the entropy density is nonnegative; $h_\mu \geq 0$.

Eqs. (4.2), (4.3), and (4.4) give different expressions for the entropy density h_μ. These are all equivalent in the present setting, though they need not be for nonequilibrium or nonstationary processes [65].

The entropy density is a property of the system as a whole; only in special cases will the isolated-spin uncertainty $H(1)$ be equal to h_μ. This leads us to consider how random the spin chain appears when finite-length spin blocks are considered.

This finite-length apparent randomness is given by

$$h_\mu(L) \equiv H(L) - H(L-1),\ L = 1, 2, \ldots\ , \tag{4.5}$$

the incremental increase in uncertainty in going from $(L-1)$-blocks to L-blocks. Thus, since we've imposed the "boundary condition" $H(0) = 0$, we have $h_\mu(1) = H(1)$.

Comparing Eq. (4.5) with Eqs. (4.3) and (4.4), we see that $h_\mu(L)$ may be viewed as the finite-L approximation to the entropy density h_μ. Graphically, $h_\mu(L)$ is the two-point slope of the $H(L)$ vs. L curve; in other words, $h_\mu(L)$ is the discrete derivative of $H(L)$. The convergence of $h_\mu(L)$ to h_μ is illustrated in Fig. (4.2). The entropy density h_μ is indicated by a horizontal dashed line.

4.3 Density, Rate, and Algorithmic Complexity

Coming back to the issue of describing the observations of spin configurations we note that the entropy density h_μ is equivalent to the growth rate of the Kolmogorov-Chaitin (KC) complexity of spin configurations, averaged over a given ensemble [30, 94]. The KC complexity of an individual configuration is defined as the length of the minimal program that, when run, will cause a universal Turing machine (UTM) to produce the configuration and then halt. The KC complexity is sometimes referred to as the algorithmic (or deterministic) complexity because it demands a deterministic accounting

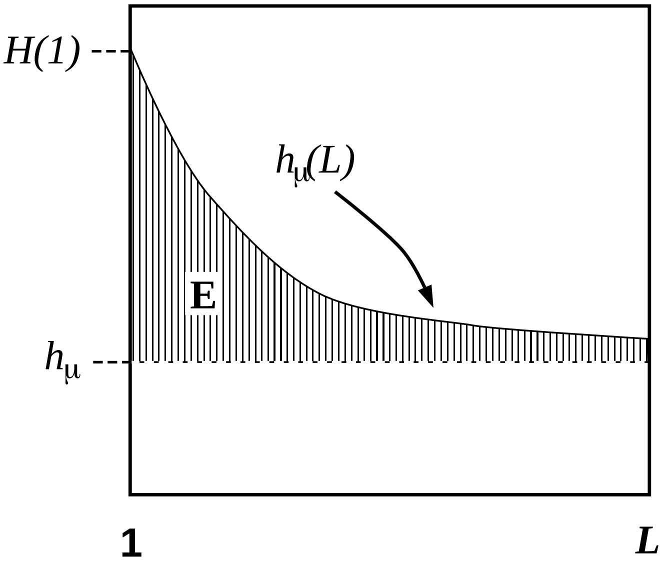


Figure 4.2: Entropy density convergence: A schematic plot of $h_\mu(L)$ versus L using the typical $H(L)$ shown in Fig. (4.1). The entropy density asymptote h_μ is indicated by the horizontal dashed line. The shaded area is the excess entropy $\mathbf{E}$.

for every spin in the configuration. A random configuration by definition possesses no regularities so it cannot be compressed. As a result, a random configuration's shortest description is the configuration itself. Hence, we see that the KC complexity is maximized by random configurations, as is the entropy density h_μ.

4.4 Redundancy

If $h_\mu < \log_2 |\mathcal{A}|$, the full information carrying capacity of the alphabet is being underutilized. Said in a complementary fashion, in this case the information source produces sequences that have correlations. One measure of these correlations is the *redundancy* [142]:

$$\mathcal{R} \equiv \log_2 |\mathcal{A}| - h_\mu \ . \tag{4.6}$$

There is no redundancy in a completely random source, since by definition such a source has $\Pr(s^L) = U(s^L)$, $L = 1, 2, \ldots$, and so $h_\mu = \log_2 |\mathcal{A}|$, where U is the uniform distribution.

4.5 Shannon's Coding Theorems

As mentioned above in Sec. 3.2.6, loosely speaking, the sequence of yes-no questions leading to the identification of a particular outcome x of the random variable X defines a code for that outcome. One can show that the average (per symbol) length of the optimal, uniquely decodable binary encoding for the information source X lies between h_μ and $h_\mu + 1$ [30]. If one tries to encode N copies of the variable X, the average length of the code approaches $N h_\mu$ as $N \to \infty$. Thus, the entropy rate h_μ of the random variable X can also be interpreted as the average number of bits of memory needed to store information about the values X takes [30, 142].

This view of entropy as average code length is in harmony with the notion of entropy as uncertainty. If we are very uncertain about the outcome of an observation, on average it will take a long code word to specify the outcome when it occurs. If we are fairly certain what the outcome will be, we can take advantage of this knowledge by using short code words for the frequently occurring outcomes. This strategy is employed in Morse code, where the most frequently occurring English letter “E” is encoded using the shortest symbol, one “dot”. Put somewhat colloquially, then, the entropy rate measures the average length required to describe observations of a random variable.

In order to state Shannon’s coding theorems more carefully, we need first to state the definition of a channel’s information carrying capacity. If the source is denoted X and the output of the channel is denoted Y, then the channel capacity C is defined as

$$C = \sup_{\{X\}} I[X;Y] \ , \tag{4.7}$$

where the supremum is understood to be taken over all information sources X—i.e., over all distributions of X. The picture here is that the sender transmits the signal X which is received as Y by the receiver. As a particularly simple example, consider a noiseless binary channel in which $X, Y \in \{0, 1\}$ and Y is determined by X; the signal is transmitted without errors. Thus, $H[Y|X] = 0$ and $C = 1$ bit, where the supremum occurs when $\Pr(X = 0) = \Pr(X = 1) = 1/2$ [30].

Now that the entropy density (or rate) and the channel capacity have been defined we can quickly mention Shannon’s coding theorems again in order to show the utility of the various entropies just discussed and also to highlight one difference in motivation between information theory and statistical mechanics. The first coding theorem states that if $h_\mu > C$, the information source cannot be transmitted without errors. The second says that if $h_\mu < C$, then there exists an encoding of the source messages that produces a new source whose rate is less than, but arbitrarily close to C. And so, by the first theorem, the source messages can be carried error free in a noisy channel of capacity C and correctly decoded. Exactly how one finds these encoding schemes is not specified by the theory, though many techniques have been developed since information theory's introduction.

Before continuing, we mention an additional example of a communication channel which serves to illustrate the importance and non-triviality of Shannon's coding theorems. Consider now a binary symmetric channel [30, pp. 186-7]. Here the signals X and Y are binary, and there is a probability p that any particular bit of the signal X is flipped. That is, $\Pr(Y = 0|X = 1) = \Pr(Y = 1|X = 0) = p$. This is an example of a noisy channel. Indeed, no single bit is reliable—each symbol in the message X has a non-zero probability of being transmitted wrong (i.e., flipped). Is it possible to reliably transmit any message using such a communication channel? Remarkably, Shannon's theorems let us conclude that such a channel can be used reliably.

It is not hard to show that the channel capacity for the binary symmetric channel is given by $1 - H[F]$ [30, p. 187], where $H[F] \equiv -p \log_2 p - (1-p) \log_2 (1-p)$ is the entropy associated with the probability that a single bit is flipped. Shannon's second coding theorem then states that any source with an entropy rate $h_\mu < 1 - H[F]$ can be encoded such that the new source (the encoding) can be transmitted error free along this communication channel. This is quite a noteworthy result; Shannon's theorem tells us that this noisy channel, which is, in some sense, completely unreliable, can nevertheless be used to transmit signals error free.

4.6 Excess Entropy

The entropy density h_μ measures the per-spin unpredictability of infinite configurations. However, h_μ says little about how difficult it is to perform this prediction. For example, consider two periodic configurations: one of period 4 and one of period 1969. Both have zero entropy density, indicating that once the periodic pattern is gleaned there is no uncertainty about the subsequent spins. But there are important (and obvious) differences between the two configurations. It seems clear that, in some sense, the period-1969 configuration is "harder" to predict than the period-4 configuration; a distinction that is missed by stating $h_\mu = 0$. For example, one would imagine that the configuration with the longer period requires much more memory to predict than that with the short period. How can we formalize the notions of "memory" and "difficulty" of prediction? For the remainder of this chapter and the following one, we shall be concerned with stating this question more clearly and then answering it.

We begin our consideration of memory by observing that the length-L approximation to the entropy density $h_\mu(L)$ overestimates the entropy density h_μ. Specifically, $h_\mu(L)$ overestimates h_μ by an amount $h_\mu(L) - h_\mu$ that measures how much more random single spins appear knowing the finite L-block statistics than knowing the statistics of the infinite configurations $\overleftrightarrow{S}$. In other words, this excess randomness tells us how much additional information must be gained about the configurations in order to reveal the actual per-spin uncertainty h_μ. More precisely, the difference $h_\mu(L) - h_\mu$ is a form of redundancy, as discussed in section 4.4 above. Though the source appears more random at length L by this amount, this amount is the information-carrying capacity in the L-blocks that is not actually random, but is due instead to correlations. We conclude that entropy convergence is related to a type of memory.

There are many alternative ways in which the finite-L approximations $h_\mu(L)$ converge to their asymptotic value h_μ. Recall Fig. (4.2). Fixing the values of $H(1)$ and h_μ, for example, does not determine the form of the $h_\mu(L)$ curve. At each L we obtain additional information about how $h_\mu(L)$ converges, information not contained in the values of $H(L)$ and $h_\mu(L)$ at smaller L. Thus, roughly speaking, each $h_\mu(L)$ is an independent indicator of the manner in which $h_\mu(L)$ converges to h_μ.

Given that each increment $h_\mu(L) - h_\mu$ is an independent contribution in the sense just described, we sum up the individual L-redundancies to obtain our candidate measure of memory. The resulting quantity is the *total excess entropy* [64, 43, 152, 144, 104, 96, 99]:

$$\mathbf{E} \equiv \sum_{L=1}^{\infty} [h_\mu(L) - h_\mu] \ . \tag{4.8}$$

Graphically, $\mathbf{E}$ is the shaded area in Fig. (4.2). If one inserts Eq. (4.5) into Eq. (4.8), the sum telescopes and one arrives at an alternate expression for the excess entropy:

$$\mathbf{E} = \lim_{L \to \infty} \left[H(L) - h_\mu L \right] \ . \tag{4.9}$$

Hence, $\mathbf{E}$ is the y-intercept of the straight line to which $H(L)$ asymptotes, as indicated in Fig. (4.1). For stationary sources the excess entropy is nonnegative, $\mathbf{E} \geq 0$.

Looking at Eq. (4.8), we see that, informally, $\mathbf{E}$ is the amount in bits, above and beyond h_μ, of *apparent* randomness that is eventually "explained" by considering increasingly longer spin blocks. Conversely, to see the actual (asymptotic) randomness at rate h_μ, we must extract $\mathbf{E}$ bits of information from observations of spin blocks. We expect a large $\mathbf{E}$ to indicate a large amount of structure: $\mathbf{E}$ is large if there are long-range correlations that account for the apparent randomness observed in distributions over small spin blocks.

This interpretation is strengthened by noting that $\mathbf{E}$ may be expressed as the mutual information I, Eq. (3.24), between the two semi-infinite halves of a configuration;

$$\mathbf{E} = I[\overleftarrow{S}; \overrightarrow{S}] \ . \tag{4.10}$$

Note that this form makes it clear that $\mathbf{E}$ is spatially symmetric. Recalling that the mutual information can also be written as the difference between a joint and a conditional entropy:

$$I[\overleftarrow{S};\overrightarrow{S}] = H[\overleftarrow{S}] - H[\overleftarrow{S} \mid \overrightarrow{S}] \ , \tag{4.11}$$

we see that $\mathbf{E}$ measures the average reduction in uncertainty $H[\overleftarrow{S}]$ of the left-half configuration $\overleftarrow{S}$, given knowledge of $\overrightarrow{S}$. One must interpret Eqs. (4.10) and (4.11) with care since they contain entropy contributions, like $H[\overleftrightarrow{S}]$, that individually may be infinite—even for a fair coin process.

Eqs. (4.10) and (4.11) allow us to interpret $\mathbf{E}$ as a measure of how much information one half of the spin chain carries about the other. In this restricted sense $\mathbf{E}$ measures the spin system's apparent spatial memory. If the configurations are perfectly random or periodic with period 1, then $\mathbf{E}$ vanishes. Excess entropy is positive between the two extremes of ideal randomness and trivial predictability. This property ultimately derives from its expression as a mutual information, since the mutual information between two variables vanishes either (i) when the variables are statistically independent or (ii) when they have no entropy or information to share. These extremes correspond to $\mathbf{E}$ vanishing in the cases of ideal randomness and trivial predictability, respectively. Finally, $\mathbf{E}$ measures the average degree of statistical independence of the two halves of a spin chain—how "indecomposable" the chain is.

Note that all three expressions for the excess entropy, Eqs. (4.8), (4.9), and (4.10), indicate that $\mathbf{E}$ carries units of *bits*. This is clear in Eq. (4.10), since the mutual information has units of bits. The entropy density, h_μ has units of bits per site, and L, the length of a spin block, has units of lattice sites. Hence, both terms on the right hand side of Eq. (4.9) have units of bits, so it follows that the left hand side, $\mathbf{E}$, must have units of bits as well. Lastly, Eq. (4.8) tells us that $\mathbf{E}$ is the shaded area in Fig. (4.2). The y-axis of Fig. (4.2) has units of bits per site while the x-axis has units of lattice sites. Since $\mathbf{E}$ is an area on Fig. (4.2), it has units of bits.

It follows immediately that any periodic sequence of period $\mathcal{P}$ has $\mathbf{E} = \log_2 \mathcal{P}$. Returning to the example at the beginning of this section, then, we see that a period-1969 sequence has an excess entropy $\log_2 1969 \approx 10.94$ bits, while the period-4 sequence has an excess entropy of $\log_2 4 = 2$ bits. Thus, as anticipated, the period-1969 sequence does indeed possess more memory than the period-4 sequence.

4.7 Correlation Information

In the previous section we interpreted the excess entropy as the total amount of information that must be extracted from measuring L-blocks in order to recover the asymptotic entropy density—that is, to see just how random each spin is. The entropy convergence plot, Fig. (4.2), is the discrete derivative, with respect to block length L, of the entropy growth curve $H(L)$ of Fig. (4.1). What if we take more discrete derivatives of $H(L)$? It turns out that the second derivative of $H(L)$ recovers the correlation informations $k(L)$ of Refs. [100, 104] and allows for another interpretation of excess entropy.

We follow Refs. [100] and [104] and define the correlation information $k(L)$ of order L as

$$k(L) \equiv \sum_{\{s^L\}} \Pr(s^L) \log_2 \frac{\Pr(s_{L-1}|s^{L-2})}{\Pr(s_{L-2}|s^{L-3})} \ . \tag{4.12}$$

It is not hard to see that $k(L) = h_\mu(L) - h_\mu(L-1)$. Thus, the correlation informations are indeed a discrete derivative of the entropy convergence function $h_\mu(L)$. These quantities have a useful interpretation as the information gain between distant spins [100] and so are somewhat similar to the two-point mutual information just introduced.

From the boundary conditions on $H(L)$, we see that $k(1) = h_\mu(1) = H(1)$. Note that $\lim_{L\to\infty} k(L) = 0$. Under suitable assumptions about the source's structure, it follows from the definition that the excess entropy is directly related to the correlation informations according to

$$\mathbf{E} = \sum_{L=1}^{\infty} L k(L) \ . \tag{4.13}$$

In this form, $\mathbf{E}$ appears as type of correlation length: $\mathbf{E}$ is an average length in which the average is weighted by the correlation informations [100, 104].

4.8 Two-Spin Mutual Information

Finally, we mention that we can use the mutual information to define an information theoretic analogue of the two-spin correlation functions discussed above. The two-spin mutual information is defined by

$$I(r) \equiv I[S_0; S_r] \ , \tag{4.14}$$

and measures the information shared between two spins separated by r sites. Using the translation invariance of spin configurations it follows that:

$$I(r) = 2H[S_0] - H[S_0, S_r] \ . \tag{4.15}$$

Note that $I(0) = H(1)$ and that for a typical source $I(r)$ is monotone decreasing, $I(r) \geq I(r+1)$. For the special case of binary sequences in which $\Gamma(r)$ vanishes as $r \to \infty$, $I(r) \sim \Gamma^2(r),\ r \gg 1$ [95].

4.9 Information Theoretic Approaches to Structure in Dynamics, Statistical Physics, and Elsewhere

The total excess entropy was used by Crutchfield and Packard in Refs. [43, 42, 41, 122] to examine the entropy convergence for noisy discrete-time nonlinear mappings. They developed a scaling theory for the entropy rate convergence: $h_\mu(L) - h_\mu \propto 2^{-\gamma L}$, where, for Markovian finite-memory chains, the excess entropy and entropy convergence exponent γ are simply related: $\mathbf{E} = (H(1) - h_\mu)/(1 - 2^{-\gamma})$. Analytical calculations of entropy convergence for some simple discrete-time nonlinear maps were carried out by Szépfalusy and Györgyi [152]. Subsequently, Csordás and Szépfalusy [46] explored a generalized version of the excess entropy, which they call the “reduced Rényi information”, also in the context of discrete-time nonlinear mappings. The reduced Rényi entropy is also discussed by Z. Kaufmann [80]. Excess entropy was recoined “stored information” by Shaw [144] and subsequently “effective measure complexity” by Grassberger [64]. These two authors emphasize the entropy growth view shown in Fig. (4.1). Entropy growth has also been considered by Eriksson and Lindgren in [52]. $\mathbf{E}$ has been discussed in the context of cellular automata by Grassberger [64] and by Lindgren and Nordahl [104]. Excess entropy was also mentioned briefly by Lindgren in Ref. [99]. The quantity is simply called “complexity” as applied to several stochastic automata by Li [96].

There have been other prior discussions of using information theory to measure a source's structure. For example, mutual information has been proposed as a measure of self-organization [26, 137]. Watanabe [161] and Kolmogorov [87] take approaches that are different, yet again. The latter is particularly notable, though brief, for how its discussion of source structure parallels the philosophies of model inference by minimum message length [160] and minimum description length [133] found in the theories of model order estimation and universal source coding. Both of these approaches address the discovery of source structure, though not as directly as concerns us here.

A number of the above notions have also recurred in more recent discussions of modeling information sources [16, 61]. See also the references in [55] and the critical evaluation there of information theoretic notions of complexity and structure.

It should be emphasized that there are subtle but significant differences in these works' notions of effective complexity, memory, and information. For example, except for Refs. [160] and [133] which are concerned with inductive inference, almost none of the above references pay attention to minimal representations. Minimality is crucial for being able to conclude that a given quantity estimated from a model actually describes an intrinsic structural property of a process and is not an artifact of some unarticulated representational choice—a key issue to which we shall return repeatedly in the following.

The preceding observations on the nature of entropy convergence stay within the framework of information theory—a largely statistical view of “structure”. These quantities and a number of the preceding observations have been known for at least a decade and a half, if not longer.

Chapter 5

Computational Mechanics

In the previous chapter we saw that the excess entropy **E** provides a measure of the apparent spatial memory stored in configurations. However, excess entropy and the apparatus of information theory tell us nothing about *how* the system's memory is organized and utilized. Computational mechanics [33] addresses this issue by paralleling and extending the architectural analyses found in discrete computation theory. For an introduction to discrete computation theory, see, e.g., Refs. [21] or [73]. In explicitly considering how the system produces the apparent spatial memory **E**, we shall be led to put forth another measure of intrinsic memory, known as the statistical complexity, defined as the minimum amount of memory needed to statistically reproduce the original configuration (and the ensemble from which it comes). This is a different interpretation of memory than given to the excess entropy. Chapters 5.6.6 and 7.1.6, however, will show that these two notions of memory are related. This additional set of theoretical tools allows us to describe structure and information processing at a more detailed and complete level than possible via information theory alone.

5.1 Intrinsic Computation

Like statistical mechanics, computational mechanics is concerned with a large system consisting of many individual components. However, computational mechanics addresses very different issues. The motivating questions of computational mechanics center around *how* a system processes information: How is information stored, transmitted, and transformed? How much memory is needed to statistically reproduce an ensemble of configurations and how is this memory organized? In general, we are interested in inferring the intrinsic computation being performed by the system itself.

By *intrinsic* computation [32] we mean something very different than "computation" as the word is typically applied either in reference to the use of modern digital computers as tools for simulation or for symbolic manipulation (e.g., as found in the Journal of Computational Physics) or in reference to the use of a device to perform useful information processing for some person or machine (as in updating a spreadsheet or determining the five billionth digit of π). Useful computation usually entails fixing the initial conditions and control parameters of a dynamical system so that the outcome contains some information of interest to us, as outside interpreters of the result [34]. For example, we might employ the mapping

$$x_{n+1} = \frac{1}{2}(x_n + \frac{a}{x_n}) \ , \ x_0 = 1 \ , \tag{5.1}$$

which has the useful property that $\lim_{n\to\infty} x_n = \sqrt{a}$ [83]. This iterative procedure for increasingly accurate estimates of roots was reported by Hero of Alexandria [118] in the first century B.C.

In contrast, when we ask about intrinsic computation, we are interested not in manipulating a system to produce an output that is useful to us—which is akin to an engineering stance towards nature. Instead, we are interested in examining the information processing that the system itself performs and the underlying mechanisms that support it—which is more of a scientific stance: exploring how nature works on its own terms.

As a concrete example, consider the two-dimensional nearest-neighbor Ising model at the critical temperature. Here the correlations between spins decay with a power law as a function of distance, yet the total magnetization of the system remains zero. Computational mechanics is concerned with what sorts of effective information processing the system must perform to reach and maintain the critical state. How much historical and spatial memory is required? How is the memory organized internally? What spatial patterns result? Are the critical configurations in any way "harder" to reach than those found at low or high temperatures? More informally, how does the system balance up and down spins so that the correlations decay as a power law, while keeping zero magnetization?

Whereas statistical mechanics starts with a system's Hamiltonian or a description of its constituents' local space-time behavior and interactions, computational mechanics begins with the joint probability distribution over the state space trajectories. With knowledge of this joint distribution, the intrinsic computation being performed by the system can be determined. By not requiring a Hamiltonian, computational mechanics can be applied in a wide range of contexts, including those where an energy function for the system may not be manifest.

In any case, as noted above, the two microscopic starting points in the many-body setting—a Hamiltonian or the joint probabilities of configurations over time—are related (at equilibrium) to each other by the usual canonical ensemble,

$$\Pr(\mathcal{C}) \propto e^{-\beta\mathcal{H}(\mathcal{C})} , \tag{5.2}$$

where $\mathcal{C}$ is a configuration, β the inverse temperature, and $\mathcal{H}$ the system's Hamiltonian.

5.2 Effective States: Preliminary Examples

Rather than launching into the mathematical development, we begin our review of computational mechanics with several very simple examples. These will lead quite naturally to the definitions put forth in the subsequent section.

The questions we shall be addressing for each example are: How can one statistically reproduce a given infinite configuration using the minimal amount of memory? In particular, how much information about the left half must be remembered to produce the right half? Here *statistically reproduce* refers to the ability to generate infinite configurations whose finite-length spin blocks occur with the same probabilities as those in the original, infinite configuration.

Another, equivalent way of stating these questions is: How much memory is needed to optimally predict configurations? And, how is this memory organized? We define an *optimally predictive* model in the following way. Suppose we have a model that uses a given amount of historical information to make its predictions. Then we observe L spins s^L in a configuration described by $\Pr(\overleftrightarrow{s})$. Using this information the model produces an estimate $\widehat{\Pr}(s|s^L)$ of the probability of the next spin s. We say that the model optimally predicts the configuration if and only if $\widehat{\Pr}(s|s^L) = \Pr(s|s^L)$ for all s^L and all s, where $\Pr(s|s^L)$ is obtained directly from $\Pr(\overleftrightarrow{s})$.

5.2.1 Fair coin configuration

Consider a string of heads (H) or tails (T) generated by a fair coin toss:

$$\overset{\leftrightarrow}{s}^{\alpha} \equiv \cdots \mathrm{THTTTHHHHHTHTHTTHHT} \cdots \ . \tag{5.3}$$

By definition all tosses are independently distributed; the probability that any particular toss is a heads is 1/2 and any particular length-L block has probability 2^{-L}. We begin by asking: How much of the left half is needed to predict the values in the right half? Restated, imagine walking down the string from left to right, noting the state of the variables one observes. After a very long time—long enough for one to have observed as many tosses as desired—how many of the preceding variables must one keep track of in order to optimally predict those encountered later?

A moment's reflection reveals that one does not need to keep track of any variables. Since the coin tosses are independent, knowledge of previous tosses does not reduce the uncertainty about the next toss. As a result, for this particularly simple example no memory is required to optimally predict subsequent variables. Here, the predictions are as good as they can be (i.e., optimal), which admittedly is not good at all. The uncertainty about the next coin toss is complete. The result could be either heads or tails with equal probability, as reflected by an entropy rate h_μ of 1 bit per toss for the fair coin.

What must one do in order to perform this optimal prediction? Equivalently,

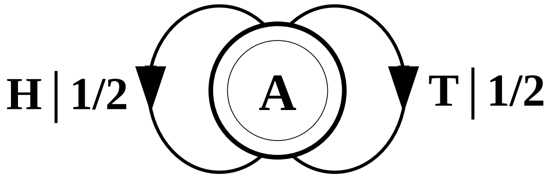


Figure 5.1: The probabilistic finite-state machine for fair coin tosses. This machine is a model of the original configuration in the sense that a random walk through the machine—making state-to-state transitions following the edges, denoted $s|p$ according the their labeled probability p—produces a sequential configuration of symbols $s_i \in \mathcal{A}$ with the same statistical properties as the original $\overset{\leftrightarrow}{s}^{\alpha}$. For more discussion, see text.

how can one statistically reproduce the configuration? The answer to these questions is illustrated in the probabilistic finite-state machine of Fig. 5.1, which compactly tells us how to reproduce strings with the same statistics as the original configuration.

The machine operates as follows. Start in state **A**. With probability 1/2 generate an H and return to state **A**. And with probability 1/2 generate a T and also return to state **A**. A random walk through the machine following these rules results in a string of H's and T's that is statistically identical to $\overset{\leftrightarrow}{s}^{\alpha}$. In this sense we say that the machine constitutes a model of the original fair coin process.

It is important to emphasize that no larger machine—i.e., with more states or edges—is required to reproduce all strings in the class of which $\overset{\leftrightarrow}{s}^{\alpha}$ is one realization. Nor is any smaller machine capable of doing so.

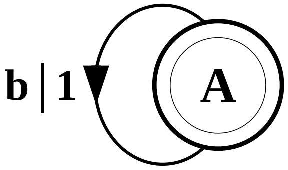


Figure 5.2: The finite-state machine for a string consisting of all b's.

5.2.2 Period-1 configuration

Consider a string consisting of a sequence of all b's:

$$\overset{\leftrightarrow}{s}^{\beta} \equiv \cdots \mathrm{bbbbbbbbbbbbbbbbbbb} \cdots \ . \tag{5.4}$$

As with the fair coin, it is clear that one doesn't need to remember any of the previous symbols to perform optimal prediction. The value of the next variable will be a b regardless of the values of the previous variables.

The finite-state machine for $\overset{\leftrightarrow}{s}^{\beta}$ is shown in Fig. 5.2. From state **A**, the machine always outputs the symbol b and returns to state **A**. In this way the machine statistically reproduces $\overset{\leftrightarrow}{s}^{\beta}$. For this example the prediction is error free, as reflected in the fact that $h_\mu = 0$.

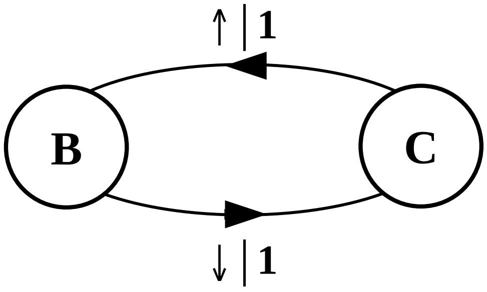


Figure 5.3: The recurrent portion of the finite-state machine for the period-2 configuration $\overleftrightarrow{s}^{\gamma}$ of Eq. (5.5). Note that this machine has two states while the machines of Figs. 5.1 and 5.2 have only one state. This is an indication that the $\cdots\uparrow\downarrow\uparrow\downarrow\cdots$ configuration requires more memory to reproduce.

5.2.3 Period-2 configuration

Now consider an infinite, alternating spin configuration:

$$\overleftrightarrow{s}^{\gamma} \equiv \cdots\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow\cdots\ . \tag{5.5}$$

Again, we begin by asking: How much of the left half is needed to predict spins in the right half? Here, some memory is needed to keep track of the phase of the alternating spin pattern. As long as this phase is remembered, one can optimally and exactly predict all the subsequent spins. As with the period-1 configuration, $\overleftrightarrow{s}^{\gamma}$ can be predicted with certainty since its entropy density is also $h_\mu = 0$. But to achieve this certainty, one must distinguish between the pattern's two different phases. As a result, the state machine for $\overleftrightarrow{s}^{\gamma}$ has (at least) two states, as indicated in Fig. 5.3.

How can we use the machine of Fig. 5.3 to reproduce $\overleftrightarrow{s}^{\gamma}$? Unlike the previous examples, it is not clear where to begin: state **B** or state **C**? One response—consonant with assumptions implicit in equilibrium statistical mechanics—is that it doesn't matter. If we run the system for infinitely long we will statistically reproduce the original configuration. The starting state is just a "boundary condition" whose effects are negligible in the thermodynamic limit.

However, in another sense, the state in which we start most definitely *does* matter. Suppose we choose to start always in state **B**. We then examine all the length-3 spin blocks generated by this choice. We see that the string ↓↑↓ is generated with probability 1. Yet in the original configuration $\overleftrightarrow{s}^{\gamma}$ we observe $\Pr(\uparrow\downarrow\uparrow) = 1/2$ and $\Pr(\downarrow\uparrow\downarrow) = 1/2$. The machine of Fig. 5.3 doesn't correctly predict the statistics of the configuration.

There is an easy remedy for this situation: start in state **B** half the time and state **C** half the time. We can achieve this by adding a *start state* to the model. This is shown in Fig. 5.4. We now always begin operating our model in the unique start state **A**. In Fig. 5.4 and all subsequent figures the start state is indicated by a double circle. The new, improved model generates spin blocks that exactly reproduce the distribution of finite-length spin blocks observed in the original configuration.

In this example, the start state is a transient state. It is never revisited after the machine outputs the first spin value and moves to state **B** or state **C**. The states **B** and **C** in Fig. 5.4 are recurrent, being visited infinitely often as the machine is operated. When examining machines obtained from one-dimensional spin-1/2 Ising

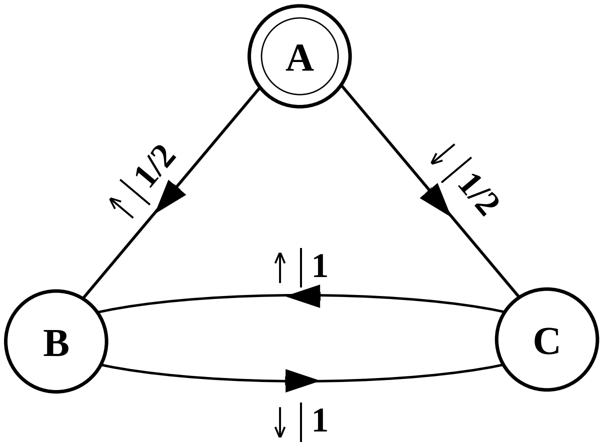


Figure 5.4: The full probabilistic finite-state machine for the period-2 configuration $\overset{\leftrightarrow}{s}^{\gamma}$. The start state **A** is indicated by the double circle. **A** is a transient state; it is never visited again after the machine outputs the first spin. States **B** and **C** are recurrent; they are visited infinitely often as the machine outputs an infinite spin configuration.

systems, we shall encounter examples where the start state is a recurrent state; as was done, but not mentioned, in Figs. 5.1 and 5.2. We shall also see machines that have more than one transient state and that have more complex transient transition structures.

5.2.4 Noisy period-2 configuration

Finally, consider an infinite binary string in which every other symbol sampled from the alphabet $\mathcal{A} = \{0, 1\}$ is a 0, but otherwise the symbols are unconstrained:

$$\overset{\leftrightarrow}{s}^{\delta} \equiv \cdots 01010001010100010000000100010 \cdots \; . \qquad (5.6)$$

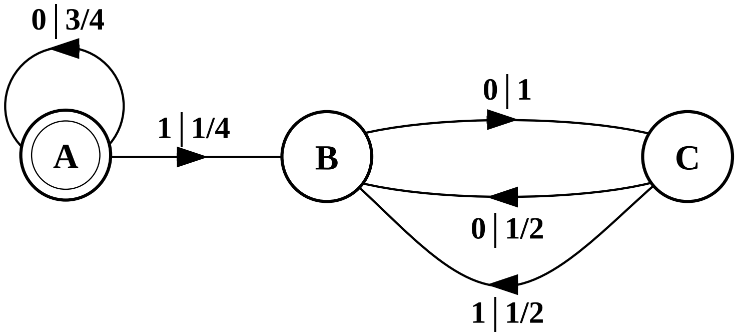


Figure 5.5: The probabilistic finite-state machine for the noisy period-2 configuration $\overset{\leftrightarrow}{s}^{\delta}$. Again, the start state **A** is a transient state and states **B** and **C** are recurrent.

Figuring out how to build a model capable of reproducing this configuration is perhaps not as straightforward as in the previous examples. The key realization is that once we observe a single 1 we are "synchronized" to the pattern. That is, after seeing a 1, a 0 *must* follow, since the configuration never exhibits two adjacent 1's. After seeing the 0 that follows the first 1, a 0 or a 1 can occur with equal probability—the only rule is that every other symbol is 0. The probabilistic finite-state machine for this configuration is shown in Fig. 5.5.

Note that the uncertainty associated with predicting the next symbol changes as one moves back and forth between state **B** and state **C**. From state **B** there is no uncertainty—a 0 is always the next symbol. From **C** there is an associated uncertainty of 1 bit, since the next symbol is equally likely to be a 0 or a 1. Thus, the entropy density is $h_\mu = 1/2$ bit per symbol. It should be not be immediately obvious how we determined the probabilities for transitions leaving state **A**. For this, we will need to review the general procedure for building such machines, as will be done below.

5.2.5 Summary of examples

Despite the examples' simplicity, a few summarizing remarks are in order before moving on to formalize the notion of "effective" state that we've just used implicitly.

First, note that the coin-toss configuration $\overset{\leftrightarrow}{s}^{\alpha}$ and the period-1 configuration $\overset{\leftrightarrow}{s}^{\beta}$ both result in a machine with only one state, an indication that we don't need to remember any information about the previous symbols to predict the values of the next. Thus, predicting a perfectly random process and a process with a very simple configuration are both "easy" tasks in the sense that they require small machines.

Second, note that entropy rate h_μ manifests itself (roughly) as the degree of branching in the machines, measured as the logarithm of the ratio of the number of edges to the number of states, in the recurrent portion of the machines. For example, in Fig. 5.1 there are two edges leaving one state. The entropy rate is 1 bit per symbol.

Third, note that the structure of the machines does not depend on the variables' values—all that matters are the probabilities over configurations. For example, if the symbols H and T are changed to $\uparrow$ and $\downarrow$, the machine of Fig. 5.1 will output different symbols, but its overall structure remains unchanged.

Fourth, transient states tell one how the machines synchronize. For the period-2 example, we argued that the transient state **A** of the machine in Fig. 5.4 was necessary so that the machine would faithfully reproduce the distribution over finite-size blocks. Equivalently, the transient states are necessary for synchronizing the machine if one is reading in data from the configuration. Before any symbols are parsed, one does not know in which internal state the process was as it produced symbols in the configuration. This state of ignorance corresponds to the start state. Transitions are then taken from the start state corresponding to the symbols observed as the configuration is parsed. The number and structure of the transient states determine how difficult it is to synchronize—i. e., to determine in which recurrent state the system is as it produces each symbol.

Lastly, note that we have taken care to construct minimal machines. That is, the machines we've put forth are such that if one removes any state or transition then one can no longer exactly statistically reproduce the configuration. This notion of minimality will be made more precise below. In complementary fashion, in each example one gains no predictability by elaborating any of the machines by adding states or transitions.

5.3 Causal States and ϵ-Machines

The preceding section considered models capable of reproducing the distribution of all finite length blocks observed in several translationally invariant configurations. This section presents a general procedure for constructing such models—minimal, optimally predictive, probabilistic state machines.

First, we need to formalize the intuitive arguments through which the "effective" states of the four example systems were discovered. The key step is identifying the notion of effective state with the conditional probability distribution over right-half configurations. And a central criterion is that the resulting model be minimal. When constructing an optimally-predictive, minimal state-machine description, there is no need to distinguish between different left-half configurations that give rise to an identical state of knowledge about the right-half configurations that follow it. Maintaining a distinction between two such states adds to the model's size without increasing its predictive ability. Therefore, we will be looking for the smallest set of predictive states.

To make these ideas precise, consider the probability distribution of all possible right halves $\overrightarrow{s}$ conditioned on a particular left half $\overleftarrow{s}_i^L$ of length L at site i: $\Pr(\overrightarrow{s}\,|\overleftarrow{s}_i^L)$, $0 \leq L \leq \infty$. For $L = 0$, $\overleftarrow{s}_i^L$ is the empty string, denoted by λ. That is, $\Pr(\overrightarrow{s}\,|\overleftarrow{s}_i^0) \equiv \Pr(\overrightarrow{s}\,|\lambda) = \Pr(\overrightarrow{s})$ denotes the probability of observing $\overrightarrow{s}$ unconditioned on any spins in the left half of the configuration.

We now use this form of a conditional probability to define an equivalence relation $\sim$ on the space of all left halves. We say that two left-half configurations at different lattice sites are equivalent (under $\sim$) if and only if they give rise to distributions over right-half configurations, conditioned on those left-halves, that are identical. Formally, we define the relation $\sim$ by:

$$\overleftarrow{s}_i^M \sim \overleftarrow{s}_j^L \text{ iff } \Pr(\overrightarrow{s} | \overleftarrow{s}_i^M) = \Pr(\overrightarrow{s} | \overleftarrow{s}_j^L) \ , \tag{5.7}$$

for all $\overrightarrow{s}$, where $L, M = 0, 1, 2, \ldots$. The induced equivalence classes are subsets of the set of all allowed $\overleftarrow{s}_i^L$. Appendix C reviews various properties of equivalence relations.

In a setting in which the conditional probabilities $\Pr(\overrightarrow{s} | \overleftarrow{s}_i^L)$ aren't known exactly, it becomes necessary as a practical matter to introduce some tolerance into the equivalence relation defined by Eq. (5.7). Implementing this is not a straightforward task, since if one adds a tolerance δ and writes

$$\overleftarrow{s}_i^M \sim \overleftarrow{s}_j^L \text{ iff } \Pr(\overrightarrow{s} | \overleftarrow{s}_i^M) = \Pr(\overrightarrow{s} | \overleftarrow{s}_j^L) + \delta \ , \tag{5.8}$$

the equivalence relation is destroyed because $\sim$ is no longer transitive; see App. C. We will address the issues surrounding the implementation of a tolerance elsewhere. For now, suffice it to say that the basic difficulty this introduces is common to other inference problems that involve statistical clustering and classification [112, 141]. Since we are focusing here on processes for which one can perform the necessary calculations analytically, statistical estimation will not be a concern.

The equivalence classes over

$$\{\overleftarrow{s}_i^L, \ i = \ldots, -2, -1, 0, 1, 2, \ldots \ , L = 0, 1, 2, \ldots\} \tag{5.9}$$

induced by this relation are called *causal states* and denoted $\mathcal{S}_\alpha, \ \alpha = 0, 1, 2, \ldots$. The $\mathcal{S}_\alpha$ are the "effective states" of the previous section. Two $\overleftarrow{s}^L$ belong to same causal state if, as measured by the probability distribution of subsequent spins conditioned on having seen each particular left-half configuration, they give rise to the same degree of certainty about the configurations that follow to the right.

We shall use the convention that causal states $\mathcal{S}_\alpha$ are generically indexed using Greek letters. Spin variables shall continue to be indexed with Roman letters. The equivalence class associated with $\Pr(\overrightarrow{s} \, | \lambda)$ is always the start state, since this distribution corresponds to the knowledge about right-half configurations before any spins are observed. The start state is denoted $\mathcal{S}_0$.

The causal states, as determined by the equivalence classes induced by Eq. (5.7), give transient as well as recurrent states. Although they may be visited an infinite number of times, transient states are those causal states that have vanishing probability in the $L, M \to \infty$ limit. In contrast, recurrent states are those visited infinitely often and have positive probability in the same limit. That is, the recurrent states are those equivalence classes obtained when the $L, M \to \infty$ limit is considered in Eq. (5.7). By considering how the process synchronizes—i.e., how it reaches the recurrent states as successively longer blocks are generated—it is possible to construct the transient states and their transitions from knowledge of the recurrent states alone. The procedure by which this is done is given in App. D.

We denote the set of causal states by $\mathcal{S} = \{\mathcal{S}_\alpha, \alpha = 0, \ldots, k-1\}$. For the processes considered here $\mathcal{S}$ is discrete and $k = |\mathcal{S}|$ is finite—neither of which necessarily holds in a general setting [33, 154]. Let $\mathcal{S}^{(T)}$ denote the set of transient states and $\mathcal{S}^{(R)}$ denote the set of recurrent states. Note that $\mathcal{S} = \mathcal{S}^{(T)} \cup \mathcal{S}^{(R)}$.

There is a mirror image definition of causal states obtained by scanning the lattice in the opposite direction (right to left), which thus uses distributions conditioned on right-half configurations. Since we will study a restricted class of systems that respect this symmetry, the causal states will be the same regardless of the scanning direction. In the general case, in which this reversal symmetry need not hold, it is possible to find different causal states if one scans $\overset{\leftrightarrow}{s}$ in different directions [32].

As we saw above, for the period-2 system there are 3 causal states, denoted in Fig. 5.4 by **A**, **B**, and **C**. These causal states are subsets of the allowed $\overset{\leftarrow}{s}^L$:

$$\mathbf{A} \;=\; \{\lambda\}\ , \tag{5.10}$$

$$\begin{aligned}\mathbf{B} \;&=\; \{\overset{\leftarrow}{s}^L | s_{-1} =\uparrow, s_{-2} =\downarrow, s_i = s_{i+2}, L \geq 1\} \\ &=\; \{\uparrow, \downarrow\uparrow, \uparrow\downarrow\uparrow, \downarrow\uparrow\downarrow\uparrow, \uparrow\downarrow\uparrow\downarrow\uparrow, \ldots\}\ ,\end{aligned} \tag{5.11}$$

and

$$\mathbf{C} \;=\; \{\overset{\leftarrow}{s}^L | s_{-1} =\downarrow, s_{-2} =\uparrow, s_i = s_{i+2}, L \geq 1\}$$

$$= \quad \{\downarrow, \uparrow\downarrow, \downarrow\uparrow\downarrow, \uparrow\downarrow\uparrow\downarrow, \downarrow\uparrow\downarrow\uparrow\downarrow, \ldots\} \ . \tag{5.12}$$

Here $\mathcal{S} = \{\mathbf{A}, \mathbf{B}, \mathbf{C}\}$, $\mathcal{S}^{(T)} = \{\mathbf{A}\}$, and $\mathcal{S}^{(R)} = \{\mathbf{B}, \mathbf{C}\}$.

Once the set $\mathcal{S}$ of causal states has been identified, we determine the transition probabilities $T^{(s)}_{\alpha\beta}$ between states upon seeing symbol $s \in \mathcal{A}$. That is, we need to find:

$$T^{(s)}_{\alpha\beta} \equiv \Pr(\mathcal{S}_\beta, s|\mathcal{S}_\alpha) \ . \tag{5.13}$$

To understand the nature of the transition probabilities better, we rewrite Eq. (5.13):

$$\Pr(\mathcal{S}_\beta, s|\mathcal{S}_\alpha) = \Pr(\mathcal{S}_\beta|s, \mathcal{S}_\alpha)\Pr(s|\mathcal{S}_\alpha) \ . \tag{5.14}$$

Knowledge of the next spin's value s uniquely determines the subsequent causal state $\mathcal{S}_\beta$. To see this, note that moving one step to the right corresponds to moving from $\overleftarrow{s}_i$ to $\overleftarrow{s}_{i+1} = \overleftarrow{s}_i s$. Since the causal states partition the set of left-half configurations, the new left-half configuration $\overleftarrow{s}_{i+1}$ is associated with one and only one causal state. Hence, observing the next spin value s determines the next causal state $\mathcal{S}_\beta$, as the chain is parsed from left to right.

This is the sense in which the causal state representation is *deterministic.* A transition from state α to state β while outputting a symbol s is uniquely determined by α and s. That is, $\Pr(\mathcal{S}_\beta|s, \mathcal{S}_\alpha) = 1$, assuming the transition is allowed. To illustrate this, consider the noisy period-2 machine of Fig. 5.5. From state $\mathbf{B}$, outputting a 0 leads one to state **C**. And from state **C**, seeing a 1 determines that the next state will be **B**. Note, however, that knowledge of the initial and final causal states does *not* determine what symbol was produced as the transition was made. For example, either a 0 or a 1 can be produced upon a transition from state **C** to **B**.

Eq. (5.14) indicates how to obtain the transition probabilities $T^{(s)}_{\alpha\beta} = \Pr(s|\mathcal{S}_\alpha)$ from the joint probabilities over configurations. Let $\overrightarrow{s}^L = s_0 s_1 \cdots s_{L-1}$ be a spin block that leads to, and belongs to, the causal state $\mathcal{S}_\alpha$. Then:

$$T^{(s)}_{\alpha\beta} = \Pr(s|\mathcal{S}_\alpha) = \frac{\Pr(s_0 s_1 \ldots s_{L-1} s)}{\Pr(s_0 s_1 \cdots s_{L-1})} \ . \tag{5.15}$$

where β indexes the causal state $\mathcal{S}_\beta$ to which one is taken on s. In other words, $s_0 s_1 \ldots s_{L-1} s \in \mathcal{S}_\beta$.

Summing over the spin values s, we obtain the stochastic connection matrix $\mathrm{T} = \sum_{s \in \mathcal{A}} T^{(s)}$, a matrix whose components $\mathrm{T}_{\alpha\beta}$ give the probability of a transition from the α^{th} to the β^{th} causal state;

$$\mathrm{T}_{\alpha\beta} \equiv \Pr(\mathcal{S}_\beta|\mathcal{S}_\alpha) \ . \tag{5.16}$$

Since the probabilities are normalized, $\sum_\beta \mathrm{T}_{\alpha\beta} = 1$, and so T is a stochastic matrix. That is, the probability of leaving a state is unity. The probability $\Pr(\mathcal{S}_\alpha)$ of finding this "internal" Markov chain in the α^{th} causal state after the machine has been scanning infinitely long is the left eigenvector of T associated with eigenvalue 1, normalized in probability. That is, $\Pr(\mathcal{S}_\alpha)$ is given by:

$$\sum_{\alpha=0}^{k-1} \Pr(\mathcal{S}_\alpha)\mathrm{T}_{\alpha\beta} = \Pr(\mathcal{S}_\beta) \ . \tag{5.17}$$

Again, the asymptotic probability of all transient states is zero;

$$\Pr(\mathcal{S}_\alpha) = 0, \ \mathcal{S}_\alpha \in \mathcal{S}^{(T)} \ . \tag{5.18}$$

For the period-2 machine of Fig. 5.4 we have:

$$T^{(s=\uparrow)} = \begin{pmatrix} 0 & 1/2 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \tag{5.19}$$

and

$$T^{(s=\downarrow)} = \begin{pmatrix} 0 & 0 & 1/2 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} , \tag{5.20}$$

where we take column and row labels to correspond to causal states in the natural way: $1 \mapsto \mathbf{A}, 2 \mapsto \mathbf{B}$, and so on. We add Eqs. (5.19) and (5.20) to obtain the machine's stochastic connection matrix:

$$\mathrm{T} = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} . \tag{5.21}$$

Note that T is stochastic and its dominant left eigenvector, normalized in probability, is $(0, 1/2, 1/2)$. Hence, $\Pr(\mathcal{S} = \mathbf{A}) = 0$ and $\Pr(\mathcal{S} = \mathbf{B}) = \Pr(\mathcal{S} = \mathbf{C}) = 1/2$. The asymptotic probability of the transient state **A** is zero.

The set $\mathcal{S}$ together with the dynamic $\{T^{(s)}, s \in \mathcal{A}\}$ constitute a model—referred to as an *ϵ-machine* [44]—of the original process. The four example machines of the previous section, Figs. 5.1, 5.2, 5.3, and (5.5, are all ϵ-machines. An ϵ-machine is the minimal representation that captures the intrinsic computation being performed by the system under study in the sense that it explicitly lays out how information in the left-half configuration is stored in the causal states and determines the range of right-halves that can be seen. In other words, an ϵ-machine shows how much memory a process has, how it is organized, and how it is used to generate the pattern exhibited by the process.

The minimality of the ϵ-machine follows immediately from the definition of the equivalence relation, Eq. (5.7). The equivalence classes induced by the relation are associated with the causal states. The procedure of forming equivalence classes ensures that we distinguish between only those states that give rise to different predictive information. As a result,

$$\Pr(s|\mathcal{S}_\alpha) \neq \Pr(s|\mathcal{S}_\beta) \ , \tag{5.22}$$

for $\alpha \neq \beta$ and for at least one value of s. Recall that we demand that the ϵ-machine be capable of statistically reproducing the original configuration. If we make our machine smaller by merging two states, say α and β, then it follows immediately from Eq. (5.22) that the machine will no longer be able to exactly statistically reproduce the original configuration since it fails to distinguish between the different conditional probabilities of Eq. (5.22). Thus, we conclude that an ϵ-machine is minimal.

The "ϵ" in the ϵ-machine signifies that, in general, the measurement values $s \in \mathcal{A}$ are not direct indicators of the observed process's internal states [32]. For example, the symbols may be discretizations of variables that are continuous in state, space, or time. For spin systems these concerns are not at issue, since we know by definition the full set of elementary measurement values, i. e., the range of spin values at each site.

In chapter 7, we determine ϵ-machines beginning with the Hamiltonian assumed for our model spin systems. However, as mentioned above, a Hamiltonian is not necessary. The determination of an ϵ-machine does not depend on knowledge of the dynamics or rule through which the configurations were generated. Moreover, the causal states and their transition probabilities may be calculated within two different paradigms; one mathematical, the other empirical. In the first, one begins with the joint distribution over all the system variables. In the second, one is given configurations from which the joint and various conditional distributions are estimated. The overall procedure in the second setting is referred to as ϵ-machine *reconstruction*. In either case, the goal is to factor the joint distribution over spin configurations into the causal state conditional distributions. The result is an ϵ-machine that consists of the components $\{\mathcal{S}, \{T^{(s)}\}, \mathcal{A}, \mathcal{S}_0\}$, where $\mathcal{S}_0 \in \mathcal{S}$ is the ϵ-machine's unique start state.

5.4 Related Computational and Statistical Model Classes

Restricting attention to ϵ-machines for 1D finite-range spin systems, if we strip off their transition probabilities, leaving just the allowed transitions, we change the ϵ-machine representation into a special class of deterministic discrete-state automata [21, 73]. Unlike the general class of automata, these nonprobabilistic ϵ-machines have the following properties: (i) a unique start state, (ii) all states are accepting, (iii) all recurrent states form a single strongly connected component in the machine's state transition graph.

A further restriction on these nonprobabilistic ϵ-machines is that there is a specific relationship between the structure of the transient states and the recurrent states. (This relationship does not hold in general for discrete-state automata.) That is, the nonprobabilistic ϵ-machine's transient states can be constructed from knowledge of the recurrent causal states alone. Appendix D gives a procedure that determines this relationship for the unrestricted, probabilistic case.

Unlike discrete-state automata, however, ϵ-machine transitions are labeled with conditional probabilities $T^{(s)}_{\alpha\beta}$. Said differently, an ϵ-machine represents a configuration distribution, not just a set of allowed configurations, as the automata do. Therefore, in important ways ϵ-machines are a richer class of representations.

For the spin systems considered in subsequent chapters, ϵ-machines can also be viewed as a type of Markov chain. First, the stochastic connection matrix T, which describes only the state-to-state transitions unconditioned by spin values, is a Markov chain over the causal states. Second, and more directly, the full ϵ-machine, including spin labelings, is a subset of models called variously *functions of Markov chains* [18] or *stochastic deterministic finite automata* [33, 154], since the output (spin) alphabet $\mathcal{A}$ differs from the internal (causal) state alphabet $\mathcal{S}$. To be more specific, an ϵ-machine is a function of a Markov chain that has a unique start state and one recurrent component. These, in turn, are a subclass of *hidden Markov models* [51].

It is important to emphasize that only in some cases do ϵ-machines reduce to functions of a Markov chain or, similarly, to probabilistic analogs [125] of discrete-state automata. ϵ-machines are best considered on their own terms—a different model class that captures different types of structure.

5.5 What Do ϵ-Machines Represent?

Given that ϵ-machines can be related to this range of statistical and computational model classes, it is important to note that an essential distinguishing feature of computational mechanics is its hierarchical inductive framework. It begins by trying to model the original process using the *least* powerful model class; probabilistic finite-memory machines are employed first. However, using a finite-memory representation may not yield a finite-size model: the number of causal states could turn out to be countably infinite, as noted above, or to lie in a fractal set or in a continuum [33, 154]. If this is the case, a more powerful model than a finite-state machine must be used. One proceeds by trying to use the next most powerful model classes in a hierarchy of machines known as the causal hierarchy [33]. The latter is an analog of the Chomsky discrete-computation hierarchy of formal language theory [21, 73].

It was suggested in Sec. 2.4 that, in a statistical mechanics context, using the most compact mathematical entity that provides a complete description of a system is an important way to distinguish between systems that are structured in different ways. The determination of an ϵ-machine may be thought of as a formalization of this process of detection and classification of structure. An ϵ-machine, the set of causal states and their transitions, provides a direct description of the structure present in the joint probabilities over the system's internal degrees of freedom. In particular, the ϵ-machine's organization shows how this joint distribution factors into conditionally independent components. Thus, determining the class of ϵ-machine that provides a finite description of the original configuration allows one to distinguish between systems that are organized in fundamentally different ways.

Furthermore, an ϵ-machine gives a minimal description of the pattern or regularities in a system in the sense that the pattern *is* the algebraic structure determined by the causal states and their transitions. If, for example, the ϵ-machine has an algebraic structure that is a group, then it captures a symmetry: for example, translation or spin-flip. That is, it captures the "pattern" exhibited in the system's configurations. Generally, though, the algebraic structure is a semigroup—and a stochastic one at that—and so not obviously interpreted in terms of symmetries. The appropriate mathematical descriptions are given in terms of measure semi-groups [82]. Despite a lack of familiar interpretations, the algebraic structure still captures the intrinsic "pattern" [171]. Examples will be given in Chapter 9 as concrete illustrations of this algebraic view of pattern.

In summary, an ϵ-machine is a model of a system's allowed-configuration ensemble. From this model, we can proceed to define and calculate macroscopic, global properties that reflect the characteristic average information processing capabilities of the system. We now turn to discuss just what features are calculable from an ϵ-machine.

5.6 Global Spatial Properties from ϵ-Machines

5.6.1 Statistical complexity

An ϵ-machine is a model capable of statistically reproducing a process's configurations. How much memory is needed on average to operate this machine? Similarly, how little internal memory could the generating process itself have used? Motivated by these questions we now define a new quantity.

To predict successive spins as one scans a configuration from left to right, one must track in which causal state the process is, since knowledge of the causal state gives the required conditional distribution for optimal prediction. Thus, the informational size of the distribution $\Pr(\mathcal{S}_\alpha)$ over causal states, as measured by the Shannon entropy, gives the minimum average amount of memory needed to optimally predict the right-half configurations. This quantity is the *statistical complexity* [44]:

$$C_\mu \equiv H[\mathcal{S}] = -\sum_{\alpha=0}^{k-1} \Pr(\mathcal{S}_\alpha) \log_2 \Pr(\mathcal{S}_\alpha) \ , \tag{5.23}$$

where, again, $\Pr(\mathcal{S}_\alpha)$ is given by Eq. (5.17). Like the excess entropy $\mathbf{E}$, the statistical complexity C_μ is a measure of memory and has units of bits. Note, however, that the two measures of memory have different interpretations. The excess entropy measures the *apparent* memory stored in the configurations, since it is determined directly from the spin configuration distribution; that is, from the spin observables. In contrast, C_μ

measures the minimal amount of (hidden) memory needed to statistically reproduce the configuration ensemble. As we shall see below, these two measures of memory, though related, typically are *not* equal.

Another, coarser measure of the ϵ-machine's size is simply the number of recurrent causal states. This motivates the definition of the *topological complexity* C_0 [33]:

$$C_0 = \log_2 |\mathcal{S}^{(R)}| \, . \tag{5.24}$$

The topological complexity gives a simple "counting" upper bound on the statistical complexity: $C_\mu \leq C_0$. This follows from a basic maximization property of Shannon entropy applied to a uniform distribution over the causal states.

5.6.2 Block distributions and entropies

We claimed above that an ϵ-machine is a model of a configuration in the sense that it reproduces the spin-block distributions $\Pr(s^L)$. We now show explicitly how these distributions follow from the recurrent portion of an ϵ-machine. In subsequent sections we will then be able to easily calculate the various information theoretic and statistical mechanical quantities defined above.

First, note that the sequence of causal states is Markovian. The probability of a transition from recurrent state $\mathcal{S}_\alpha$ to recurrent state $\mathcal{S}_\beta$ is given by $\mathrm{T}_{\alpha\beta}$. Hence,

$$\Pr(\mathcal{S}_\alpha, \mathcal{S}_\beta) = \Pr(\mathcal{S}_\alpha)\mathrm{T}_{\alpha\beta} \ . \tag{5.25}$$

The probability that the particular sequence $\mathcal{S}_{\alpha_0}, \cdots, \mathcal{S}_{\alpha_{L-1}}$ occurs is given by:

$$\Pr(\mathcal{S}_{\alpha_0}, \cdots, \mathcal{S}_{\alpha_{L-1}}) = \Pr(\mathcal{S}_{\alpha_0}) \prod_{i=0}^{L-2} \mathrm{T}_{\alpha_i \alpha_{i+1}} \ . \tag{5.26}$$

However, we are interested in the distribution of spin blocks, as well as sequences of causal states. Recall that $T^{(s)}_{\alpha\beta} \equiv \Pr(\mathcal{S}_\beta, s|\mathcal{S}_\alpha)$ is the probability of making a transition from state α to state β while producing the spin s. Each (α, β)-entry in the word matrix

$$T^{(s^L)} \equiv T^{(s_0)} T^{(s_1)} \cdots T^{(s_{L-1})} \tag{5.27}$$

gives the probability of seeing word $s^L = s_0 s_1 \ldots s_{L-1}$ starting in state α and ending in state β. (In Eq. (5.27), the matrices on the right hand side are understood to be multiplied together.) Using this matrix we can easily write down an expression for the probabilities over spin blocks:

$$\Pr(s^L) = \sum_{\alpha,\beta=0}^{k-1} \Pr(\mathcal{S}_\alpha) T^{(s^L)}_{\alpha\beta} \ . \tag{5.28}$$

Here we sum over the probabilities of all sequences of $L+1$ causal states, selecting only those for which the particular spin sequence $s_0, \cdots, s_{L-1}$ occurs.

Given the joint distribution over spins blocks, Eq. (5.28), the block entropies $H(L)$ follow immediately from Eq. (4.1).

5.6.3 Two-spin mutual information and correlation function

Recall that using the translation invariance of the configurations, the two-spin mutual information of Sec. 4.8 was given by:

$$I(r) = 2H[S_0] - H[S_0, S_r] \, . \tag{5.29}$$

The second entropy term on the right-hand side requires calculating the joint distribution $\Pr(s_0, s_r)$ and the first requires $\Pr(s_0)$. In the previous section we derived an expression for $\Pr(s_0)$, Eq. (5.28). Thus, to calculate $I(r)$ we need to develop an expression for $\Pr(s_0, s_r)$.

$\Pr(s_0, s_r)$ is easy to obtain by summing over all intervening spins in Eq. (5.28):

$$\begin{aligned}
\Pr(s_0, s_r) &= \sum_{s_1, \cdots, s_{r-1}} \Pr(s_0, s_1, \cdots, s_{r-1}, s_r) && (5.30)\\
&= \sum_{\alpha,\beta=0}^{k-1} \Pr(\mathcal{S}_\alpha) \sum_{s_1, \cdots, s_{r-1}} T_{\alpha\beta}^{(s_0 \ldots s_r)} && (5.31)\\
&= \sum_{\alpha,\beta,\gamma,\delta=0}^{k-1} \Pr(\mathcal{S}_\alpha) T_{\alpha\beta}^{(s_0)} \mathrm{T}_{\beta\gamma}^{r-1} \mathrm{T}_{\gamma\delta}^{(s_r)} \, , && (5.32)
\end{aligned}$$

since $\mathrm{T}^{r-1}_{\alpha\beta} = \sum_{s_1,\cdots,s_{r-1}} T^{(s_1 \cdots s_{r-1})}_{\alpha\beta}$ and where T^r denotes the r^{th} power of the connection matrix T. The last equality follows since the summation over the α's has the effect of multiplying together the T matrices.

The Shannon entropy of $\Pr(s_0, s_r)$ is $H[S_0, S_r]$ and the entropy of $\Pr(s_0)$ is $H[S_0]$ and so $I(r)$, Eq. (4.15), follows immediately.

Using these same distributions it is now possible to calculate $\Gamma(r)$, the two-spin correlation function of Eq. (2.14), since

$$\Gamma(r) = \sum_{s_0,s_r} s_0 s_r \Pr(s_0, s_r) - \left(\sum_{s_0} s_0 \Pr(s_0) \right)^2 . \tag{5.33}$$

The structure factors and susceptibility follow directly from $\Gamma(r)$. Thus, all of these quantities can be readily calculated once an ϵ-machine is in hand.

5.6.4 ϵ-Machine entropy rate

Recall from Eq. (4.4) that the entropy density h_μ can be expressed as the entropy of one spin conditioned on all those spins preceding it. Using this, it is not hard to show that the entropy density can be expressed as the next-spin uncertainty averaged over the causal states:

$$h_\mu = -\sum_{\alpha=0}^{k-1} \Pr(\mathcal{S}_\alpha) \sum_{s\in\mathcal{A}} \Pr(s|\mathcal{S}_\alpha) \log_2 \Pr(s|\mathcal{S}_\alpha) \ , \tag{5.34}$$

where $\Pr(\mathcal{S}_\alpha)$ is given by Eq. (5.17) and $\Pr(s|\mathcal{S}_\alpha)$ is given by Eq. (5.15).

This result is similar to, but not the same as, that originally given in App. 4 of Ref. [142] for Markov chains. We derive it here in App. B. Our result is not surprising given the definition of causal states, which groups together left-half configurations that lead to the same conditional distribution over possible right-half configurations. As a result, to calculate the entropy density h_μ one only need consider the entropy of a single spin conditioned on the current causal state.

The entropy rate is invariant under a change in the direction in which the configuration is scanned [32]. This fact is quite general and holds for any one-dimensional stationary process, a class of systems much broader than the spin systems considered here.

5.6.5 ϵ-Machine excess entropy

The excess entropy $\mathbf{E}$ can also be calculated from the probabilities of the causal states and their transitions. In the most general setting there is no compact formula for $\mathbf{E}$ in terms of $\Pr(\mathcal{S})$ and $\Pr(s|\mathcal{S})$, as there was for h_μ. However, we shall see in Sec. 7.1.5 that for the special case of finite-range spin systems, it is possible to write down a relatively simple formula for $\mathbf{E}$ in terms of an ϵ-machine.

5.6.6 Relationships between measures of memory

As remarked above, the excess entropy and the statistical complexity are different measures of a system's memory. However, it turns out that the excess entropy sets a lower bound on the statistical complexity:

$$\mathbf{E} \leq C_\mu \ . \tag{5.35}$$

This result holds for any translationally invariant infinite configuration [45]. Thus, the memory needed to perform optimal prediction of the right-half configurations can exceed the mutual information between the left and right halves themselves. This relationship reflects the fact that, in the general setting, a process's internal state sequences are not in one-to-one correspondence with L-block or even ∞-length configurations.

5.6.7 The examples analyzed quantitatively

In Table 5.1 we show the results of calculating (by direct inspection) the entropy density h_μ, the excess entropy $\mathbf{E}$, and the statistical complexity C_μ for the example processes of Sec. 5.2.

Process	h_μ	**E**	C_μ
Fair Coin	1	0	0
Period 1	0	0	0
Period 2	0	1	1
Noisy Period 2	1/2	1	1

Table 5.1: The entropy density h_μ, the excess entropy $\mathbf{E}$, and the statistical complexity C_μ for the four example processes of section 5.2.

5.6.8 Scan-direction invariance

Interestingly, one can show that for some classes of systems (not including the finite-range spin systems here) C_μ and C_0 are not scan-direction invariant [32]. That is, the causal states, and as a result C_μ and C_0, may be different depending the direction in which the configuration is scanned: left to right or right to left. However, the values of the entropy rate h_μ, the excess entropy $\mathbf{E}$, and the two-spin mutual information $I(r)$ *are* independent of the direction in which the configuration is observed. This scan-direction invariance derives from these quantities' definitions and is not a result which is particular to spin systems.

5.6.9 Related, or not, "complexity" measures

As noted above, an ϵ-machine is a model of the original process that uses the least powerful computational class admitting a finite model [33]. In contrast, Kolmogorov-Chaitin (KC) complexity, discussed in Sec. 4.3, characterizes symbol sequences by considering their representation in terms of the most powerful of the discrete computational model classes, the universal Turing machines (UTMs).

Note that $C_\mu > 0$ and $\mathbf{E} > 0$ do not imply that memory resources are expended trying to account for the randomness or thermal fluctuations present in a system. Thus, these measures of structural complexity depart markedly from the KC (deterministic UTM) complexity. As noted above, the ensemble-averaged per-site KC complexity is h_μ [30, 94]. And so, the KC complexity is dominated by random components in a process. It does not strongly reflect the algebraic symmetries or structural properties, unless the entropy rate is zero.

One unfortunate shortcoming of KC complexity, and its UTM-based framework, is that it is in general uncomputable [30, 94]. That is, unlike statistical complexity and excess entropy, there exists no general algorithm for its calculation. It should be noted, however, that in special cases such as finite-state Markov chains [30] or continuous-state dynamical systems with an absolutely continuous invariant measure [22], the *average value of the growth rate* of the Kolmogorov-Chaitin complexity can be calculated and is equal to the Shannon entropy rate h_μ of the process.

A quantity more closely related to statistical complexity and excess entropy is the *logical depth* of Bennett [13]. Whereas the Kolmogorov-Chaitin complexity of a symbol string is defined in terms of deterministic-UTM program length, the logical depth is defined as the time needed for the UTM, running the minimal program, to produce the string. On the one hand, if a configuration, like $\overleftrightarrow{s}^{\alpha}$, is random, the shortest UTM program that reproduces it is the program "Print($\overleftrightarrow{s}^{\alpha}$)". This is a relatively long program but takes very little time to run: a time proportional to the length of $\overleftrightarrow{s}^{\alpha}$. On the other hand, if a configuration has a simple pattern, like $\overleftrightarrow{s}^{\beta}$'s string of all b's, then the program to reproduce it also takes a short time to run: a time proportional to the number of b's to print. The minimal program is also short: all the UTM needs to do is loop over the command "Print b", counting up to the desired string length. But if a spin configuration has a great deal of intricacy—for example, if the spins code for the binary expansion of π—then the minimal program to reproduce it will involve many operations, many more than the number of desired spins.

As a result, like excess entropy and statistical complexity, the logical depth captures a property—being low for both simple and random configurations—that is distinct from randomness and from those properties captured by the entropy rate and Kolmogorov-Chaitin complexity. While there are superficial similarities, however, C_μ and $\mathbf{E}$ are measures of memory while logical depth is a measure of run time. A shortcoming of logical depth shared with KC complexity is that it is in general uncomputable [30, 94]. That is, there exists no general algorithm for its calculation.

For other approaches to statistical complexity and correlational structure see Refs. [5, 55, 61, 74, 90, 105] and citations therein.

5.6.10 ϵ-Machine thermodynamics

As a final note, we mention that ϵ-machines also provide a direct way to calculate the thermodynamic potentials for a process. These are also known as the fluctuation spectrum, the Rényi entropy, the spectrum of singularities, $S(U)$ curves, and $f(\alpha)$ curves [12, 68, 172]. The fluctuation spectrum provides a measure of how likely a system is to deviate from its average behavior and is closely related to more modern methods, as found in the theory of large deviations [24, 119], to describe a process's behavior outside of the range of validity of the law of large numbers.

In Ref. [172] it was shown that calculating the fluctuation spectrum by first determining the ϵ-machine and then proceeding to calculate the spectrum from the machine yields significantly more accurate results than estimating the spectrum directly from configurations by using histograms to estimate spin-block probabilities. Finally, one can analyze the fluctuation spectra of causal state sequences themselves by replacing the Shannon entropy in the definition of statistical complexity with the Rényi entropy.

5.7 Summary and a Look Ahead

In this and the previous two chapters we have reviewed the tools used by statistical mechanics, information theory, and computational mechanics to measure correlation and structure. The main quantities from statistical mechanics, discussed in Chapter

2, are correlation functions $\Gamma(r)$, the correlation length ξ, and the structure factors $S(q)$. In Chapter 4 we saw that information theory provides a measure, h_μ, of the randomness or unpredictability of a system and also provides measures of the apparent spatial memory of a configuration, the excess entropy $\mathbf{E}$ and the coarser two-spin mutual information $I(r)$.

However, information theory tells us little about how a system utilizes its memory nor whether the apparent memory ($\mathbf{E}$) is equal to the minimum amount of memory (C_μ) actually required internally to produce configurations. To help address this concern, computational mechanics was put forth as a way to discover and quantify the intrinsic computational capability of a system. By constructing a model (an ϵ-machine) that statistically reproduces the system's configurations, we obtain an explicit description of the architecture of the minimal information processing apparatus needed to produce the configuration ensemble. One consequence is that the statistical mechanical and information theoretic quantities can be calculated directly.

Let's now return to the theme of this dissertation: discovering structure and quantifying patterns. Do ϵ-machines capture our intuitive notion of pattern? If so, in what sense? And how is the architectural analysis of information processing provided by computational mechanics related to the notion of a pattern? Addressing these questions forms the backbone of Part II.

It is not obvious *a priori* that examining the intrinsic computation of a system is a sensible approach to describing patterns. However, we will demonstrate in Chapter 9 that ϵ-machines provide a more explicit representation of all the patterns, symmetries, and regularities in a spin configuration than is provided by either information theory or statistical mechanics. To do so, we shall calculate statistical mechanical, information theoretic, and computational mechanical quantities for some short-range one-dimensional Ising systems. After a brief review of one-dimensional spin systems and transfer matrix techniques in Chapter 6, we present in Chapter 7 exact, analytic techniques for calculating the information theoretic and computational mechanical quantities of interest. In Chapters 8 - 10 we then proceed to a direct comparison of the three different approaches to discovering and quantifying patterns.

Part II

Results for Finite-Range, One-Dimensional Spin Systems

Chapter 6

Brief Review of Statistical Mechanics of One-Dimensional Spin Systems

In this chapter we review the statistical mechanics of the one-dimensional Ising model. The Ising model is perhaps the most-studied statistical mechanical system. A treatment of the one-dimensional, nearest-neighbor (nn) version of the model can be found in almost any textbook on statistical mechanics, e.g., Refs. [17, 27, 130, 136, 170]. We begin in Sec. 6.1 with a brief examination of the life of Ernst Ising, the man for whom the model is named. (This section is not essential for the rest of the chapter; it is, however, an interesting story.) In Section 6.2 we illustrate the transfer matrix method by using transfer matrices to calculate the partition function of the one-dimensional spin-1/2 nearest-neighbor Ising model. We then discuss how the transfer matrix can be generalized to apply to systems with longer range interactions. Finally, in Sec. 6.3 we state methods for determining the other statistical mechanical quantities using transfer matrix techniques. We conclude by discussing and analyzing these results.

6.1 A Brief History of Ising and his Model

The following biographical comments are largely based on Ref. [84], a very brief biography of Ising by Kobe. For a much more thorough discussion of the Ising model's history, the reader is referred to Ref. [23].

The Ising model was originally proposed in 1920 by Lenz [92] as an attempt to understand ferromagnetism. Ernst Ising, Lenz's student, in his 1924 dissertation solved the one-dimensional version of the model, considered in this chapter. As we shall see, the model fails to "order"—acquire a net magnetization in the absence of an external field—and thus fails to account for ferromagnetism [75]. (Unknown to Ising at the time, the two-dimensional model *does* order.)

After receiving his doctorate, Ising taught at a high school in Strausberg, and then in Crossen on the river Oder [84]. When Hitler assumed power in January of 1933, Ising, being a Jew, lost his position as a civil servant. After a year of unemployment, he taught at a school for Jewish children near Potsdam. He eventually became headmaster of the school, only to witness the school's destruction in a 1938 pogrom.

Ising and his wife Johanna then went to Luxembourg, hoping to emigrate to the United States. However, they were unable to do so, due to U.S. emigration quotas; the Isings endured the war in Europe. Ising was forced to work for the German army from April 1943 until the liberation by allied forces almost a year later. After the war, Ising took a position at the State Teacher's College in Minot, North Dakota. After two years in North Dakota, he accepted a position as a professor of Physics at Bradley University in Peoria, Illinois, a position which he held from 1948 until 1976.

Kobe [84] writes that the model was first referred to as "The Ising Model" in a 1936 paper by Peierls [126]. Amazingly, it wasn't until 1949 that Ising realized that the model he studied for his dissertation had become an important, much-studied system. The Ising model has now achieved the status of a canonical model of an interacting many-body system. A keyword search using the word "Ising" in the INSPEC database from January 1969–March 1998 returns $14,514$ papers! As of 1995, Ernst Ising, at the age of 95, was still alive and living in Peoria.

6.2 The Transfer Matrix Method

We consider now one-dimensional Ising models of the following form:

$$\mathcal{H}(s^N) \equiv -J_1 \sum_{i=1}^{N} s_i s_{i+1} \; - \; J_2 \sum_{i=1}^{N} s_i s_{i+2} \; \cdots \; - B \sum_{i=1}^{N} s_i \tag{6.1}$$

where, for a spin-K system, $s \in \{-K, -K+1, \cdots, K\}$ for $K > 1/2$. If $K = 1/2$, then, by convention, $s \in \{+1, -1\}$. We shall assume that the interactions are of finite range. That is, there is some n_0 such that for all $i > n_0$ $J_i = 0$.

As is well known, the partition function Z_N, Eq. (2.7), for any one-dimensional spin system with finite-range interactions can be expressed in terms of a finite-dimensional matrix, known as the transfer matrix V [89]. Namely, $Z_N = \text{Tr } V^M$, where V^M is the M^{th} power of V and M is proportional to the system size N. The transfer matrix may be viewed as a function of the values of blocks of consecutive spins, with the required block size depending on the interaction range. The dimensionality of the transfer matrix is chosen to be large enough so that the sum over all spin configurations in the partition function, in Eq. (2.7), can be reexpressed as a product of transfer matrices. Hence, the transfer matrix approach effectively decomposes a configuration into a concatenation of contiguous spin blocks.

We will illustrate the transfer matrix method by first showing how the partition function can be calculated using transfer matrices for the spin-1/2, nearest-neighbor model. Then, after briefly discussing how to generalize the transfer matrix method, we will give a few additional results of thermodynamic functions that can be calculated using transfer matrices.

6.2.1 Nearest-Neighbor Case

We consider now the nearest neighbor (nn) spin 1/2 case of the one-dimensional Ising model; $J_i = 0$ in Eq. (6.1) except for $i = 1$ and $s_i \in \{+1, -1\}$. For convenience, we shall assume periodic boundary conditions and identify s_{N+i} with s_i. This choice of a boundary condition does not effect the behavior of the model in the thermodynamic limit, $N \to \infty$ [11].

We begin our calculation of the partition function Z_N by plugging the Hamiltonian into the expression for the partition function, Eq. (2.7):

$$Z_N = \sum_{s^N} \left[e^{\beta J_1 s_1 s_2} e^{\beta B s_1} e^{\beta J_1 s_2 s_3} e^{\beta B s_2} \cdots e^{\beta J_1 s_{N-1} s_N} e^{\beta B s_{N-1}} e^{\beta J_1 s_N s_1} e^{\beta B s_N} \right] . \qquad (6.2)$$

The $J_1 s_N s_1$ term in the exponential arises as a result of the periodic boundary conditions. We then rewrite this expression in the following, suggestive way:

$$\begin{aligned} Z_N = \sum_{s^N} \Big[& e^{\frac{1}{2}\beta B s_1} e^{\beta J_1 s_1 s_2} e^{\frac{1}{2}\beta B s_2} \; e^{\frac{1}{2}\beta B s_3} e^{\beta J_1 s_2 s_3} e^{\frac{1}{2}\beta B s_3} \; \cdots \\ & \cdots \; e^{\frac{1}{2}\beta B s_{N-1}} e^{\beta J_1 s_{N-1} s_N} e^{\frac{1}{2}\beta B s_N} \;\; e^{\frac{1}{2}\beta B s_N} e^{\beta J_1 s_N s_1} e^{\frac{1}{2}\beta B s_1} \Big] . \qquad (6.3) \end{aligned}$$

We now arrive at the crucial step that simplifies this calculation. We define the *transfer matrix* for this system by

$$V(s_i, s_{i+1}) \equiv \exp[J_1 \beta s_i s_{i+1} + \frac{1}{2} B\beta (s_i + s_{i+1})] \,. \qquad (6.4)$$

The transfer matrix V is a 2 x 2 matrix whose row and column indices correspond respectively to the values of a single spin and its nearest neighbor to the right. In matrix notation, V is given by:

$$V \; = \; \begin{pmatrix} e^{\beta(J_1+B)} & e^{\beta J_1} \\ e^{\beta J_1} & e^{\beta(J_1-B)} \end{pmatrix} \qquad (6.5)$$

We adopt the convention that the first (second) row of V corresponds to $s_i = -1(+1)$ in Eq. (6.4); the first (second) column corresponds to $s_{i+1} = -1(+1)$,

It is now not hard to see that Eq. (6.3) may be rewritten using the transfer matrix defined in Eq. (6.4):

$$Z_N \; = \sum_{s^N} \Big[\; V(s_1, s_2) V(s_2, s_3) \cdots V(s_{N-1}, s_N) V(s_N, s_1) \Big] \,. \qquad (6.6)$$

Notice that the summation has the effect of multiplying the transfer matrices together:

$$\sum_{s_{i+1}} V(s_i, s_{i+1}) V(s_{i+1}, s_{i+2}) = V^2(s_i, s_{i+2}) \,. \qquad (6.7)$$

This observation enables us to rewrite Eq. (6.6) as:

$$Z_N = \sum_{s_1} V^N(s_1, s_1) = \mathrm{Tr}[\mathrm{V^N}] \, . \tag{6.8}$$

Equation (6.8) is of use because the trace of a product of matrices is basis independent. Since the V of Eq. (6.4) is symmetric, we can diagonalize V and evaluate the trace in this basis. We obtain:

$$Z_N = \lambda_1^N + \lambda_2^N \, , \tag{6.9}$$

where λ_1 and λ_2 are the eigenvalues of V.

The Perron-Frobenius theorem guarantees that the largest eigenvalue of a positive matrix whose components are finite is always real, positive, and nondegenerate, and that the components of the left and right eigenvectors corresponding to the largest eigenvalue are real. We denote this largest eigenvalue by λ_1. (For a physicist's "proof" of the Perron-Frobeinus theorem, see Ref. [17, p. 66]. For a mathematician's "proof" see Ref. [138].) Since the transfer matrix V is positive and finite, the Perron-Frobenius theorem applies and V's largest eigenvalue is nondegenerate. This nondegeneracy enables us to determine the free energy per site F, Eq. (2.9);

$$F = -\frac{T}{N} \log \left[\lambda_1^N + \lambda_2^N \right] \, , \tag{6.10}$$

$$= \; -\frac{T}{N}\log\lambda_1^N\left[1+(\frac{\lambda_2}{\lambda_1})^N\right]\;. \tag{6.11}$$

In the $N \to \infty$ limit, the second term on the right hand side vanishes since $\lambda_1 > \lambda_2$ and we see that

$$F \;=\; -T\log\lambda_1\;. \tag{6.12}$$

Summarizing, introducing the transfer matrix enabled us to reduce the partition function to the trace of 2 x 2 matrix. This results in Eq. (6.12); the free energy per site is given by the product of the temperature and the logarithm of the dominant eigenvalue of V. From the free energy, other thermodynamic functions, such as the magnetization m, susceptibility χ, and the entropy $\mathbf{S}$ readily follow.

6.2.2 Generalizing the Transfer Matrix Method

We have seen that for the special case of a spin-1/2 system with nn interactions, the partition function can be evaluated by grouping the terms in the Hamiltonian so that the partition function can be rewritten as a product of 2 x 2 matrices. For longer range and/or higher spin models, this technique still works, but higher-dimensional matrices become necessary. For range-R interactions the spins must be grouped into blocks of R consecutive spins. For a spin-K system, the row and column indices then run over the $(2K+1)^R$ possible values the spins can assume in a block of R sites. We shall denote the R-spin blocks by η; the composite variable η can thus assume

$(2K+1)^R$ possible values. Only for the special case of a nearest neighbor ($R = 1$) interaction does $\eta = s$, a single spin. Subscripts on the spin blocks will indicate lattice site, not the particular value of the spin block. The transfer matrix connecting the i^{th} and $(i+1)^{\text{st}}$ spin blocks is denoted by $V(\eta_i, \eta_{i+1})$.

For higher dimensional transfer matrices it turns out that the matrix is not always symmetric, and hence we cannot guarantee that it can be diagonalized. Nevertheless, an extension of the Perron-Frobenius theorem due to Ruelle [138] allows us to conclude that Eq. (6.12) still holds. That is, the largest eigenvalue of V still dominates the trace of Eq. (6.8), despite the nonexistence of a basis in which V is diagonal. Moreover, the dominant left and right eigenvectors have nonnegative components.

As mentioned above, for longer range systems the transfer matrix grows in size, and it is usually not possible to determine an analytic expression for V's dominant eigenvalue and eigenvectors. Instead, the transfer matrix's eigenvalues and eigenvectors must be found numerically. Nevertheless, the key results mentioned in this chapter—Eq. (6.12) above and Eqs. (6.19) and (6.25) below—still hold. These expressions for the free energy density, the two-spin correlation function and the correlation lengths are valid for *any* system that can be described by a finite-dimensional transfer matrix.

For a given system there are a number of ways to construct a transfer matrix that describes its statistical mechanics; see, e.g., Ref. [170]. A general method for forming the transfer matrix for a one-dimensional spin system is elegantly described by Dobson in Ref. [49]. Below, we shall assume that V has been constructed to add on the effects of R spins per matrix operation, where R is again the system's interaction range. By $u_1^{\mathcal{R}}$ ($u_1^{\mathcal{L}}$) we denote the right (left) eigenvector corresponding to V's largest eigenvalue λ_1, normalized so that the inner product of $u_1^{\mathcal{R}}$ and $u_1^{\mathcal{L}}$ is unity. In the following, we shall drop the subscript 1 in situations where doing so results in no ambiguity.

6.3 Results for the One-Dimensional, Nearest Neighbor Model

We conclude our review of the statistical mechanics of one-dimensional spin systems by stating some additional results obtainable via transfer matrices for the nn case. In Sec. 6.3.1 we then discuss these results. The results mentioned below are well-known and can be found in almost any statistical mechanics text; we recommend Yeomans [170] for a particularly clear presentation.

For the nn one-dimensional, spin-1/2 Ising model, the transfer matrix is given by Eq. (6.4). The eigenvalues of V are given by

$$\lambda_{1,2} = e^{\beta J_1}\cosh\beta B \pm \sqrt{e^{2\beta J_1}\sinh^2\beta B + e^{-2\beta J_1}}\ . \qquad (6.13)$$

The eigenvectors of V are given by

$$u_1^{\mathcal{L}} = u_1^{\mathcal{R}} = \begin{pmatrix} a_+ \\ \\ a_- \end{pmatrix} , \tag{6.14}$$

and,

$$u_2^{\mathcal{L}} = u_2^{\mathcal{R}} = \begin{pmatrix} a_- \\ \\ -a_+ \end{pmatrix} , \tag{6.15}$$

where,

$$a_\pm = \frac{1}{2}\left(1 \pm \frac{e^{\beta J_1}\sinh\beta B}{\sqrt{e^{2\beta J_1}\sinh^2\beta B + e^{-2\beta J_1}}}\right) . \tag{6.16}$$

Using Eq. (6.12), one can obtain the free energy per site;

$$F = -T\log\left(e^{\beta J_1}\cosh\beta B + \sqrt{e^{2\beta J_1}\sinh^2\beta B + e^{-2\beta J_1}}\right) \tag{6.17}$$

We see in Eq. (2.13) that the magnetization follows from F by taking the derivative of F with respect to B. One obtains

$$m = \frac{e^{\beta J_1}\sinh\beta B}{\sqrt{e^{2\beta J_1}\sinh^2\beta B + e^{-2\beta J_1}}} . \tag{6.18}$$

It is also possible to obtain the two-spin correlation function $\Gamma(r)$, Eq.(2.14) via the transfer matrix technique. The calculation is somewhat involved; see Refs. [170]

or [17] for details. Here, we simply state the result:

$$\Gamma(r) \;=\; \sum_{i=2}^{\mathrm{dim}V} \left(\frac{\lambda_i}{\lambda_1}\right)^r (u_1^{\mathcal{L}}, \widehat{S}u_i^{\mathcal{R}})\,(u_i^{\mathcal{L}}, \widehat{S}u_1^{\mathcal{R}}) \;, \tag{6.19}$$

where $(\bullet, \bullet)$ indicates inner product and $\widehat{S}$ is the spin operator; $\widehat{S}$ applied to $u_{\mathcal{R}}$ returns a vector whose i^{th} component corresponds to the value of the left-most spin of the i^{th} particular value of the R-spin block. For the spin-1/2, nn case, where $s_i \in \pm 1$,

$$\widehat{S} \;=\; \left(\begin{array}{cc} 1 & 0 \\ 0 & -1 \end{array}\right) \;. \tag{6.20}$$

In Eq. (6.19), u_i refers to the i^{th} eigenvector, not the i^{th} component of the dominant eigenvector.

Using Eq. (6.19), it is not hard to obtain an expression for the structure factors $S(q)$, defined as the discrete Fourier transform of the correlation function $\Gamma(r)$. To do so, one must use the well known relation (see, for example formula 19.43 of Ref.[150]):

$$\sum_{n=0}^{\infty} a^n \cos n\alpha \;=\; \frac{1 - a\cos\alpha}{1 - 2a\cos\alpha + a^2} \;. \tag{6.21}$$

One finds, after some simplification:

$$S(q) \;=\; A\frac{1 - a\cos q}{1 - 2a\cos q + a^2} \;, \tag{6.22}$$

where

$$A = \frac{e^{-2J_1/T}}{(e^{2J_1/T}\sinh^2 B/T) + e^{-2J_1/T}} \; , \tag{6.23}$$

and

$$a = \frac{\lambda_2}{\lambda_1} \; . \tag{6.24}$$

From the expression for the two-spin correlation, Eq. (6.19), the correlation length, the exponential decay rate of $\Gamma(r)$ defined by Eq. (2.17), follows immediately:

$$\xi^{-1} = -\log\left(\frac{\lambda_1}{\lambda_2}\right) \; . \tag{6.25}$$

Note that a calculation of $\Gamma(r)$ requires a determination of all the eigenvalues and eigenvectors of V whereas only the two largest eigenvalues are needed for a calculation of the correlation length ξ.

6.3.1 Discussion of Results

What do these results tell us about the one-dimensional Ising model? Let's begin by considering the magnetization m. Taking the $B \to 0$ limit in Eq. (6.18), we see that the magnetization vanishes for all $T > 0$ in the absence of an external field B. In other words, the system does not acquire a net magnetic moment unless there is an external field to bias the spins. Hence, we see that the one-dimensional version of this model does not admit ferromagnetism.

In more modern language, we say that the one-dimensional finite-range Ising model fails to exhibit *spontaneous symmetry breaking.* The Ising Hamiltonian (in any dimension) with zero external field B possesses a global spin-flip symmetry; the energy (i.e. the Hamiltonian) is unchanged if all the values of the spins are flipped. Flipping a spin is equivalent to multiplying its value by -1. Despite this symmetry, in the two-dimensional Ising model the magnetization assumes a nonzero value for all temperatures below the (nonzero) critical temperature T_c. A nonzero magnetization indicates an excess of either up or down spins. Hence, a system with $m \neq 0$ does not posses a spin flip symmetry.

For the one-dimensional nearest-neighbor model, the fact that the zero-field magnetization vanishes for all non-zero temperatures tells us that the spin-flip symmetry of the Hamiltonian is not broken. As such, there is no continuous phase transition to an ordered, ferromagnetic state. (Phase transitions will be reviewed in Chapter. 12.) Accordingly, the correlation length ξ is finite for all $T > 0$. This can easily be seen by looking at Eq. (6.25), which tells us that ξ is proportional to the logarithm of the ratio of the two largest eigenvalues of the transfer matrix. As noted above, the Perron-Frobenius theorem asserts that the largest eigenvalue of the transfer matrix will always be unique. Thus ξ is finite for all positive temperatures. This result holds for any one-dimensional finite-range Ising system, since the transfer matrix for any such system is finite and positive, and hence falls under the jurisdiction of the Perron-Frobenius theorem. The conclusion, then, is that no finite-range one-dimensional Ising system exhibits a continuous phase transition.

The lack of a phase transition for the one-dimensional Ising model is often cited as evidence for dismissing it as uninteresting or irrelevant. However, despite this lack of a phase transition, the one-dimensional Ising model nevertheless does undergo some important structural changes as system parameters are varied. Here, we will briefly examine the behavior of the two-spin correlation function $\Gamma(r)$ and the correlation length ξ for the nearest neighbor one-dimensional Ising model. In chapters 8–10, the structure present in the Ising model will be discussed at length.

In Fig. 6.1 we have plotted the two-spin correlation function $\Gamma(r)$ as a function of the spin separation r for the parameter values $J_1 = 1.0$, $B = 0.05$, $T = 0.5$. We can see that $\Gamma(r)$ is not zero—a clear indication that there are correlations present in the system. More importantly, note the relatively slow rate at which $\Gamma(r)$ decreases as the separation r between spins increases. The correlation length ξ, defined in Eq. (2.17) as the exponential rate of decay of $\Gamma(r)$ is approximately 4.92 lattice sites for the system shown in Fig. 6.1. This indicates that the configurations contain ordered clusters of, on average, around 5 sites.

What is particularly noteworthy about Fig. 6.1, however, are the questions it leaves unanswered. What is the nature of the ordered clusters? Are the spins all aligned, do they alternate up and down, or do they exhibit an even more complicated pattern? Statistical mechanics typically answers these questions by appealing to the structure factors $S(q)$, defined Eq. (2.22) as the discrete Fourier transform of the

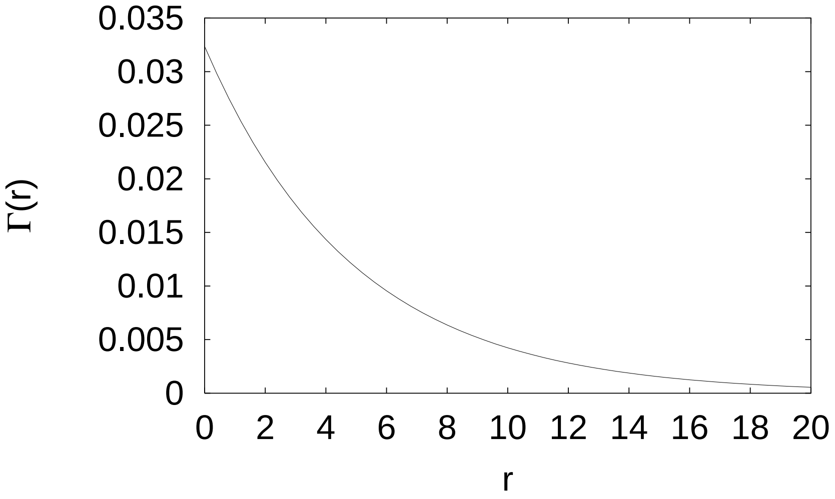


Figure 6.1: The two-spin correlation function $\Gamma(r)$ versus r. The parameter values are $J_1 = 1.0$, $B = 0.05$, and $T = 0.5$. Note the exponential decay of $\Gamma(r)$. For these parameters $\xi \approx 4.92$ lattice sites.

two-spin correlation function. However, we shall see in chapters 8–10, the structure factors do not fully describe the structure present in the one-dimensional Ising system. To adequately capture the Ising system's patterns, the information and computation theoretic techniques of chapters 4 and 5 are needed. In the following chapter, we illustrate how the transfer matrix methods can be used to calculate these information and computation theoretic quantities for finite-range one-dimensional spin systems.

Chapter 7

Calculational Methods and Illustrative Results

In this chapter we present a technique for calculating the information theoretic measures of memory and structure presented in Chapters 4 and 5. To do so, we shall employ the transfer matrices introduced in the previous chapter. In Secs. 7.1.1 and 7.1.2 we determine the recurrent causal states and the state-to-state transition probabilities for finite-range, one-dimensional spin system. Then, in Secs. 7.1.3-7.1.5 we determine the spin system's statistical complexity, entropy density, and excess entropy.

We then illustrate our results for nearest-neighbor (nn) and next nearest-neighbor (nnn) systems in Sec. 7.2 and 7.3. We discuss the nn case in considerable detail, considering, in turn, para-, ferro- and antiferromagnetic couplings. In Chapters 8-10 we shall carry out a direct comparison of information theoretic, statistical mechanical, and computational measures of structure and pattern.

7.1 Calculational Methods

7.1.1 Determination of recurrent causal states

We begin by determining the recurrent causal states from a range-R spin system determined by V. To do this we will need to form conditional probabilities as in Eq. (5.7). In particular, we must find an expression for the probability that L consecutive spins take on the particular values $s_i, s_{i+1}, \cdots, s_{i+(L-1)}$. For convenience, we let $L = RL'$ where $L' > 0$ is an integer that indexes the contiguous R-blocks in the lattice; that is,

$$\eta_i = s_{Ri} s_{Ri+1} \cdots s_{R(i+1)-1} , \; i = 0, 1, \ldots, L' - 1 \; . \tag{7.1}$$

This choice does not affect the results, but simplifies the following derivations. After constructing a transfer matrix that adds on the effects of R spins per matrix operation (see Sec. 6.2, one can use the Boltzmann distribution of Eq. (2.1) to obtain:

$$\Pr(s_0, s_1, \cdots, s_{L-1}) = \frac{u^{\mathcal{R}}_{\eta_{L'}} u^{\mathcal{L}}_{\eta_1}}{\lambda^{(L'-1)}} \prod_{i=1}^{L'-2} V(\eta_i, \eta_{i+1}) \; . \tag{7.2}$$

That is, for a given block of L spins, the probability is a product of components of the transfer matrix and its principal eigenvalue and eigenvectors. Each particular configuration $s_0 s_1 \cdots s_{L-1}$ specifies unique values of the contiguous R-spin blocks $\eta_0, \eta_1, \cdots, \eta_{L'-1}$ in the configuration. To evaluate the right-hand side of Eq. (7.2), the components of the matrices and vectors are chosen by the η variables that correspond to the particular spin variables on the left-hand side; that is, according to Eq. (7.1).

Consider an infinite configuration split at s_0 and left- and right-half configurations of length L on either side:

$$\overset{\leftarrow}{s}^L \equiv s_{-L} s_{-L+1} \ldots s_{-2} s_{-1} \tag{7.3}$$

and

$$\vec{s}^L \equiv s_0 s_1 \ldots s_{L-2} s_{L-1} \; . \tag{7.4}$$

Now,

$$\Pr(\vec{s}^L | \overset{\leftarrow}{s}^L) = \frac{\Pr(\vec{s}^L, \overset{\leftarrow}{s}^L)}{\Pr(\overset{\leftarrow}{s}^L)} \; . \tag{7.5}$$

Using Eq. (7.2), the definitions of $\overset{\leftarrow}{s}^L$ and $\vec{s}^L$, and Eqs. (7.3) and (7.4) in Eq. (7.5) we have, after some simplifying,

$$\Pr(\vec{s}^L | \overset{\leftarrow}{s}^L) = \frac{u^{\mathcal{R}}_{\eta_{L'-1}}}{\lambda^{L'} u^{\mathcal{R}}_{\eta_{-1}}} \prod_{i=-1}^{L'-2} V(\eta_i, \eta_{i+1}) \; . \tag{7.6}$$

Recall that we view this as a function over all possible length-L right-half configurations $\overrightarrow{s}^L$ conditioned on a particular length-L left-half configuration $\overleftarrow{s}^L$. Analyzing this equation is the key step in determining the causal states.

Notice that Eq. (7.6) indicates that of all the spin blocks in $\overleftarrow{s}^L = \eta_{-L'}, \ldots, \eta_{-1}$, $\Pr(\overrightarrow{s}^L|\overleftarrow{s}^L)$ only depends on the single spin block η_{-1}. All the other spin blocks η in Eq. (7.6) are members of $\overrightarrow{s}^L$. That is, the probability distribution over right-half configurations depends only on the value of the left-most (closest) neighboring block. This result holds for any $L > R$. Hence,

$$\Pr(\overrightarrow{s}^L|\overleftarrow{s}^L) = \Pr(\overrightarrow{s}^L|\eta_{-1}) \; . \tag{7.7}$$

Expressed informally, the values of $s_0, \ldots, s_{L-1}$ are "shielded" from $s_{-L}, \ldots, s_{-R-1}$ by spin block $\eta_{-1} = s_{-R} \cdots s_{-1}$, $\overrightarrow{s}^L$'s leftmost R neighboring spins. This observation was made in a different context by Baker [9].

This somewhat surprising result can be explained physically as a direct consequence of the range-R interactions in the Hamiltonian. The probability of a right-half configuration $\overrightarrow{s}$ depends only on its energy. Of all the spins in $\overleftarrow{s}^L$, only the spins in the η_{-1} block—i. e., the block that neighbors $\overrightarrow{s}$—contributes to the energy and hence to the probability of $\overrightarrow{s}$.

Recall that two left-half configurations are considered equivalent if and only if they give rise to the same distribution of right-half configurations conditioned on having seen those particular left halves. The equivalence classes induced by this relation are identified as the causal states. Thus, Eq. (7.6) tells us that for a spin-K system with range R interactions there are at most $(2K+1)^R$ recurrent causal states corresponding to the $(2K+1)^R$ possible values of a single spin block: $|\mathcal{S}^{(\mathcal{R})}| \leq (2K+1)^R$. Recall that the recurrent causal states are the equivalence classes in the limit that $K, L \to \infty$ in Eq. (5.7). In determining the causal states, say by successively increasing L from 0, Eq. (7.6) shows us that the set of causal states will not change once $L > R$. For any $L > R$, the conditional distribution of Eq. (7.6) depends only on η_{-1}, as indicated by Eq. (7.7).

To complete our determination of the recurrent causal states, we must make sure that each different value of η_{-1} actually gives rise to a *different* $\Pr(\overrightarrow{s}^L|\eta_{-1})$. That is, we must check for all different spin-block pairs, $\eta_{-1} \neq \eta'_{-1}$, that

$$\Pr(\overrightarrow{s}^L|\eta_{-1}) \neq \Pr(\overrightarrow{s}^L|\eta'_{-1}) \ , \tag{7.8}$$

for at least one $\overrightarrow{s}^L \in \mathcal{A}^L$. Note that $\overrightarrow{s}^L$ can be decomposed into a telescoping product over its η_i's; that is,

$$\Pr(\overrightarrow{s}^L|\eta_{-1}) = \prod_{i=-1}^{L'-2} \Pr(\eta_{i+1}|\eta_i) \ . \tag{7.9}$$

Looking at the future conditional probability distribution, Eq. (7.6), we see that the

Eq. (7.8) may be written as

$$\frac{V(\eta_{-1}, \eta_0)}{u^{\mathcal{R}}_{\eta_{-1}}} \neq \frac{V(\eta'_{-1}, \eta_0)}{u^{\mathcal{R}}_{\eta'_{-1}}} \; , \tag{7.10}$$

for at least one η_0, the next spin-block.

It should be emphasized that we are fixing two particular values for the rightmost spin block in the left half, η_{-1} and η'_{-1}, and comparing the distribution over all possible values of $\vec{s}^L$ or its surrogate block η_0. When $\eta_{-1} \neq \eta'_{-1}$, if Eq. (7.10) holds for at least one η_0, then the conditional distributions are distinct. This, in turn, means that the causal states are in a one-to-one relation with the values of R-spin blocks. If this is not the case, then we've found two distinct blocks, η_{-1} and η'_{-1}, that lead to the same conditional distribution $\Pr(\vec{s}^L|\bullet)$. Therefore, (i) η_{-1} and η'_{-1} are in the same equivalence class and (ii) there are fewer recurrent causal states than there are spin blocks. For the Ising systems considered here, condition Eq. (7.10) is almost always met; if system parameters are randomly chosen, Eq. (7.10) holds with probability 1. The immediate conclusion is that the set of R-spin blocks $\{\eta\}$ is the set of recurrent causal states. The exceptions, though, are notable, as we will see later on.

We can simplify the notation of Eq. (7.10) by dropping the R-block index when referring to the transfer matrices and its eigenvectors. We can then express Eq. (7.10) in terms of the components of the transfer matrix and the eigenvectors.

That is, we change

$$V(\eta_{-1}, \eta_0) \to V_{ij} \ , \tag{7.11}$$

where $i, j = 0, 1, \ldots, (2K+1)^R - 1$ index the rows and columns of V and correspond to the values of η_{-1} and η_0, respectively. Using this simpler notation, the condition for distinct conditional distributions Eq. (7.10) may be rewritten as

$$V_{ik}/u_i^{\mathcal{R}} \neq V_{jk}/u_j^{\mathcal{R}} \ , \tag{7.12}$$

for at least one k. If this is satisfied, then the recurrent causal state probabilities are given by:

$$\Pr(\mathcal{S}_i^{(R)}) = u_i^{\mathcal{R}} u_i^{\mathcal{L}} \ , \ i = 0, 1, \ldots, (2K+1)^R - 1 \ . \tag{7.13}$$

We have now determined an upper bound on C_0 and C_μ for a spin-K system with R^{th} nearest neighbor interactions: $C_\mu \leq C_0 \leq R \log_2(2K+1)$. This result indicates that this class of spin systems is a severely restricted subset of ϵ-machines. For example, the number of causal states is finite for all parameter values.

7.1.2 Causal state transitions

Now that we have found the recurrent causal states, our task is to determine the probabilities for transitions between them: $T_{\alpha\beta}^{(s)} = \Pr(\mathcal{S}_\beta, s | \mathcal{S}_\alpha)$, where the indices run over only the recurrent causal states.

Recall that transitions between causal states are *deterministic* in the sense that knowledge of the next spin determines the next causal state; that is, $\Pr(\mathcal{S}_\beta|s, \mathcal{S}_\alpha) = 1$ if the transition is allowed and zero otherwise. Thus, it suffices to know $\Pr(s|\mathcal{S}_\alpha)$ to determine $T^{(s)}_{\alpha\beta}$.

It follows from Eq. (7.2) that

$$\Pr(\eta_0|\eta_{-1}) = \frac{1}{\lambda} V(\eta_{-1}, \eta_0) \frac{u^{\mathcal{R}}_{\eta_0}}{u^{\mathcal{R}}_{\eta_{-1}}} \ , \tag{7.14}$$

where $\eta_0 = s_0 s_1 \cdots s_{R-1}$ is the contiguous, non-overlapping R-spin block to the right of $\eta_{-1} = s_{-R} s_{-R+1} \cdots s_{-1}$. To obtain $\Pr(s_0|\eta_{-1})$ from $\Pr(\eta_0|\eta_{-1})$ we must sum over all the spin variables in η_0 except for s_0. Hence:

$$\Pr(s_0|\eta_{-1}) = \sum_{s_1} \cdots \sum_{s_{R-1}} \frac{1}{\lambda} V(\eta_{-1}, \eta_0) \frac{u^{\mathcal{R}}_{\eta_0}}{u^{\mathcal{R}}_{\eta_{-1}}} \ . \tag{7.15}$$

For a next nearest-neighbor ($R = 2$) system we have, for example,

$$\begin{aligned} \Pr(\uparrow|\eta_{-1}) &= \Pr(\uparrow\downarrow|\eta_{-1}) + \Pr(\uparrow\uparrow|\eta_{-1}) \\ &= \frac{1}{\lambda} V(\eta_{-1}, \uparrow\downarrow) \frac{u^{\mathcal{R}}_{\uparrow\downarrow}}{u^{\mathcal{R}}_{\eta_{-1}}} + \frac{1}{\lambda} V(\eta_{-1}, \uparrow\uparrow) \frac{u^{\mathcal{R}}_{\uparrow\uparrow}}{u^{\mathcal{R}}_{\eta_{-1}}} \ . \end{aligned} \tag{7.16}$$

$T^{(s)}_{\alpha\beta}$ follows immediately from Eq. (7.15). The transient states and their transition probabilities can be determined using the method given in App. D. Taken altogether, then, the recurrent and transient states plus their transition probabilities constitute an ϵ-machine for the spin system described by the transfer matrix V. We shall give examples of spin system ϵ-machines in the following sections.

7.1.3 Spin system statistical complexity

In a previous section we identified the recurrent causal states as the possible values of the spin blocks η, assuming Eq. (7.12) is satisfied. Recalling that the statistical complexity is the Shannon entropy over the asymptotic causal state distribution, we may use Eqs. (5.23) and (7.13) to obtain:

$$C_\mu = - \sum_{i=0}^{|\mathcal{S}^{(R)}|-1} u_i^{\mathcal{R}} u_i^{\mathcal{L}} \log_2 (u_i^{\mathcal{R}} u_i^{\mathcal{L}}) \ . \tag{7.17}$$

Eq. (7.17) is equivalent to setting $L = R$ in Eq. (4.1),

$$C_\mu = H(R) \ . \tag{7.18}$$

That is, the statistical complexity is the Shannon entropy $H(R)$ of the R-spin block distribution. As already noted, $H(R)$, divided by R to give a density, is not the entropy density h_μ, even if $R = 1$.

7.1.4 Spin system entropy density

Since the probability of a spin block depends only on the value of that block's nearest neighbor, Eq. (4.4) for the entropy density reduces to

$$h_\mu = \frac{1}{R} H[\eta_i | \eta_{i-1}] \; , \tag{7.19}$$

a form for h_μ discussed in App. B. Using Eq. (7.2) we find that

$$h_\mu = \frac{1}{R} \left(\log_2 \lambda - \lambda^{-1} \sum_{i,j=0}^{|\mathcal{S}^{(R)}|-1} u_i^{\mathcal{R}} u_j^{\mathcal{L}} V_{ji} \log_2 [V_{ji}] \right) \; . \tag{7.20}$$

Although not apparent, it is straightforward to show that, for systems described by a finite-dimensional transfer matrix, Eq. (7.20) is equivalent to the more familiar expression for entropy density,

$$h_\mu = - \lim_{N \to \infty} \frac{\partial}{\partial T} \left(- \frac{T}{N} \log Z\right) \; , \tag{7.21}$$

using $Z = \text{Tr } V^N$. Our method for calculating the entropy density by using Eq. (4.4) has also been employed by Lindgren [99].

7.1.5 Spin system excess entropy

The range-R interactions also lead to a compact expression for the excess entropy. Recall that $\mathbf{E}$ may be expressed as the mutual information between the left and right halves of a configuration. Because only the neighboring R-spin block of one half influences the distribution over the other half, it follows that:

$$\mathbf{E} = I[\overleftarrow{S};\overrightarrow{S}] = I[\eta_i;\eta_{i+1}] = I[\mathcal{S}_0;\mathcal{S}_R] \ , \tag{7.22}$$

where $\mathcal{S}_0, \mathcal{S}_1, \ldots, \mathcal{S}_R$ is a spatial sequence of recurrent causal states, so that $\mathcal{S}_R$ denotes the causal state seen R spins after seeing $\mathcal{S}_0$. To calculate $\mathbf{E}$ using Eq. (7.22), we need the marginal distribution of $\mathcal{S}$ and the joint distribution of $\mathcal{S}_0$ and $\mathcal{S}_R$. The former is just the asymptotic distribution over causal states, $\Pr(\mathcal{S})$, as given by Eq. (7.13). The joint distribution follows by applying the stochastic connection matrix T;

$$\begin{aligned}
\Pr(\mathcal{S}_0, \mathcal{S}_R) &= \sum_{\mathcal{S}_1,\cdots,\mathcal{S}_{R-1}} \Pr(\mathcal{S}_0) \mathrm{T}_{\mathcal{S}_0\mathcal{S}_1} \mathrm{T}_{\mathcal{S}_1\mathcal{S}_2} \cdots \mathrm{T}_{\mathcal{S}_{R-1}\mathcal{S}_R} \\
&= \Pr(\mathcal{S}_0) \mathrm{T}^{R-1}_{\mathcal{S}_0\mathcal{S}_R} \ ,
\end{aligned} \tag{7.23}$$

where T^{R-1} is the $(R-1)^{\mathrm{th}}$ power of T. From Eqs. (7.13), (7.22), and (7.23), $\mathbf{E}$ follows readily. We should emphasize, however, that Eq. (7.22) is not completely general—it applies only to R-range one-dimensional spin systems for which Eq. (7.12) holds.

In terms of the transfer matrix, an expression for $\mathbf{E}$ follows by inserting

Eq. (7.2) into Eq. (7.22) and simplifying;

$$\mathbf{E} \;=\; -\log_2 \lambda + \lambda^{-1} \sum_{i,j=0}^{|\mathcal{S}^{(R)}|-1} u_i^{\mathcal{R}} u_j^{\mathcal{L}} V_{ji} \log_2[V_{ji}] \;-\; \sum_{j=0}^{|\mathcal{S}^{(R)}|-1} u_j^{\mathcal{R}} u_j^{\mathcal{L}} \log_2[u_j^{\mathcal{R}} u_j^{\mathcal{L}}] \; . \tag{7.24}$$

One can also calculate $\mathbf{E}$ and h_μ by determining an expression for $H(L)$ in terms of V and using Eq. (4.9). Doing this, we get formulae that agree with those derived above.

7.1.6 Relationships between spin system memory measures

Note that these results—Eqs. (7.17), (7.20), and (7.24)—establish an explicit version of the equality in Eq. (5.35) between $\mathbf{E}$ and C_μ mentioned above; namely,

$$C_\mu = \mathbf{E} + R h_\mu \; , \tag{7.25}$$

where R is the range of interaction and again assuming that Eq. (7.12) is satisfied. Eq. (7.25) is a consequence of the information theoretic identity,

$$H[\eta_i] = I[\eta_i ; \eta_{i+1}] + H[\eta_{i+1} | \eta_i] \; . \tag{7.26}$$

This result, Eq. (7.25), also applies to finite-step Markov chains in which blocks over the observed alphabet $\mathcal{A}$ are in $1-1$ correspondence with the internal-state blocks $\mathcal{S}^R$. Note that the fair coin and the noisy period-2 examples violate this condition and so Eq. (7.25) does not hold for them. (See Table 1.)

Eq. (7.25) shows us that $\mathbf{E}$, C_μ, and h_μ are not independent for the finite-range systems considered here. As such, we will focus mostly on $\mathbf{E}$ and h_μ for the remainder. When discussing ϵ-machines below, we will consider mainly their detailed structure and will not focus on the single number C_μ.

Finally, since mutual information, and thus $\mathbf{E}$, is a nonnegative quantity, we note that

$$C_\mu \geq h_\mu \ , \tag{7.27}$$

recalling the restrictions on Eq. (7.25). It might seem puzzling that the amount of information carried by the ϵ-machine—"Which causal state is the process in?"—is *larger* than the information available (on average) from individual spin observations. However, C_μ and h_μ simply measure different types of information.

7.2 Spin-1/2 Nearest-Neighbor Systems

Starting with the transfer matrix, the preceding section developed a general method for determining the causal states, constructing an ϵ-machine, and calculating the statistical complexity, entropy density, and excess entropy. The results describe all finite-range one-dimensional spin systems. In this section we apply these results to the simple case of nearest-neighbor (nn) spin-1/2 systems; the Hamiltonian is given by Eq. (6.1) with all the J_i's except for J_1 set to zero. The main goal of this section is to illustrate the use of our methods and to allow the reader to gain familiarity with the quantities we've defined in earlier sections. In subsequent chapters we shall consider longer range models, compare and contrast **E** and ϵ-machines with the measures of structure found in statistical mechanics, and, then, draw some general conclusions about the behavior of these different quantities.

7.2.1 ϵ-Machines for the spin-1/2, nearest-neighbor Ising model

For the special case of a spin-1/2 system with nearest-neighbor interactions, and for those parameter values where Eq. (7.12) holds, the corresponding ϵ-machine is shown in Fig. 7.1. The transition probabilities are obtained from Eq. (7.15) and the transient state construction technique of App. D. State **A** is the start state and is the only transient state. States **B** and **C** are recurrent.

The transition matrices $\{T^{(s)} : s \in \mathcal{A}\}$ are given by:

$$T^{(\downarrow)} = \begin{pmatrix} 0 & \Pr(\downarrow) & 0 \\ 0 & \Pr(\downarrow \mid \downarrow) & 0 \\ 0 & \Pr(\downarrow \mid \uparrow) & 0 \end{pmatrix} \tag{7.28}$$

and

$$T^{(\uparrow)} = \begin{pmatrix} 0 & 0 & \Pr(\uparrow) \\ 0 & 0 & \Pr(\uparrow \mid \downarrow) \\ 0 & 0 & \Pr(\uparrow \mid \uparrow) \end{pmatrix} . \tag{7.29}$$

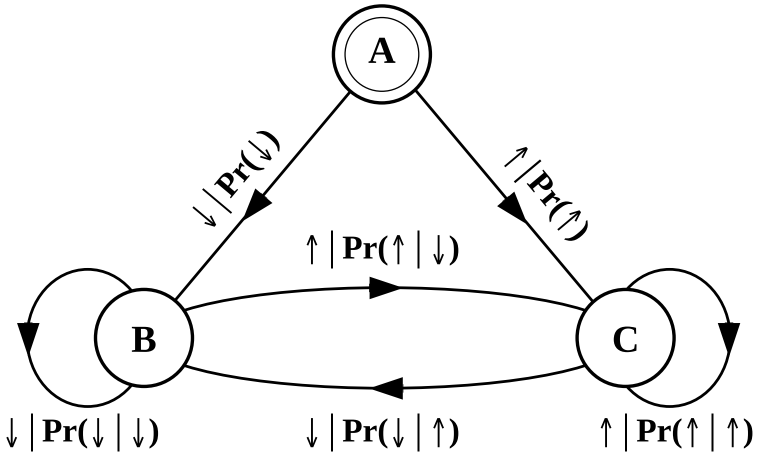


Figure 7.1: The spin-1/2 Ising ϵ-machine. The double-circled causal state **A** is the start state. It is a transient state, never visited again after the first transition. The two initial transitions give the probabilities of isolated up and down spins.

The stochastic connection matrix T, being the sum of the above two matrices, is:

$$\mathrm{T} = \begin{pmatrix} 0 & \Pr(\downarrow) & \Pr(\uparrow) \\ 0 & \Pr(\downarrow \mid \downarrow) & \Pr(\uparrow \mid \downarrow) \\ 0 & \Pr(\downarrow \mid \uparrow) & \Pr(\uparrow \mid \uparrow) \end{pmatrix} . \tag{7.30}$$

The components $\mathrm{T}_{\alpha\beta}$ give the probability of making a transition from causal state α to causal state β. As before, we use the convention that the numerical values of the index α correspond to the alphabetical indices of the causal states in the natural way; $\alpha = 0$ corresponds to causal state **A**, $\alpha = 1$ to **B**, etc. For example, the probability of making a transition from **B** to **C** is given by $T^{\uparrow}_{12} = \mathrm{T}_{12} = \Pr(\uparrow \mid \downarrow)$. The matrices given by Eqs. (7.28) and (7.29) are equivalent to the ϵ-machine shown in Fig. 7.1.

In terms of the system parameters J_1, B, and T, the elements of the stochastic connection matrix can be calculated explicitly via Eq. (7.15) and the transient state construction technique of App. D. We find:

$$\mathrm{T} = \begin{pmatrix} 0 & \frac{1+m}{2} & \frac{1-m}{2} \\ 0 & \kappa^{-1} e^{\frac{-B+J_1}{T}} & 1 - \kappa^{-1} e^{\frac{-B+J_1}{T}} \\ 0 & 1 - \kappa^{-1} e^{\frac{B+J_1}{T}} & \kappa^{-1} e^{\frac{B+J_1}{T}} \end{pmatrix} . \tag{7.31}$$

The normalization factor κ is given by:

$$\kappa = e^{\frac{J_1}{T}} \cosh(\frac{B}{T}) + \sqrt{e^{\frac{-2J_1}{T}} + e^{\frac{2J_1}{T}} \sinh^2(\frac{B}{T})} \ , \tag{7.32}$$

and m, the magnetization, is:

$$m = \frac{e^{\frac{J_1}{T}} \sinh(\frac{B}{T})}{\sqrt{e^{\frac{-2J_1}{T}} + e^{\frac{2J_1}{T}} \sinh^2(\frac{B}{T})}} \ . \tag{7.33}$$

7.2.2 Paramagnet

We now apply these results to some particular cases, beginning with a paramagnet, $J_1 = 0$. It is easy to check that Eq. (7.12) is not satisfied. Hence, there is only one causal state, and $C_\mu = 0$ for all temperatures and all values of the external field B. Physically, this is because at $J_1 = 0$ there is no coupling at all between the spins; a spin exerts no influence on the value of its neighbors. As a result, there is only one conditional distribution:

$$\Pr(\overrightarrow{s}^L|s_{-1} = \downarrow) \;=\; \Pr(\overrightarrow{s}^L|s_{-1} = \uparrow)\ . \tag{7.34}$$

The ϵ-machine for the paramagnet is shown in Fig. 7.2. In terms of B and T,

$$\Pr(\uparrow) = (1/2)\frac{e^{B/T}}{\cosh(B/T)}\ , \tag{7.35}$$

and $\Pr(\downarrow) = 1 - \Pr(\uparrow)$. The start state is recurrent for this particularly simple process. If there is no external field B to bias the spins, then $\Pr(\downarrow) = \Pr(\uparrow) = 1/2$ and the ϵ-machine of Fig. 7.2 is identical to the the fair coin machine of Fig. 5.1.

Since knowledge of a spin carries no information about the value of its neighbors, the excess entropy also vanishes for the paramagnet. If there is no external field B to bias the spins, all configurations are equally likely and $h_\mu = 1$. As B increases from 0, the configurations are biased toward ↑, and the entropy density monotonically decreases. Note that for $|B| < \infty$ and $T > 0$, h_μ can take on all possible values except for 0; $0 < h_\mu \leq 1$. Yet for all B, $C_\mu = \mathbf{E} = 0$. This simple example illustrates how excess entropy and statistical complexity are measuring a property that is clearly distinct from the entropy density—they are different from randomness. We can also see how the process of determining the causal states factors out the randomness in the system.

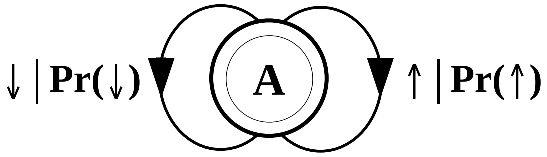


Figure 7.2: The ϵ-machine for a paramagnet. Pr(↑) and Pr(↓) depend on B and T. However, $C_\mu = \mathbf{E} = 0$, for $T > 0$.

7.2.3 Ferromagnet

For ferromagnetic coupling, $J_1 > 0$, Eq. (7.12) holds for all temperatures except zero and infinity. In this range the recurrent causal states may be identified with the values of a single spin. An ϵ-machine for typical parameter values is shown in Fig. 7.3.

At infinite temperature, thermalization dominates and the spins effectively decouple; all configurations are equally likely. Thus, as for the paramagnet, there is only one conditional distribution,

$$\Pr(\overrightarrow{s}^L | s_{-1} = \downarrow) = \Pr(\overrightarrow{s}^L | s_{-1} = \uparrow) \; . \tag{7.36}$$

As a result, there is only one causal state and C_μ and $\mathbf{E}$ vanish. The ϵ-machine for the infinite temperature ferromagnet, shown in Fig. 7.4, is identical to the fair coin machine, Fig. 5.1.

At zero temperature there are no thermal fluctuations and the spins are locked in their ferromagnetic ground state: all spins align with the external field. This

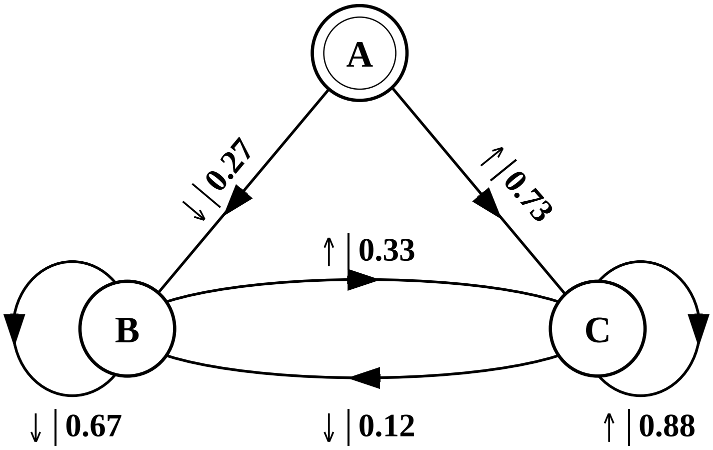


Figure 7.3: A typical ϵ-machine for a ferromagnet. $J_1 = 1.0$, $B = 0.3$, and $T = 1.50$. $C_\mu = 0.72$ bits, $\mathbf{E} = 0.16$ bits, and $h_\mu = 0.56$ bits per site. Note the high probability of "self-transitions" from **B** to **B** and from **C** to **C**, a manifestation of the relatively large ferromagnetic interaction.

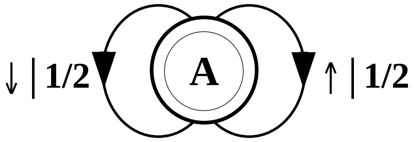


Figure 7.4: The ϵ-machine for $T = \infty$. There is no memory, of any type: $C_\mu = \mathbf{E} = 0$ bits. The infinite temperature machine is identical for the ferro-, para-, and antiferromagnets so long as J_1 and B remain finite.

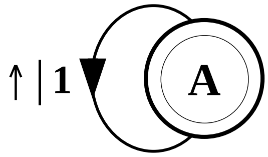


Figure 7.5: The ϵ-machine for the ferromagnetic ground state, $T = 0$. $C_\mu = \mathbf{E} = 0$ bits for $0 < B < \infty$ and $0 < J_1$.

situation is exactly the same as the period-1 example considered in Sec. 5.2.2. Thus, as shown in Fig. 7.5, there is only one causal state and only one transition. The excess entropy, statistical complexity, and entropy density all vanish for this simple, trivially predictable system.

The statistical complexity and excess entropy for a nearest-neighbor ferromagnet are plotted as a function of temperature in Fig. 7.6. C_μ increases monotonically as a function of temperature until $T = \infty$ (not shown). There, as mentioned above, the couplings between spins become negligible, causing the two causal states to merge into one, yielding $C_\mu = 0$. The monotonic increase in between these two extremes is due to the distribution over the two causal states, **B** and **C**, becoming more uniform as the temperature is increased. Since C_μ is the Shannon entropy of the causal state distribution, it is maximized when the distribution is uniform. This distribution is approached as one nears, but is not at, $T = \infty$.

In Fig. 7.7 we plot C_μ and $\mathbf{E}$ parametrically as a function of the randomness, as measured by h_μ. This plot is referred to as the complexity-entropy diagram [44].

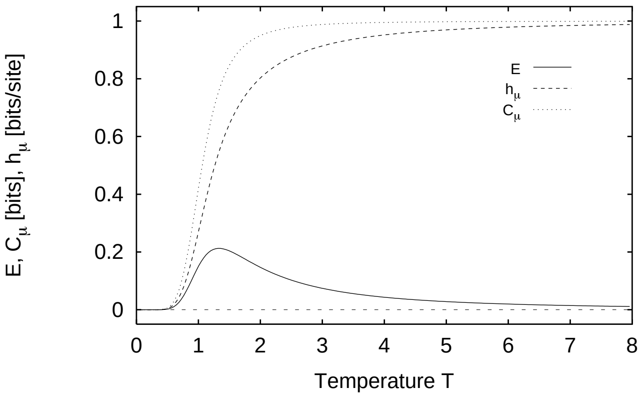


Figure 7.6: C_μ, h_μ, and $\mathbf{E}$ as a function of T for the nn spin-1/2 ferromagnet. B was held at 0.20 and $J_1 = 1$.

The benefit of this type of plot is that it is free of the external control parameters—temperature, coupling strength, and external field. Thus, the complexity-entropy diagram gives direct access to a system's information processing capabilities and provides a common set of coordinates with which to compare the information processing properties of systems with different architectures and control parameters. For example, in Ref. [38] we used the complexity-entropy diagram to compare the configurations generated by one-dimensional Ising systems with the sequences generated by the symbolic dynamics of the logistic map. The complexity-entropy diagram is also discussed in Chapter 11, which was originally published as Ref. [55].

For the ferromagnet, we see in Figs. (7.6) and (7.7) that $\mathbf{E}$ has a maximum in a region between total randomness ($h_\mu = 1$) and complete order ($h_\mu = 0$). At low temperatures (and, hence, low h_μ) most of the spins align with the magnetic field. At

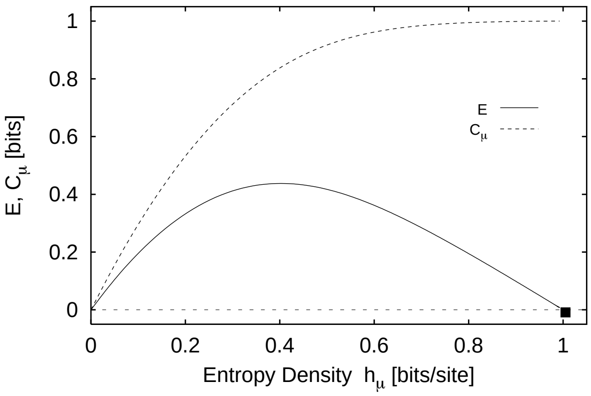


Figure 7.7: The complexity-entropy diagram for a ferromagnet; C_μ and $\mathbf{E}$ plotted parametrically against h_μ. The coupling constant and field were fixed—$J_1 = 1.0$ and $B = 0.05$—as T was varied. At $h_\mu = 1$ ($T = \infty$), $C_\mu = 0$ bits; this is denoted by the square token.

high temperatures, thermal noise dominates and the configurations are quite random. In both regimes one half of a configuration contains very little information about the other half. For low h_μ, the spins are fixed and so there is no information to share. For high h_μ, there is much information at each site; a roughly equal number of spins point up and down, so the single spin uncertainty is quite high. However, this information is uncorrelated with all other sites. Thus, the excess entropy is small in these temperature regimes. In between the extremes, however, $\mathbf{E}$ has a single maximum at the temperature where spin coupling strength balances the thermalization. The result is a maximum in the system's spatial memory.

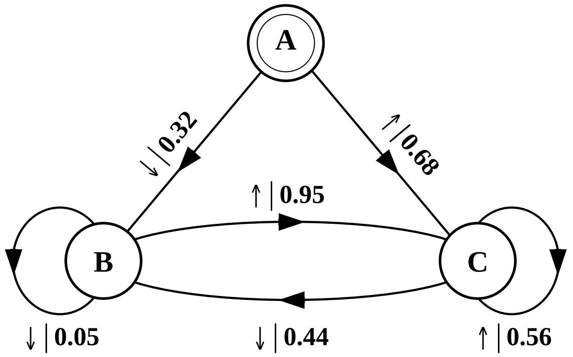


Figure 7.8: A typical ϵ-machine for an antiferromagnet (AFM). $J_1 = -1.0$, $B = 1.8$, and $T = 1.5$. Giving $C_\mu = 0.92$ bits, $\mathbf{E} = 0.27$ bits, and $h_\mu = 0.65$ bits per site. Note the relatively strong interstate coupling.

7.2.4 Antiferromagnet

An ϵ-machine for a typical antiferromagnetic (AFM) coupling ($J_1 < 0$) is shown in Fig. 7.8. Note that "topologically" the AFM machine is identical to the FM machine of Fig. 7.3—the states and their connectivity are identical. The difference between the two systems lies in the probabilities of the transitions. The FM shows a high probability for self-transitions; $\Pr(\mathbf{B} \to \mathbf{B}) = 0.67$ and $\Pr(\mathbf{C} \to \mathbf{C}) = 0.88$. These self-loops are responsible for the FM pattern: aligned spins. For the AFM, the self-loops are relatively weaker; $\Pr(\mathbf{B} \to \mathbf{B}) = 0.05$ and $\Pr(\mathbf{C} \to \mathbf{C}) = 0.56$, with the high value of the latter being due partially to the high $B(= 1.8)$. This indicates a stronger tendency for spins to be anti-aligned, as expected for a system with AFM couplings.

The high temperature behavior of the excess entropy is similar for both the AFM and FM. (See Fig. 7.9.) Thermal fluctuations destroy all correlations and $\mathbf{E}$

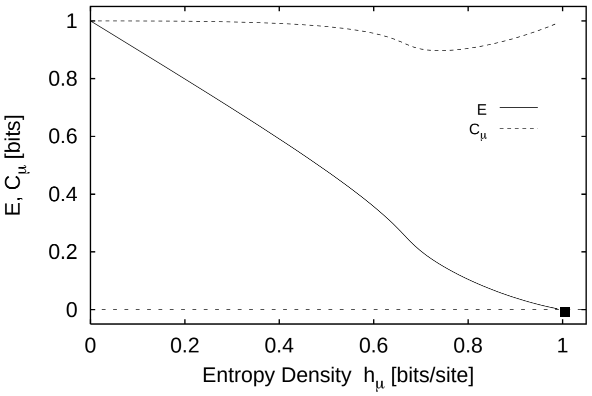


Figure 7.9: The complexity-entropy diagram for an antiferromagnet. The temperature was varied as J_1 and B were held constant at $J_1 = -1.0$ and $B = 1.80$. As was the case for the ferromagnet, at $h_\mu = 1$ ($T = \infty$), $C_\mu = 0$ bits; this is denoted by the square token.

vanishes. The low T behavior differs, however, as one might expect given the different ground states exhibited by models with ferro- and antiferromagnetic couplings: in the FM ground state, all the spins are aligned, while the ground state of the AFM consists of alternating up and down spins. The latter is, of course, the period-2 configuration given by Eq. (5.5) that we considered back in Sec. 5.2 with Fig. 5.4. In Fig. 7.9 we chose the antiferromagnetic coupling to be strong enough so that an antiferromagnetic ground state persists despite the presence of an external field. In the antiferromagnetic ground state the spatial configurations thus store one bit of information about whether the odd or even sites contain up spins. Accordingly, as can be seen in Fig. 7.9, $\mathbf{E} \to \log_2 2 = 1$ bit as $h_\mu \to 0$.

7.2.5 General remarks

For different couplings and field strengths a range of $\mathbf{E}$ vs. h_μ relationships can be realized but their form is similar to those shown in Figs. (7.7) and (7.9); $\mathbf{E}$ either shows a single maximum or decreases monotonically. It is always the case, though, that $\mathbf{E}$ is bounded from above by $1 - h_\mu$, which follows immediately if C_μ is set equal to its maximum value, 1 bit, in Eq. (7.25).

We have demonstrated this upper bound explicitly in Ref. [55][1], where we show a plot of the excess entropy versus entropy density for a nn Ising system with randomly chosen system parameters. Li [96] has performed a similar study using several probabilistic automata.

In Sec. 5.6, C_μ was presented as a measure of structure. It is perhaps surprising, then, that it behaves so differently from $\mathbf{E}$. As h_μ increases, one might expect C_μ to reach a maximum, as does $\mathbf{E}$, and then decrease as the increased thermalization merges causal states that were distinct at lower temperatures. However, Fig. 7.7 shows a monotonic increase in C_μ with h_μ for the FM. To understand this, recall that the number of recurrent causal states does not change as T is varied between zero and infinity. For the nn spin-1/2 Ising model, the number of causal states remains fixed at two. What *does* change as T is varied are the causal state probabilities $\Pr(\mathcal{S}_\alpha)$. For the FM, as the temperature rises the distribution $\Pr(\mathcal{S}_\alpha)$ becomes more uniform

[1] Contained here as Chapter 11

and so C_μ grows. This growth continues until T becomes infinite, since only there do the causal states collapse into one, at which point C_μ vanishes.

For the AFM the situation is a little different. At $T = 0$ there are two recurrent causal states corresponding to the two spatial phases of the alternating up-down pattern. The probability of these causal states is equal. Hence, we see a low temperature statistical complexity of $C_\mu = 1$ bit. At high (but finite) temperatures, the thermal fluctuations dominate; the anti-ferromagnetic order is lost, but the distribution over causal states is still relatively uniform, so the statistical complexity remains large. (As with the FM, at $T = \infty$ the two causal states merge and C_μ falls to zero.) Between these extremes there is a region where the influence of the external field dominates, biasing the configurations. This is reflected in a bias in the causal state probabilities and C_μ dips below 1, as seen in Fig. 7.9.

The tendency for C_μ to remain large for large values of h_μ is due to a more general effect, which follows from Eq. (7.25): $C_\mu = \mathbf{E} + Rh_\mu$. The memory needed to model a process (or for the process to produce its configurations) depends not only on the observed memory of the configurations generated, as measured by $\mathbf{E}$, but also on its randomness, as measured by h_μ. It is important to note, however, that C_μ is driven up by thermalization *not* because the model attempts to account for random spins in a configuration and not because the process must develop substantial memory resources to produce random spin values. Rather, C_μ rises with h_μ because $\Pr(\mathcal{S}_\alpha)$ becomes more uniform as the temperature increases. This reflects the fact that knowing in which causal state the process is becomes more informative in this regime.

7.3 Spin-$1/2$ next-nearest neighbor Ising system

We now discuss the causal states and ϵ-machines for a spin system with nearest and next-nearest neighbor (nnn) interactions. That is, the Hamiltonian is given by

$$\mathcal{H}(s^N) \equiv -J_1 \sum_{i=1}^{N} s_i s_{i+1} - J_2 \sum_{i=1}^{N} s_i s_{i+2} - B \sum_{i=1}^{N} s_i \tag{7.37}$$

This system, capable of richer behavior than the nn system discussed above, is an important addition for discussing the detection of patterns and structure in subsequent sections. It will serve as our primary example when we compare computational mechanical, statistical mechanical, and information theoretic approaches to structure.

For the nnn system the recurrent causal states are, assuming Eq. (7.12) is satisfied, in a one-to-one relation with the four possible values of a block of two spins. A general ϵ-machine for this system is shown in Fig. 7.10. The connection matrix for this system is given by:

$$\mathrm{T} = \begin{pmatrix} 0 & \Pr(\downarrow) & \Pr(\uparrow) & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \Pr(\uparrow \mid \downarrow) & 0 & \Pr(\downarrow \mid \downarrow) & 0 \\ 0 & 0 & 0 & 0 & \Pr(\uparrow \mid \uparrow) & 0 & \Pr(\downarrow \mid \uparrow) \\ 0 & 0 & 0 & 0 & \Pr(\uparrow \mid \downarrow\uparrow) & 0 & \Pr(\downarrow \mid \downarrow\uparrow) \\ 0 & 0 & 0 & 0 & \Pr(\uparrow \mid \uparrow\uparrow) & 0 & \Pr(\downarrow \mid \uparrow\uparrow) \\ 0 & 0 & 0 & \Pr(\uparrow \mid \downarrow\downarrow) & 0 & \Pr(\downarrow \mid \downarrow\downarrow) & 0 \\ 0 & 0 & 0 & \Pr(\uparrow \mid \uparrow\downarrow) & 0 & \Pr(\downarrow \mid \uparrow\downarrow) & 0 \end{pmatrix} . \quad (7.38)$$

There are several features to note about the ϵ-machine of Fig. 7.10. First, the machine has more states than the nearest-neighbor system. This is a direct consequence of the longer-range interactions in the nnn system. Second, there are four recurrent causal states, **D** through **G**, and thus the topological complexity Eq. (5.24) is $C_0 = \log_2 4 = 2$ bits. As noted in Chapter 5.6.1, the topological complexity sets an upper bound on the statistical complexity C_μ; hence, $C_\mu \leq 2$ bits. Lastly, setting h_μ to its minimum value, 0, in Eq. (7.25), we also obtain a bound on the excess entropy; $\mathbf{E} \leq 2$ bits. We will discuss the behavior of $\mathbf{E}$ for the nnn system in the next chapter.

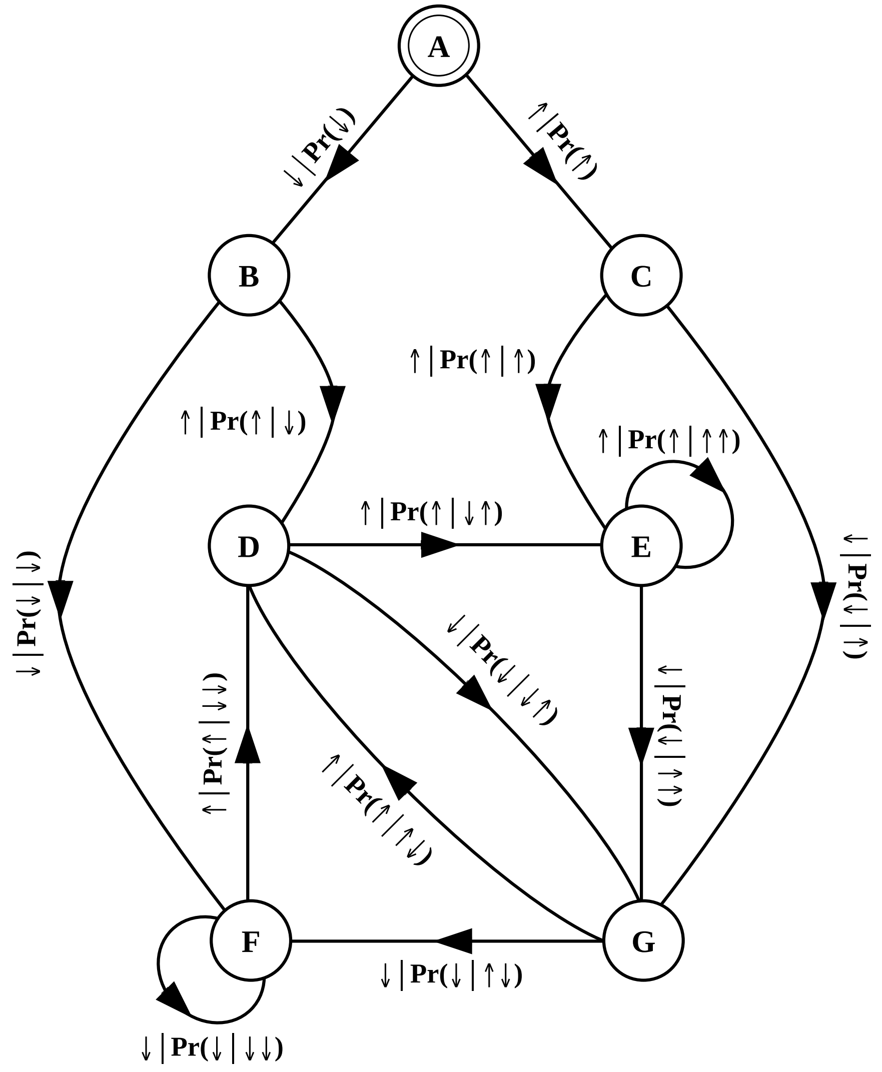


Figure 7.10: ϵ-machine for a next-nearest neighbor spin-1/2 Ising model. There are three transient states $\mathcal{S}^{(T)} = \{\mathbf{A}, \mathbf{B}, \mathbf{C}\}$ and four recurrent states $\mathcal{S}^{(R)} = \{\mathbf{D}, \mathbf{E}, \mathbf{F}, \mathbf{G}\}$. The start state is $\mathbf{A}$.

Chapter 8

Excess Entropy is a Wavelength-Independent Measure of Periodic Structure

We now begin the first of three chapters in which we explicitly compare the different approaches to structure: Information Theoretic, Computational, and Statistical Mechanical. In this chapter, we compare the excess entropy, an information theoretic measure of memory, with the statistical mechanical structure factors introduced in Chapter 2.

Superficially, it might seem that the excess entropy and the structure factors reflect the same feature of a configuration. The excess entropy measures the total mutual information between two halves of a configuration and so may be viewed as the total apparent spatial memory stored in the configuration. The structure factors $S(q)$, defined in Eq. (2.22), are a sum over all two-point correlation functions and thus, in a limited sense, can be viewed as a measure of the total correlation. However, we shall see in this chapter that **E** and $S(q)$ have several important differences.

8.1 Comparing E and $S(q)$

In Fig. 8.1 we have plotted $S(0)$, $S(\pi/2)$, $S(\pi)$, and **E** as a function of the coupling strength J_1 between nearest neighbors for the nnn Ising system just described. The field, temperature, and next-nearest neighbor coupling constant were held fixed at $B = 0.05$, $T = 1.0$, and $J_2 = -1.2$. The structure factors of wavenumbers $0, \pi/2$, and π correspond to wavelengths of spatial periods 1, 4, and 2, respectively.

Let's first analyze the behavior of the structure factors in Fig. 8.1. As J_1 goes to $-\infty$, the thermal fluctuations become negligible and the system is confined to its ground state. The nnn coupling constant J_2 is also negligible in this limit. Hence, the system's ground state is antiferromagnetic with period 2: alternating up and down spins and, in Fig. 8.1, $S(\pi)$ diverges.

For $J_1 > 0$ the nn coupling is ferromagnetic. As J_1 becomes larger than J_2, the system moves through a region of ferromagnetic structure similar to that reflected in Fig. 7.6 and indicated by the $S(0)$ peak in Fig. 8.1. As $J_1 \to +\infty$ the thermal fluctuations and the nnn coupling are again negligible and the system is fixed in its ground state. Here, since $J_1 > 0$, the ground state is ferromagnetic. All spins line up with the external field. As a result, $\langle s_0 s_r \rangle = \langle s \rangle^2$ so all the $\Gamma(r)$'s vanish, yielding a vanishing $S(0)$.

For $|J_1| \ll T$, the thermal fluctuations and the nnn coupling J_2 dominate and the lattice effectively decouples into two non-interacting chains. That is, the even and odd sites do not interact with each other. Since $J_2 < 0$, the ground state in this parameter regime is antiferromagnetic with period 4:

$$\cdots \uparrow\uparrow\downarrow\downarrow\uparrow\uparrow\downarrow\downarrow\uparrow\uparrow\downarrow\downarrow\uparrow\uparrow\downarrow\downarrow\uparrow\uparrow \cdots \tag{8.1}$$

As a result, we see a peak in $S(\pi/2)$ at $J_1 = 0$ in Fig. 8.1. The wavenumber $\pi/2$ corresponds to a period of 4. The structure factors $S(0)$ and $S(\pi)$ are insensitive to structure at this wavelength.

Fig. 8.1 shows that the system exhibits significant structural changes, as indicated by the structure factors, as the parameter J_1 is varied. Notice, however, that analyzing the configurations using only one of the structure factors misses most of the changes that occur in the configurations elsewhere. Excess entropy, however, is not limited to a particular wavelength. As can be seen in Fig. 8.1, **E** picks up the ferromagnetic and both types of antiferromagnetic structure. This feature of the excess entropy is especially noteworthy, as statistical mechanics does not possess a structure factor that is applicable to all such situations.

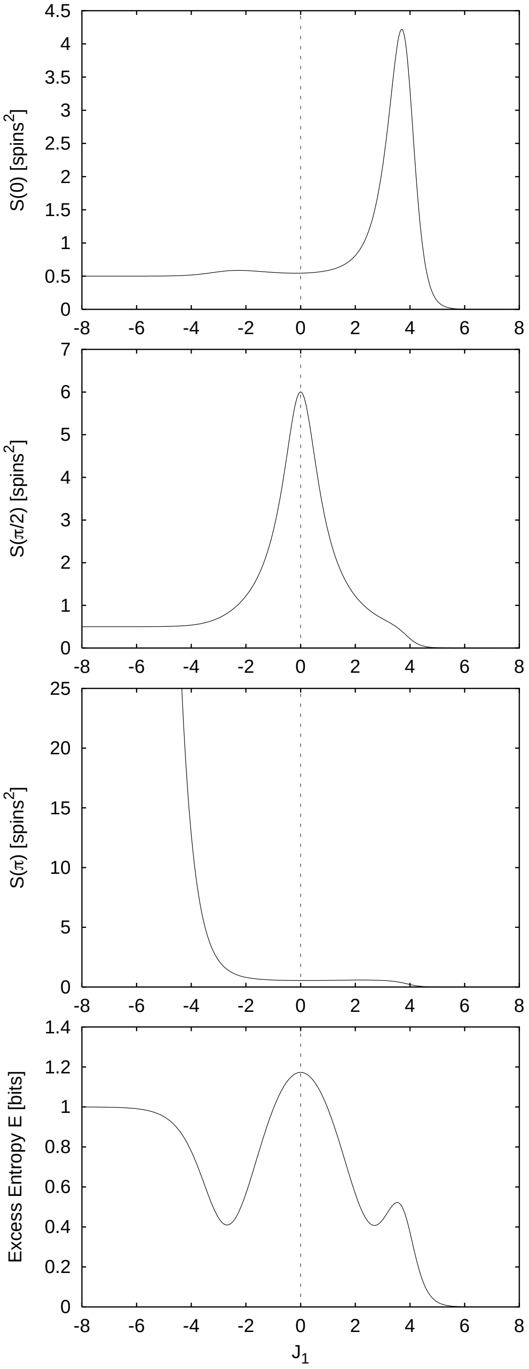


Figure 8.1: The structure factors $S(0)$, $S(\pi/2)$, and $S(\pi)$ and the excess entropy $\mathbf{E}$ versus nearest-neighbor coupling J_1 for a next-to-nearest neighbor 1D Ising system. The parameters are $B = 0.05$, $T = 1.0$, and $J_2 = -1.2$.

The excess entropy is capable of detecting structure at any wavelength because it is a much more "global" function than the structure factors. Although calculation of the structure factor involves summing over all the variables in the chain, the correlations are considered in pairs, since the two-point correlation functions $\Gamma(r)$ are summed over. Excess entropy, on the other hand, is defined as the information that the *entire* left half carries about the *entire* right half. The excess entropy treats the left half and the right half of a configuration as two (very large) composite variables; it does not break them into pairs. In this sense we say that $\mathbf{E}$ is more global than $S(q)$. Conversely, $S(q)$ is somewhat "myopic". By considering only two-point correlations modulated at some wavenumber q, $S(q)$ misses structure that occurs at other wavenumbers and that is due to more-than-two-spin correlations.

The differences between $\mathbf{E}$ and $S(q)$ can be better understood by considering the motivations behind their definition. Structure factors are designed to detect a pattern of a given periodicity. For example, if one performs a numerical experiment to determine the critical point of a paramagnet-antiferromagnet transition, then the antiferromagnetic structure factor $S(\pi)$ is the natural quantity to use to detect the onset of antiferromagnetic ordering. If, however, one is interested in the apparent spatial memory of a configuration, $\mathbf{E}$ is the natural quantity to use.

Simply put, excess entropy and structure factors measure different things: $\mathbf{E}$ measures spatial memory and $S(q)$ detects correlations of a particular periodicity. They behave similarly because the spatial memory of a configuration is relatively large if it has a periodic pattern. In fact, $\mathbf{E} = \log_2 \mathcal{P}$ for a periodic configuration of period $\mathcal{P}$. For the spin systems considered here, a configuration is periodic if and only if its entropy density h_μ vanishes. Thus, if a system has $h_\mu = 0$, this implies that is periodic with $\mathcal{P} = 2^{\mathbf{E}}$.

Since $\mathbf{E}$ and $S(q)$ measure different properties of a configuration, they carry different units. The excess entropy is measured in bits while $S(q)$ has the dimensions of spin-value squared. Note that $\mathbf{E}$ is a function of the distribution of the spin variables and, unlike $S(q)$, does not depend on the values or units of the spin variables. For example, if we were to consider a model where $s_i \in \{\pm 2\}$ instead of $s_i \in \{\pm 1\}$, $S(q)$ would increase by a factor of 4 while $\mathbf{E}$ would remain unchanged. This is a fairly trivial observation but, as has been mentioned elsewhere [95], it emphasizes how mutual information is a more flexible measure of correlation than a correlation function. In other words, "correlation" is best interpreted to be a statement concerning the joint distribution of two variables, not the values those variables can assume.

The fact that $\mathbf{E}$ and $S(q)$ carry different units means that their numerical values are interpreted differently. This is particularly clear in the $J_1 \to -\infty$ behavior of Fig. 8.1. Here, $S(\pi)$ diverges, indicating exact periodicity at $q = \pi$. The excess entropy, however, is finite: $\mathbf{E} = \log_2 2 = 1$ bit, indicating that the configurations store 1 bit of information.

8.2 Trying to Sum $S(q)$ over q

Looking at Fig. 8.1, it appears as if the sum of $S(0)$, $S(\pi/2)$, and $S(\pi)$ might behave like **E**. Indeed, summing these three functions does produce a function that behaves like **E** for this particular system. But summing up the relevant $S(q)$'s still depends on guessing the right q's. A response to this objection might be to sum $S(q)$ over *all* q's. However, if one does this the different phases destructively interfere. As a result:

$$\sum_{q=0}^{N-1} S_N(q) \equiv \sum_{q=0}^{N-1} \sum_{r=0}^{N-1} \Gamma(r) e^{2\pi irq/N} = \Gamma(0) \ . \tag{8.2}$$

All we're left with is $\Gamma(0)$, a "self-correlation" term that is a function of the distribution of a single spin and, hence, clearly is no measure of spatial structure.

Summing over the absolute value of the structure factors, as in

$$\bar{S} \equiv \sum_{q=0}^{\infty} |S(q)| \ , \tag{8.3}$$

also yields a quantity that fails to measure the total correlation of the system. For example, $\bar{S}$ fails to vanish for a paramagnet in the presence of an external field.

One also might be tempted to add together all the connected correlation functions. That is, combine the two-spin, three-spin, four-spin, etc., connected correlation functions. (For a discussion of how to extend the definition of connected correlation functions to more than two spins see, e.g., [17].) However, such a sum either diverges or is simply proportional to the free energy. Neither case leads to a measure of structure. More specifically, one can show that the connected correlation functions are, up to factors of β, the coefficients in the Taylor expansion of $\log Z = \beta F$ in powers of local coupling constants J'_i attached to each site i, where $F = U - T\mathbf{S}$ is the Helmholtz free energy [17]. Thus, a sum of the connected correlation functions corresponds to setting $J'_i = 1$ in this expansion. Since this is outside of the series' radius of convergence, we conclude that this sum will fail to converge unless the series terminates. If the series terminates, however, the resulting sum is a quantity proportional to the free energy F.

8.3 Summary

In summary, Fig. 8.1 illustrates one the central points of this dissertation: excess entropy measures the memory stored in spatial configurations and as such is sensitive to periodic structure at any period. As far as one can tell, statistical mechanics does not possess a quantity that has these properties. The structure factors are sensitive to periodic behavior, but only at particular wavenumbers. Furthermore, the structure factors do not measure memory, carrying units of [spins^2], whereas $\mathbf{E}$ carries units of bits, which often can be usefully interpreted in terms of the pattern or structure in a configuration.

Since any periodic 1D spin configuration has $h_\mu = 0$, a vanishing entropy rate together with a positive excess entropy is an unambiguous indication of periodic order. The two information processing "coordinates", h_μ and $\mathbf{E}$, provide a means for detecting periodicity. But periodicity is just one type of structure. How can we discover and describe structure that isn't periodic, i. e., structure that has a positive entropy density? Moreover, different structures can have the same period. For example, $\cdots \uparrow\downarrow\downarrow\downarrow\uparrow\downarrow\downarrow\downarrow \cdots$ and $\cdots \uparrow\uparrow\downarrow\downarrow\uparrow\uparrow\downarrow\downarrow \cdots$ both have $\mathbf{E} = 2$ bits corresponding to a period of 4.

More generally, "pattern" is not synonymous with memory as measured by $\mathbf{E}$. Knowing the amount of memory stored in spatial configurations does not specify how the memory is organized or used. Put another (obvious) way, knowledge of $\mathbf{E}$ alone does not allow one to reproduce the original configuration ensemble. We shall consider these issues at length now.

Lastly, note that, like $\mathbf{E}$, C_μ serves as a measure of low-entropy memory since since by Eq. (7.25), $h_\mu = 0$ implies $C_\mu = \mathbf{E}$.

Chapter 9

ϵ-Machines Reveal Structural Features in Entropic Processes

In the previous chapter we saw that $\mathbf{E}$ serves to detect periodic—i. e., $h_\mu = 0$, $\mathbf{E} < \infty$—structure. In this chapter we examine systems that have relatively large entropy density, yet still produce highly structured configurations. To describe these systems, $\mathbf{E}$ becomes inadequate and the full apparatus of computational mechanics becomes necessary.

9.1 Discovering and Describing Entropic Patterns

To illustrate this, we consider a spin-1/2 Ising model with next-nearest neighbor interactions, as analyzed in Sec. 7.3. We fix the coupling constants and the temperature at the following values: $J_1 = -3.0$ and $J_2 = -1.0$ at $T = 0.2$. The temperature is small compared to the external parameters. Hence the system is close to its ground state and thermal excitations are small. As B increases, the ground state for the system changes. This can be seen by considering the excess entropy and the entropy density, which are plotted along with C_μ in Fig. 9.1.

The previous section noted that a 1D spin system with $h_\mu = 0$ is periodic with period $\mathcal{P} = 2^{\mathbf{E}}$. For $B < 1/2$, we see in Fig. 9.1 that h_μ vanishes while $\mathbf{E} \approx 1$ indicating a periodic structure of period 2. Similarly, for $B > 8.5$ $\mathbf{E}$ and h_μ vanish, indicating periodicity of period 1. For $4 < B < 6$, $\mathbf{E} \approx 1.59$ and $h_\mu \approx 0$, indicating that the system is in a configuration with period 3, since $\log_2 3 \approx 1.59$. Thus, as B is varied, the system makes transitions between three different spatially periodic ground states.

The transitions between different periodic regimes can also be seen in Fig. 9.2 which shows plots of various structure factors as a function of the external field B. The period-2 structure factor $S(\pi)$ diverges as B falls below 2. (Strictly speaking, the structure factor does not diverge. Since the temperature is nonzero, the structure factor remains finite. "Diverging" structure factors here have values around 10^5.) Above $B = 8$, all structure factors vanish, an indication that the system is in a ferromagnetic ground state. For B between 2 and 8, we see in Fig. 9.2 that the period-3 structure factor $S(2\pi/3)$ diverges. Note that our being able to detect these changes in the periodic structure of the system is due to judicious choices of q for

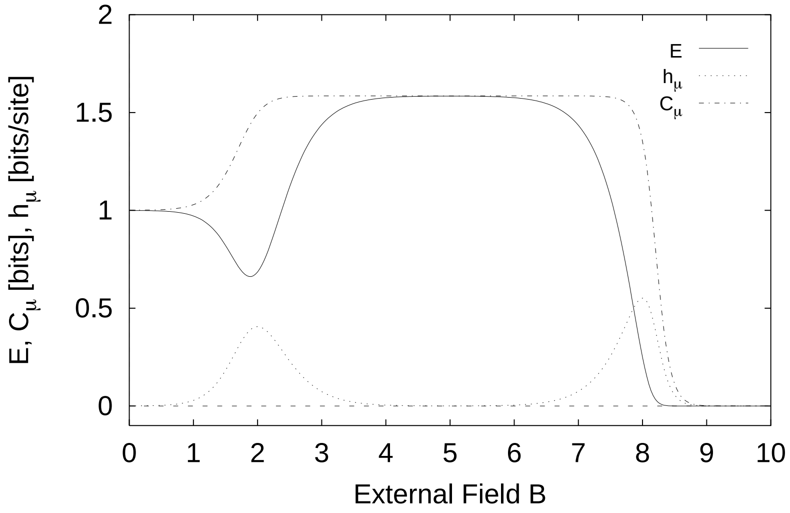


Figure 9.1: The statistical complexity C_μ, excess entropy $\mathbf{E}$, and entropy density h_μ versus external field B for a next-to-nearest neighbor 1D Ising system. The parameters are $J_1 = -3.0$, $J_2 = -1.0$, and $T = 0.2$.

the $S(q)$'s shown. This example again illustrates the utility of $\mathbf{E}$ as a wavelength-independent detector of periodicity.

We now turn our attention to the main question of this section: What is happening during the transitions between these periodic regimes? For $B = 2$ and 8 we see in Fig. 9.1 that the entropy density h_μ is large. Thus, the system is not spatially periodic in these regimes and cannot be well-described by structure factors. Presumably the configurations are some mixture of the periodic ground states that dominate on either side of each transition. But is this the case? And if they are mixtures, how do two periodicities "mix"?

The structure factors do not provide much, if any, clue. Near $B = 8$ in Fig. 9.2

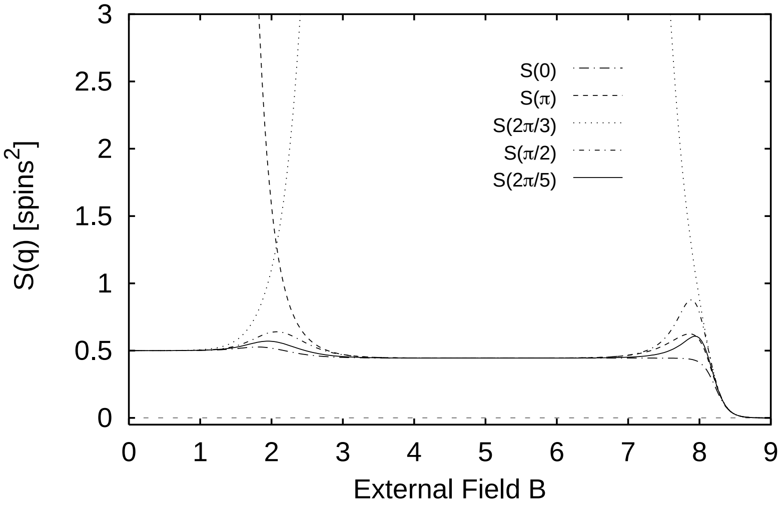


Figure 9.2: The structure factors $S(0)$, $S(\pi)$, $S(2\pi/3)$, $S(\pi/2)$, and $S(2\pi/5)$ versus external field B for a next-to-nearest neighbor 1D Ising system with $J_1 = -3.0$, $J_2 = -1.0$, and $T = 0.2$.

we see gentle peaks in $S(\pi)$, $S(\pi/2)$, and $S(2\pi/5)$, the structure factors for patterns of periods 2, 4, and 5, respectively. And near $B = 2$, we see peaks in the $S(q)$'s for periods 1, 2, and 4. What sort of configuration could produce these structure factor amplitudes? To help us answer this question, we examine the ϵ-machines for the configurations at the transition points.

Fig. 9.3 shows the ϵ-machine for $B = 8.0$. Transitions that occur with a probability of less than 10^{-6} are not shown. Note that this ϵ-machine has only 3 recurrent states, as opposed to the 4 recurrent causal states of the generic nnn ϵ-machine of Fig. 7.10. State **F** has disappeared—it is reached with a probability of less than 10^{-7} and so is not included in Fig. 9.3.

As expected, the ϵ-machine of Fig. 9.3 demonstrates that the transitional structure is indeed a "mixture" of periodic behaviors of periods 1 and 3. States **A**, **B**, and **C** are transient states. The self-loop on state **E** gives the period-1 pattern $\uparrow\uparrow\uparrow$. The $\mathbf{E} \rightarrow \mathbf{G} \rightarrow \mathbf{D} \rightarrow \mathbf{E}$ loop is the period-3 pattern $\downarrow\uparrow\uparrow$. The entropy density for the configurations described by the ϵ-machine of Fig. 9.3 is relatively high: $h_\mu \approx 0.551$ bits per spin. Nevertheless, the configurations have considerable structure—simply calling them random or "mostly random" is unnecessarily crude.

Note that Fig. 9.3 is not the only way for a period-1 and a period-3 pattern to mix. For example, extra $\uparrow$'s could be inserted at both state **E** and state **D**. That is, there could be an additional self-loop on state **D** that occurs with a different probability than the self-loop on state **E**. Thus, the ϵ-machine provides more information than just showing that the configurations are a mixture of period-1 and period-3 patterns; the machine tells us *how* the patterns combine.

Unlike the collection of statistics plotted in Figs. 9.1 and 9.2, the ϵ-machine provides a complete description of the configuration ensemble: The ϵ-machine is capable of statistically reproducing the *entire* original configuration, along with any other realizations consistent with the ensemble.

Recall that, as explained above, the ϵ-machine is a minimal description. First, the procedure of equivalence classing to determine the causal states ensures that the model has the fewest number of states while still accounting for all the causal structure of the system. Second, the model is chosen within the least powerful class that admits

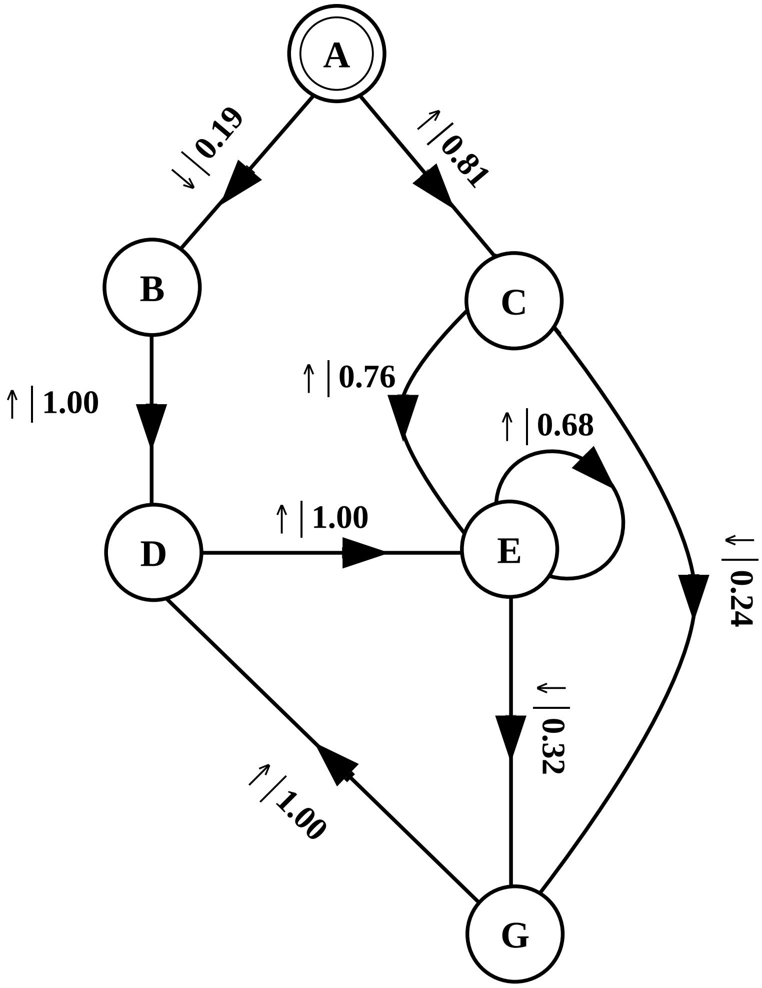


Figure 9.3: The ϵ-machine for the spin system shown in Figs. (9.1) and (9.2) with the external field fixed at $B = 8.0$. The other parameters are $J_1 = -3.0$, $J_2 = -1.0$, and $T = 0.2$.

a finite description of the original process. Thus, in analogy with the group theoretic description of exact symmetries, the ϵ-machine may be viewed as the "irreducible representation" of the approximate symmetries. In this sense we conclude that the ϵ-machine *is* the pattern. Note that to discover the pattern's structure by building the ϵ-machine, no assumptions are made, aside from the translational invariance of the original configuration. That is, determining an ϵ-machine for a system is not a transform for detecting an *a priori* given set of patterns, as is the case with Fourier analysis and the structure factors, for example. Rather, ϵ-machines enable one to discover patterns and structures not assumed beforehand.

9.2 Detecting Entropic Patterns

The ϵ-machine directly reveals important and useful structural information about a configuration. As mentioned above, the machine of Fig. 9.3 reveals how the period-1 and period-3 patterns mix to produce the configurations responsible for the complexities and structure factors observed at $B = 8.0$ in Figs. 9.1 and 9.2. Furthermore, the ϵ-machine structure often can be easily translated to provide a compact description of the configuration in natural language, as already attempted in the preceding paragraphs. In the case just analyzed, configurations may be viewed as a background pattern of $\downarrow\uparrow\uparrow$'s with one or more extra $\uparrow$'s inserted before each $\downarrow$. Typical configurations at $B = 8$ are:

$$\uparrow\uparrow\downarrow\uparrow\uparrow\downarrow\uparrow\uparrow\uparrow\uparrow\uparrow\downarrow\uparrow\uparrow\uparrow\uparrow\uparrow\uparrow\downarrow\uparrow\uparrow\uparrow\downarrow\uparrow\uparrow\uparrow\uparrow\uparrow\downarrow\uparrow\uparrow\uparrow\downarrow\uparrow\uparrow\uparrow\uparrow\downarrow \tag{9.1}$$

and

$$\uparrow\uparrow\downarrow\uparrow\uparrow\downarrow\uparrow\uparrow\uparrow\downarrow\uparrow\uparrow\downarrow\uparrow\uparrow\uparrow\uparrow\downarrow\uparrow\uparrow\uparrow\downarrow\uparrow\uparrow\uparrow\downarrow\uparrow\uparrow\uparrow\uparrow\uparrow\downarrow\uparrow\uparrow\downarrow\uparrow\uparrow\uparrow\ . \tag{9.2}$$

The probability that there are M extra ↑'s inserted is readily gleaned from the ϵ-machine in Fig. 9.3. The probability is given by:

$$\Pr(M \text{ extra} \uparrow \text{'s}) = \left\{ \begin{array}{ll} 0.32 & M = 0 \\ 0.68^M & M > 1 \end{array} \right\} . \tag{9.3}$$

Equivalently, the ϵ-machine tells us how we can construct a different machine that produces similar configurations: a machine with a single state that generates the sequence $\downarrow\uparrow\uparrow$ with probability 0.32 and ↑ with probability 0.68. Note that this is *not* an ϵ-machine since each transition made by the machine does not produce one symbol.

This alternative description lets us construct yet another machine, a *transducer*, that detects which sites are participating in the constituent subpatterns $\alpha = \uparrow^*$ or $\beta = (\downarrow\uparrow\uparrow)^*$, where w^* denotes an arbitrary number of replications of w. This filtering machine, illustrated in Fig. 9.4, defines a function from configurations of spins to sequences over the alphabet $\{\alpha, \beta\}$. Transitions are selected deterministically according to which spin state is read in from the configuration. (See Refs. [40] and [70] for more discussion of building and using these types of transducers.)

Before any symbols are read in, the transducer begins in the start state, labeled **A** in Fig. 9.4. If the first symbol read in is $\uparrow$ the transducer produces the null symbol λ and returns to state **A**. If the symbol is $\downarrow$, then transducer is synchronized to the configuration: it "knows" what causal state the process is in and it outputs the symbol β, indicating that the observed $\downarrow$ is part of the $\downarrow\uparrow\uparrow$ pattern. The next two symbols read in will be $\uparrow\uparrow$ and the transducer makes transitions from **G** to **D** and **D** to **E**. In this manner, the transducer maps the input string of $\uparrow$'s and $\downarrow$'s to a string of α's and β's. For example, the configuration of Eq. (9.1) is mapped to:

$$\lambda\lambda\beta\beta\beta\beta\beta\beta\alpha\alpha\alpha\beta\beta\beta\alpha\alpha\alpha\alpha\alpha\beta\beta\beta\alpha\beta\beta\beta\alpha\alpha\beta \; . \tag{9.4}$$

There are several features of the transducer that give it utility. First, the transducer can be viewed as giving "meaning" to individual spins [32]. Determining if a spin is part of the period-1 or period-3 pattern tells us what role that particular site is playing in the configuration. This is not a trivial observation since an isolated $\uparrow$ is a part of both of the two competing subpatterns.

Second, the transducer provides a way to recognize sequences; that is, to determine if a candidate sequence is statistically identical to the configuration from

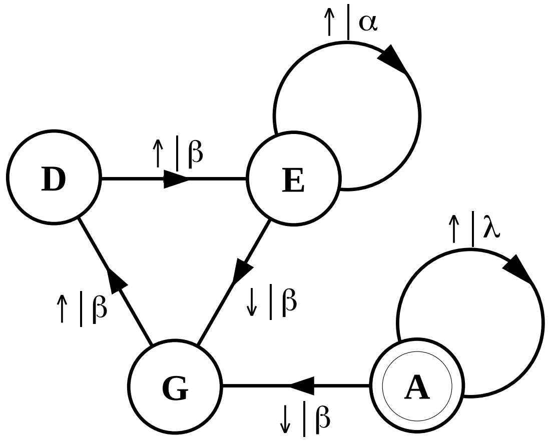


Figure 9.4: A transducer that detects the elemental spin subpatterns $\uparrow\uparrow\uparrow\uparrow \ldots$ and $\downarrow\uparrow\uparrow\downarrow\uparrow\uparrow\downarrow\uparrow\uparrow \ldots$ and labels the lattice sites with the name—α or β, respectively—of the subpattern in which each site participates.

which the ϵ-machine was originally constructed. This recognition process consists of two components; first we determine if the candidate configuration is allowed and then we check to see that the spin blocks within the candidate configuration occur with the correct probabilities.

If, as a transducer reads a configuration, a spin value is encountered for which there is no transition, then that configuration is rejected. We conclude that it is not a member of the configuration ensemble. For example, the sequence $\uparrow\uparrow\downarrow\downarrow$ is rejected since there is no transition leaving state **G** when a $\downarrow$ is read.

To conclude that a configuration $\overleftrightarrow{s}$ is *statistically* consistent with the configuration from which the ϵ-machine was built, $\overleftrightarrow{s}$ must do more than correspond to a path through the transducer. It must also produce the correct percentage of α and β subpatterns. Fig. 9.3 tells us that proper configurations are a type of a biased coin: with probability 0.68 a ↑ is generated and with probability 0.32 a ↓↑↑ is generated. If a configuration produced by the ϵ-machine of Fig. 9.3 is used as input to the transducer of Fig. 9.4, then the fraction f_α of α's in the output produced is:

$$f_\alpha = \frac{0.68}{0.68 + 3 \times 0.32} \approx 0.42 \ . \tag{9.5}$$

Thus, if a configuration is read in to the transducer and the fraction of α's produced approaches 0.42, then we conclude that it is statistically identical to the original one.

For this case, the simple form of the ϵ-machine of Fig. 9.3 lets us easily compute the finite-size scaling of the transducer output. Thus, thinking of the ϵ-machine as a biased coin, it immediately follows that the number of α's expected in a length-N configuration is:

$$\# \text{ of } \alpha'\text{s} \approx 0.42 \times N \pm \sqrt{0.42 \times 0.58 N} \ , \tag{9.6}$$

or,

$$f_\alpha = 0.42 \pm 0.49 N^{-1/2} \ . \tag{9.7}$$

Here we have ignored the number of spins needed to synchronize to the pattern. This simple calculation assumes, consistent with the law of large numbers, that the deviations from f_α are small. The correct way to estimate the fluctuations for ϵ-machines uses methods from large deviation theory, as done in [172].

Analytically calculating (as here) or empirically reconstructing ϵ-machines en-

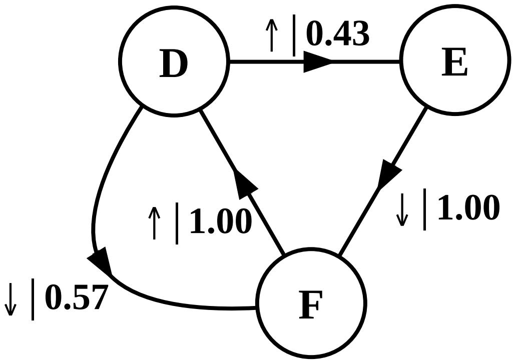


Figure 9.5: The recurrent portion of the ϵ-machine for the system shown in Figs. (9.1) and (9.2) with the external field fixed at $B = 2.0$. The other parameters are $J_1 = -3.0$, $J_2 = -1.0$, and $T = 0.2$.

ables us to discover patterns. Analyzing the ϵ-machine reveals what the patterns are. And the transducer—a simple modification of the ϵ-machine—tells us how we can detect these patterns. One can do more, such as calculate the expected error in the transducer's output for finite-length input strings, as we have just outlined.

A similar structural analysis of the configurations in the $B = 2$ transition region follows from the ϵ-machine of Fig. 9.5. For example, we again see a mixture of two periodic patterns. This time period-2 and period-3 subpatterns combine. Configurations consist of a "background" of ↑↓↑ with a period-2 component of ↓↑'s inserted between the two ↑'s of the background pattern. The probability that M period-2 blocks are inserted is given by:

$$\Pr(M \ \uparrow\downarrow -\text{blocks}) = \left\{ \begin{array}{ll} 0.43 & M = 0 \\ & \\ 0.57^M & M > 1 \end{array} \right\} . \tag{9.8}$$

In summary, the preceding subsections illustrated how ϵ-machines provide a complete, minimal description of the patterns or regularities contained in (entropic) spin configurations. Roughly speaking, they may be viewed as the irreducible representations of the statistical symmetries of the system. As such, an ϵ-machine provides a much more complete and informative description of a pattern than is available within information theory or statistical mechanics. In contrast, Figs. (9.1) and (9.2) do not strike us as being structurally very informative. It's clear from these plots that there is a transition between periodic behaviors, but the specifics of the structural changes are not at all obvious.

The dip in the excess entropy and the peak in the entropy density at the transition regions give a general indication of high-entropy, low-apparent-memory configurations. These structures are not periodic and, thus, are not compactly described by the structure factors that implicitly assume the system has strong periodic components. However, the configurations most certainly are not structureless. The ϵ-machine analyses showed that the constituent periodic patterns mix in very particular ways. This explicit analysis of patterns is not available within the existing frameworks of statistical mechanics or information theory.

Lastly, recall that our description of spin configurations began with a Hamiltonian with nearest and next-nearest neighbor interactions, which in turn led to a 4×4 transfer matrix. The Hamiltonian and the transfer matrix both determine all the information about the system in the sense that they can be used to calculate the probability, and thus the energy, of any configuration. However, neither the Hamiltonian nor the transfer matrix capture the intrinsic information processing in the explicit way an ϵ-machine does nor do they provide a minimal description of the underlying patterns.

Chapter 10

Phenomenological Comparison of Excess Entropy and Statistical Mechanical Quantities

At this point we have reached the two central conclusions we wish to draw from our comparison of information theoretic, computational, and statistical mechanical approaches to structure. First, in chapter 8 we have shown that the entropy density h_μ and the excess entropy $\mathbf{E}$ together serve to detect periodic structure at any wavelength. If $h_\mu = 0$, and $\mathbf{E} < \infty$ then 1D spin systems are exactly periodic with period $\mathcal{P} = 2^{\mathbf{E}}$. Second, in chapter 9 we have seen that the ϵ-machine *is* the underlying pattern in the sense that it is a minimal representation of all the (group and semi-group theoretic) regularities produced by a process. Excess entropy and ϵ-machines complement each other; **E** measures a system's apparent spatial memory, while an ϵ-machine gives direct access to how a system is organized and how it processes information. The causal states, part of an ϵ-machine, reveal the hidden, effective states of a process.

We conclude our comparison of different approaches to structure by explicitly comparing the excess entropy with some of the statistical mechanical functions defined in chapter 2. Specifically, we compare the excess entropy with the correlation length, the specific heat, particular structure factors, and the nearest neighbor correlation function. We shall see that none of these statistical mechanical measure the system's memory in the manner that **E** does.

10.1 Excess Entropy versus Correlation Length

We begin by comparing excess entropy with the correlation length ξ, defined by Eq. (2.17). Qualitatively, their behavior is similar, as can be seen in Fig. 10.1 where they are plotted as a function of temperature for a ferromagnetic system in an external field. These two functions have different units; **E** is measured in bits per site, while ξ is a length measured as a number of lattice sites. Thus, their relative magnitudes cannot be meaningfully compared.

However, we can compare their qualitative behavior as the temperature is varied. Looking at Fig. 10.1, we see that both quantities have a single maximum as

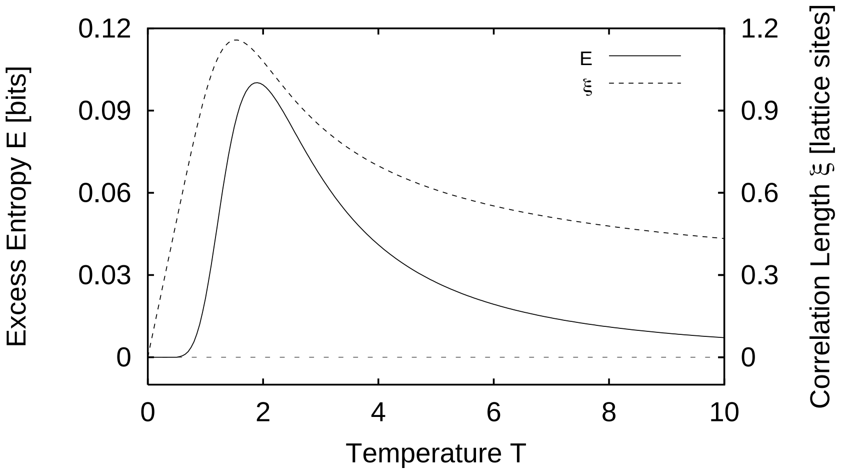


Figure 10.1: $\mathbf{E}$ and ξ versus T for a nn ferromagnet with $J_1 = 1.0$ and $B = 0.5$.

a function of temperature. However, their maxima occur at different temperatures: ξ is maximized at $T \approx 1.55$, while $\mathbf{E}$ reaches a maximum at $T \approx 1.90$. This indicates that they are not related to each other by a simple multiplicative constant. Note that ξ is linear for small T while $\mathbf{E}$ vanishes exponentially.

A more important difference, though, is that $\mathbf{E}$ and ξ have very different physical interpretations. On the one hand, the correlation length ξ measures the *rate* at which correlations between spins decay as a function of increasing distance. The decay rate provides little or no information about how much total correlation or memory is present. The excess entropy, on the other hand, measures the mutual information between two semi-infinite halves of the configuration and thus provides a measure of the total spatial memory of the system.

10.2 Excess Entropy versus Specific Heat

In Fig. 10.2 we plot the specific heat C and the excess entropy $\mathbf{E}$ as a function of temperature. As in the comparison of $\mathbf{E}$ and ξ, they carry different units and so their numerical values can't be compared. Qualitatively, their behavior is more similar. However, they are maximized at different temperatures; C reaches a maximum at $T \approx 1.55$ while $\mathbf{E}$, as above, attains its maximum value at $T \approx 1.90$.

Nevertheless, these two quantities measure very different properties of the system. As mentioned in discussing Eq. (2.26), the specific heat measures the system's energy fluctuations. While these fluctuations may be evidence of correlations between different degrees of freedom, leading to a large $\mathbf{E}$, this most certainly is not always the case. For example, a paramagnet has a nonzero specific heat that shows a single maximum just as C does in Fig. 10.2. Yet a paramagnet, by definition, has no correlations between spins. Accordingly, the excess entropy of a paramagnet vanishes for all values of the temperature and the external field. Since the specific heat does not vanish for such a system, it is clear that C cannot be viewed as providing any general indication of spatial structure.

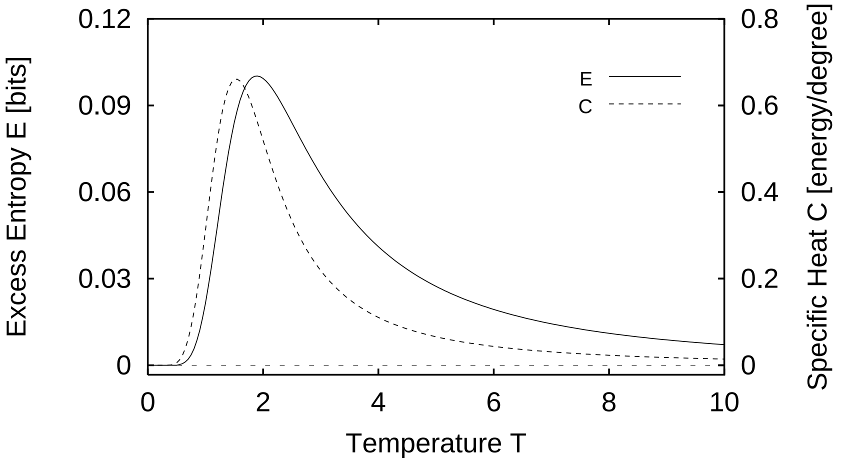


Figure 10.2: **E** and specific heat C versus T for a nn ferromagnet with $J = 1.0$ and $B = 0.5$.

10.3 Excess Entropy versus Particular Structure Factors

In Fig. 10.3 we have plotted **E** and $S(0)$ versus temperature. The system is ferromagnetic with $J = 1.0$ and $B = 0.5$. Thus, we have chosen to plot the structure factor for $q = 0$ since *a priori* we expect ferromagnetic behavior—i. e., configurations with period 1. The behavior of **E** in this case has been discussed above.

In the low temperature limit $S(0)$ vanishes. Since all the spins align with the magnetic field as $T \to 0$, $\langle s \rangle^2$ and $\langle s_0 s_r \rangle$ approach 1 for all r. Hence, $\Gamma(r) = \langle s_0 s_r \rangle - \langle s \rangle^2 = 0$ and so $S(0)$ vanishes.

In contrast, the high temperature behavior of $S(0)$ is a little surprising—based

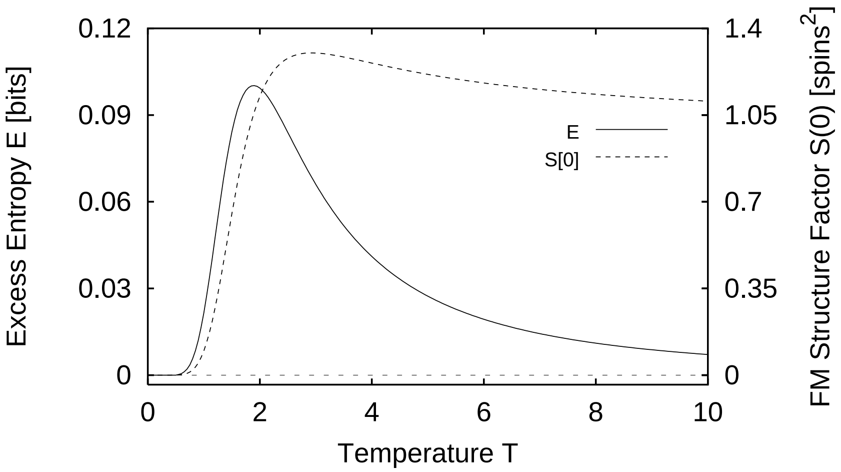


Figure 10.3: **E** and $S(0)$ versus T for a nn ferromagnet with $J = 1.0$ and $B = 0.5$.

on the above argument one would expect $S(0)$ to go to zero as T goes to infinity and as the correlations vanish. However, recall that $S(0)$ contains a "self-correlation" term, $\Gamma(0) = \langle s_0 s_0 \rangle - \langle s \rangle^2$. At high temperatures, the spins are randomly oriented so $\langle s \rangle^2 = 0$. However, $\langle s_0 s_0 \rangle = 1$ for all temperatures since $s_0 \in \{+1, -1\}$. Thus, $\Gamma(0) \to 1$ as $T \to \infty$, so $S(0) \to 1$ as $T \to \infty$.

In between these temperature extremes, there is a region where the correlation between spins is largest. Here, the system is neither random, as it is at high temperatures, nor is it trivially ordered, as it is at low temperatures. Not surprisingly, both $S(0)$ and **E** reach a single maximum in the intermediate regime.

However, note that **E** and $S(0)$ attain their maxima at different temperature values. The structure factor $S(0)$ is maximized at $T \approx 2.90$, and **E** is maximized at $T \approx 1.90$. As discussed in chapter 8, a given structure factor is designed to return a large signal if there are correlations present at that wavenumber. Its numerical value does not have a direct interpretation.

The shape of the curve in Fig. 10.3 is unchanged if either of the two modified structure factors defined in Eqs. (2.23) and (2.24) are substituted for $S(0)$. And none of these structure factors are maximized at the same temperature that maximizes $\mathbf{E}$.

10.4 Excess Entropy versus $\Gamma(1)$

As our last phenomenological comparison, Fig. 10.4 plots the nearest-neighbor correlation function $\Gamma(1)$ and the excess entropy $\mathbf{E}$ as a function of the temperature T. Like the structure factor, $\Gamma(1)$ carries units of [spins^2], not bits. As in the preceding examples, the two functions are maximized at different temperatures: $\Gamma(1)$ is maximized at $T \approx 2.50$.

That $\Gamma(1)$ and $\mathbf{E}$ reach a maximum at different temperatures is especially noteworthy since we are considering a system with only nearest-neighbor interactions. One might reasonably expect that for a system with such local, pairwise interactions, the nearest-neighbor two-spin correlation function would be sufficient to capture the system's global correlations. Fig. 10.4 shows that this is not the case. Even for a system with nn interactions, the nearest-neighbor correlation function does not measure apparent spatial memory as the excess entropy does.

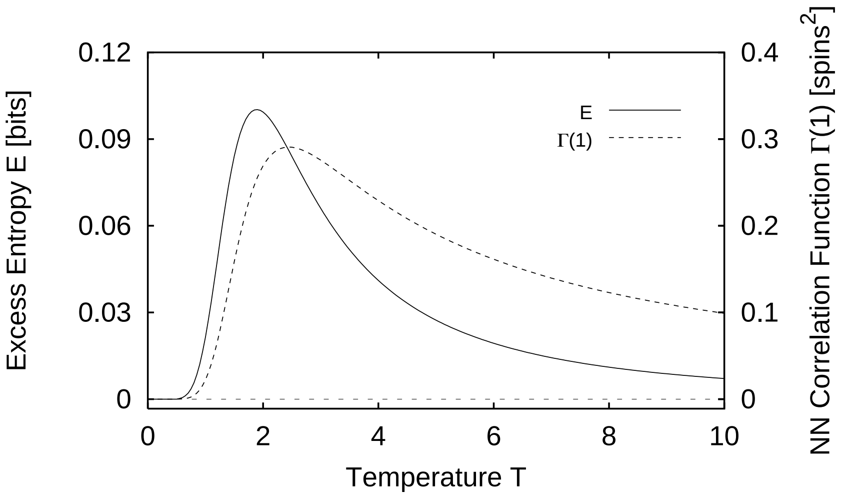


Figure 10.4: $\mathbf{E}$ and $\Gamma(1)$ versus T for a nn ferromagnet with $J = 1.0$ and $B = 0.5$.

10.5 Summary of Phenomenological Observations

Statistical mechanics possesses several functions that are similar to the excess entropy, but none can be interpreted as measures of spatial memory as $\mathbf{E}$ can be. We have seen that the correlation length, specific heat, and the structure factors exhibit behavior qualitatively similar to the excess entropy for the particular class of systems studied here. However, none of these statistical mechanical quantities returns a numerical value that quantifies memory. The excess entropy, being defined as a mutual information, carries units of bits, appropriate for this type of structural feature.

Moreover, we have seen in this section that each of these quantities reaches a maximum at different system parameters. This means that the statistical mechanical functions cannot be used to determine the parameter setting at which a given system's spatial memory is the largest. Simply put, to measure apparent spatial memory, one must use **E**.

Part III

Additional Results

Chapter 11

Measures of Statistical Complexity: Why?[1]

11.1 Statistical Complexity Measures

Theoretical physics has long possessed a general measure of the uncertainty associated with the behavior of a probabilistic process: the Shannon entropy of the underlying distribution [30, 142]—a quantity originally introduced by Boltzmann over 100 years ago. In the '50's, Kolmogorov and Sinai [85, 147] adapted Shannon's information theory to the study of dynamical systems. This work formed the foundation for the statistical characterization of deterministic sources of apparent randomness in the late '60's through the early '80's. These efforts to describe the unpredictability of

[1] **Co-written with J. P. Crutchfield and published as Ref. [55]**

dynamical systems were largely successful. The metric entropy, Lyapunov exponents, and fractal dimensions now provide widely applicable quantities that can be used to detect the presence of and to quantify the degree of deterministic chaotic behavior.

Since that time, though, it has become better appreciated that measuring the randomness and unpredictability of a system fails to adequately capture the correlational structure in its behavior. *Structure* here is taken to be a statement about the relationship between a system's components. Roughly speaking, the larger and more intricate the "correlations" between the system's constituents, the more structured the underlying distribution. Structure and correlation are not completely independent of randomness, however. It is generally agreed that both maximally random and perfectly ordered systems possess no structure [44, 64, 44]. Nevertheless, at a given level of randomness away from these extremes, there can be an enormously wide range of differently structured processes.

These realizations led to a considerable effort to develop a general measure that quantifies the degree of structure or pattern present in a process (c.f. [13, 33, 43, 44, 61, 64, 74, 96, 104, 144, 152, 158, 168]). There are many *ad hoc* methods for detecting structure, but none are as widely applicable as entropy is for indicating randomness. The quantities that have been proposed as general structural measures are often referred to as "complexity measures." To reduce confusion, it has become convenient to refer to them instead as *statistical* complexity measures. In so doing they are immediately distinguished from deterministic complexities, such as the Kolmogorov-

Chaitin complexity [25, 86, 94] which requires the deterministic accounting of every bit—random or not—in an object. In contrast, statistical complexity measures discount for randomness and so provide a measure of the regularities present in an object above and beyond pure randomness. Deterministic complexities are dominated by the random components in an object; the result is that their average-case growth rate is given by the Shannon entropy rate [30].

A number of approaches to measuring statistical complexity have been taken. One line of attack operates within information theory and examines how the Shannon entropy of successively larger subsystems converges to the entropy density of the entire system [43, 64, 96, 104, 144, 152]. These quantities can be interpreted as the average memory stored in configurations.

Another set of approaches appeals to computation theory's classification of devices that recognize different classes of formal language (a set of strings). Examples include finite memory devices (e.g. the finite-state machines) and infinite memory devices (e.g. push-down automata and Turing machines) [73]. One such computation-based measure of statistical complexity is the *logical depth* [13]. The logical depth of (say) a system's configuration is the time required for a universal Turing machine to run the minimal program that reproduces it. Another example of a computation theoretic approach is found in the statistical complexity of Refs. [33, 44], a quantity that measures the amount of memory needed, on average, to statistically reproduce a given configuration. Unlike logical depth, which assumes the use of a universal Turing machine—the most powerful discrete computational model class—this statistical complexity assumes that the simplest possible computational class is used to describe the configuration. A higher level class is used only if lower ones fail to admit a finite description. This “bottom-up” definition of hierarchical information processing has been successfully applied to the symbolic dynamics of chaotic maps [33, 44], cellular automata [40], spin systems [38], and hidden Markov models [154]. For other approaches to statistical complexity see, for example, Refs. [61, 74, 105, 158, 168].

11.2 Properties of C_{LMC}

Recently, Lòpez-Ruiz, Mancini, and Calbet proposed another measure of statistical complexity C_{LMC} [106]. Consider a discrete random variable Y that can take on N values y. We denote by $\Pr(y)$ the probability that the variable Y assumes the value y. Ref. [106] then defines a complexity measure:

$$C_{\mathrm{LMC}}[Y] \equiv H[Y]\,D[Y] \ , \tag{11.1}$$

where H is the Shannon entropy,

$$H[Y] = -\sum_{\{y\}} \Pr(y) \log_2 \Pr(y) \ , \tag{11.2}$$

in which the sum runs over all allowed values of y. The quantity D is the "disequilibrium," defined by

$$D[Y] = \sum_{\{y\}} (\Pr(y) - \frac{1}{N})^2 \,, \tag{11.3}$$

which measures the departure of $\Pr(y)$ from uniformity.

The motivation posited in Ref. [106] for the form of $C_{\rm LMC}$ is that it vanishes for distributions that correspond to perfect order and maximal randomness. Ref. [106] argues that perfect order corresponds to a vanishing Shannon entropy and notes that for $H = 0$, $C_{\rm LMC} = 0$. Maximal randomness occurs for $H = \log_2 N$, corresponding to $\Pr(y) = 1/N$. And so, by Eq. (11.3) D and hence $C_{\rm LMC}$ equal zero. Thus, by construction, $C_{\rm LMC}$ vanishes in the extreme ordered and disordered limits.

We now proceed to discuss the behavior of $C_{\rm LMC}$ in the thermodynamic limit. Anteneodo and Plastino have already reported some of $C_{\rm LMC}$'s properties in this limit [2]. However, their line of investigation is rather different from that undertaken here. In Ref. [2] the distribution that maximizes $C_{\rm LMC}$ is determined. Numerical and analytic work there indicate that the maximizing distribution for $N \to \infty$ is one in which a single event has probability 2/3 while all others are equally likely.

As an alternative to looking at the maximizing distribution, we suggest examining how a system's complexity changes as parameters—e.g temperature, coupling strength, nonlinearity, etc.—are varied. This approach is in keeping with statistical mechanics, where one typically looks at changes in the behavior of quantities in the thermodynamic limit as system parameters are varied. For example, rather than determine the distribution that maximizes the expectation value of the specific heat for a finite-sized system, one usually determines how the specific heat per site behaves in the limit of an infinite system as, say, the temperature is varied.

Consider the class of probability distributions over discrete, finite random variables generated by finite-memory Markov chains. Let $\ldots X_{-2}, X_{-1}, X_0, X_1, X_2, \ldots$ be a bi-infinite chain of random variables where each value x_i is chosen from a discrete finite alphabet of size k. We denote L consecutive variables by $X_i^L \equiv X_i, X_{i+1}, X_{i+2}, \ldots, X_{i+L-1}$. X_i^L is a system of L variables with $N = k^L$ possible configurations x_i^L. Let $\Pr(x_i)$ be the probability that the i^{th} random variable takes on the particular value x_i. We denote by $\Pr(x_i^L)$ the joint probability distribution over L consecutive random variables. We assume a shift symmetry so that $\Pr(x_i^L) = \Pr(x_0^L)$ and subsequently drop the subscript $i = 0$. The chain of discrete random variables may be viewed as a translationally invariant spin system or, equivalently, as a stationary stochastic process.

As is well known, the Shannon entropy of a block of L such variables typically grows linearly for sufficiently large L. In other words, the limit

$$h_\mu \equiv \lim_{L\to\infty} \frac{1}{L} H[X_0, X_1, \ldots, X_{L-1}] \qquad (11.4)$$

exists and, in the thermodynamic ($L \to \infty$) limit,

$$H[X^L] \propto h_\mu L \, . \tag{11.5}$$

The quantity h_μ is well-defined as the system size goes to infinity and is known as the entropy rate, the metric entropy, or the thermodynamic entropy density depending on the context. In statistical mechanics parlance, Eq. (11.4) tells us that the Shannon entropy H is an extensive quantity, so that it is possible to define a meaningful entropy density h_μ that characterizes the randomness per variable in the system.

The "disequilibrium" term, Eq. (11.3), is not so well behaved. In fact, under no circumstances does it grow linearly with system size L. One can show, quite generally, that for a system of length L described by a probability distribution over $N = k^L$ events, D is bounded above by $1 - 1/N$. To see this, we expand the square in Eq. (11.3) and, since the probability distribution is normalized, obtain

$$D[X^L] = \sum_{\{x^L\}} \Pr(x^L)^2 - \frac{1}{k^L} \, . \tag{11.6}$$

The sum is understood to run over all k^L possible values of X^L. Since $\Pr(x^L) \leq 1$, it follows that

$$\sum_{\{x^L\}} \Pr(x^L)^2 \leq \sum_{\{x^L\}} \Pr(x^L) = 1 \, . \tag{11.7}$$

Thus,

$$D[X^L] \leq 1 - k^{-L} , \tag{11.8}$$

and we see that D cannot grow linearly with the system size L. Therefore, the "disequilibrium" is not a thermodynamically extensive quantity.

In fact, it can be shown that D vanishes exponentially with increasing system size for a large class of systems. Let our chain of variables $\ldots X_{-1}, X_0, X_1, X_2, \ldots$ be chosen by a one-step Markov process with transition probabilities given by $T_{ab} = \Pr(b|a),\ a, b \in \{0, 1, \ldots, k-1\}$. That is, T_{ab} gives the probability that the variable X_{i+1} takes on the $b^{\underline{\text{th}}}$ value given that X_i takes on the $a^{\underline{\text{th}}}$ value. (What follows applies to any finite-step Markov chain, as blocks of adjacent variables can be grouped to render the process one-step.) Then, if we assume that the Markovian process is regular—*i.e.*, there exists some K such that $(T^K)_{ab} > 0$ for all a and b—then it follows that the disequilibrium of a system of L variables goes to zero exponentially fast in L. This is proved in appendix E.

As a result, C_{LMC} vanishes in the thermodynamic limit for all regular Markov chains, a class of systems that includes all finite-range, one-dimensional spin systems with finite-strength interactions. It seems to us counterintuitive that a (useful) measure of complexity vanishes for all of these systems. While these models exhibit no critical phenomena, there are considerable changes in the structure of the distributions as system parameters are varied [38]. A measure of complexity, as we envision it, should be sensitive to these changes.

As an illustration of the nonextensivity of D, consider the special case where the chain consists of variables that are independent and identically distributed (iid); *i.e.*, $\Pr(x_i, x_j) = \Pr(x_i)\Pr(x_j)$ for all X_i and X_j with $i, j = \ldots, -2, -1, 0, 1, 2, \ldots$. For a spin system, this corresponds to the case where there is no coupling between spins—a paramagnet. For convenience we assume that X_i can take on two values, (say) 0 and 1, and we denote the probability that X_i takes on the value 1 by p. Then, using the binomial theorem, we find that C_{LMC} for a system of L such variables is given by:

$$C_{\mathrm{LMC}}[X^L] = LH[X] \left((1 - 2p + 2p^2)^L - 2^{-L} \right) , \tag{11.9}$$

where $H[X]$ is the binary entropy function:

$$H[X] = -p \log_2 p - (1 - p) \log_2 (1 - p) . \tag{11.10}$$

Eq. (11.9) is rather curious. In our view, the complexity of a collection of iid binary variables should vanish regardless of their number. A set of independent variables is statistically very simple—there is manifestly no correlational structure whatsoever. Furthermore, it seems to us that if the complexity doesn't vanish for all such systems, it ought to grow linearly as a function of the number of variables. That is, six biased coins should be twice as complex as three biased coins. Eq. (11.9) shows that the size dependence of C_{LMC} is much more complicated.

In Ref. [2] it is found that the maximal value of C_{LMC} goes to $4/27$ as the system size goes to infinity. This is not at odds with the exponentially fast vanishing of D (and hence C_{LMC}) for regular Markov processes noted here, since the system that maximizes C_{LMC} is not Markovian. To see this, recall that the maximal distribution reported in Ref. [2] has one configuration with probability $2/3$ while all others have equal probability. In the thermodynamic limit, this one configuration is infinitely long. Thus, the generating process must keep track of arbitrarily long sequences in order to assign a distinct probability to one and another probability uniformly to all others. As a result, the distribution that maximizes C_{LMC} cannot be generated by a finite-memory Markov process in the thermodynamic limit.

11.3 Repairing Nonextensivity

Up to this point we have seen that C_{LMC} is not suitable for use in a statistical mechanics context. In particular, it suffers from two related deficiencies: it is not an extensive quantity and it vanishes for a large variety of structured processes. The trouble causing both of these shortcomings resides in the "disequilibrium" factor. Perhaps if one altered the definition of C_{LMC} so that D was extensive, as the Shannon entropy H is, one would obtain a more useful measure of statistical complexity.

To this end, we seek an extensive measure of a distribution's departure from uniformity. As we will be multiplying this measure by the Shannon entropy H

which carries units of bits, it also seems natural, although not necessary, to choose a "disequilibrium-like" quantity that also carries units of bits. Information theory is armed with just such a function: the relative information [30].

The relative information, also known as the information gain or the Kullback-Leibler information distance, between two distributions $\Pr(y)$ and $\widehat{\Pr}(y)$ is defined by

$$D(\Pr(y) \,||\, \widehat{\Pr}(y)) \equiv \sum_{\{y\}} \Pr(y) \log_2 \frac{\Pr(y)}{\widehat{\Pr}(y)} \; . \tag{11.11}$$

The relative information is not a true distance function—neither satisfying the triangle inequality nor being symmetric. Nevertheless, it does provide a measure of how much two distributions differ and it does carry the same units (bits) as the Shannon entropy. It is also an extensive quantity, since it grows linearly with the number of variables in the distribution's support.

So, $D(\Pr(y) \,||\, \widehat{\Pr}(y))$ where $\widehat{\Pr}(y) = 1/N$ provides an extensive measure of $\Pr(y)$'s departure from uniformity in units of bits. Using this in Eq. (11.3), we define a modified statistical complexity measure:

$$C'[Y] \equiv H[Y] \, D(\Pr(y) \,||\, 1/N) \; , \tag{11.12}$$

which has units of [bits^2]. From here on we will focus on C', the modified C_{LMC}. Note that much of C''s character is shared by C_{LMC}.

To see how this new quantity behaves, let's look more closely at $D(\Pr(y) \,||\, 1/N)$.

Consider again X^L, a Markov chain of length L. And for convenience let the X_i be binary variables. The total number N of configurations for such a system is 2^L. First, note that:

$$D(\,\Pr(x^L)\,\|\,1/N\,) \;=\; L - H[X^L]\;. \qquad (11.13)$$

Since $H[X^L]$ is extensive, we have:

$$D(\,\Pr(x^L)\,\|\,1/N\,) \;\propto\; L(1-h_\mu)\;. \qquad (11.14)$$

Thus, C' consists of the product of two extensive quantities, H and $D(\,\Pr(x^L)\,\|\,1/N\,)$. As a consequence, dividing C' by L^2 yields a quantity that is finite in the $L \to \infty$ limit; specifically,

$$\lim_{L\to\infty} \frac{1}{L^2} C' \;=\; h_\mu(1-h_\mu)\;. \qquad (11.15)$$

(See Fig. 11.1.) This is indeed a function that vanishes in the ordered ($h_\mu = 0$) and disordered ($h_\mu = 1$) extremes. However, note that it is a function *only* of the system's entropy density. That is, the modified statistical complexity measure Eq. (11.12) is a function only of the system's randomness. The relation $C' \propto L^2 h_\mu (1-h_\mu)$ strikes us as being "over-universal." For example, it's possible for a paramagnet and a system at its critical point, where the correlation length diverges, to have the same value of C'. It does not seem particularly revealing that all systems with the same entropy density have the same statistical complexity.

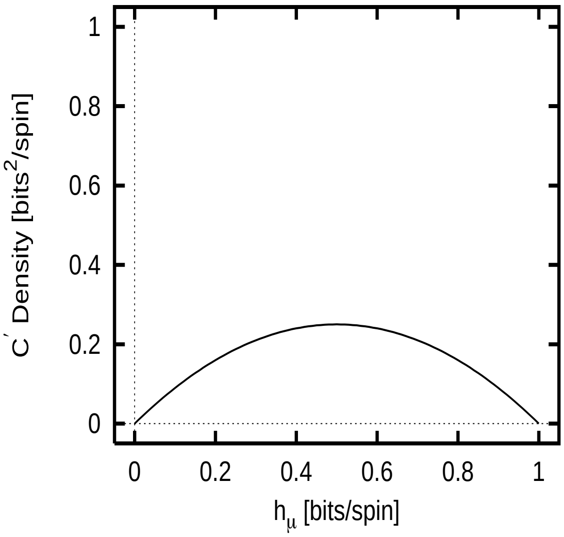


Figure 11.1: C' versus entropy density h_μ. Note that C' is a function of h_μ.

As a contrast to the C' versus h_μ behavior shown in Fig. (11.1), consider Fig. (11.2), where we show the behavior of a different measure of statistical complexity, the *excess entropy* E [43, 64, 96, 104, 144, 152]. The excess entropy of an infinite configuration may be expressed as the mutual information between two semi-infinite halves of the configuration. That is, E is the amount of spatial memory embedded in configurations. In Ref. [38] we show that E captures significant structural changes in the configurations of one-dimensional spin systems as external parameters are varied. We have plotted the excess entropy for 10^4 sets of parameter values for a 1D Ising system with nearest-neighbor coupling in the presence of an external field. Note that for *all* the Ising systems plotted in Fig. (11.2), $C_{\mathrm{LMC}} = 0$, since C_{LMC} vanishes in the thermodynamic limit. Comparing the two figures, it is clear that the excess entropy depends on the entropy density h_μ in a much more subtle way than C' does.

Comparing these two plots raises another important issue. Note that the excess entropy does not always equal zero for $h_\mu = 0$, an apparent violation of the "boundary condition" requiring that a complexity measure vanish in the perfectly ordered limit. However, $h_\mu = 0$ corresponds to perfect *asymptotic predictability*, not perfect *order.* A process with a vanishing entropy rate indicates that it can be predicted without error—it says nothing, however, about how much effort or memory is required to perform this prediction. Thus, a zero value of the entropy density is too crude a measure of order. To see this, note that *any* periodic system has $h_\mu = 0$. Yet all periodic systems aren't equally ordered: a configuration with period 1 is certainly more ordered than a configuration of period 1729—which, for example, requires more memory to produce. In fact, at the period-doubling accumulation point of the logistic map, the symbolic dynamics produce periodic configurations of diverging periodicity. Hence, the excess entropy is infinite here, while the entropy rate remains zero [33, 44, 64].

In contrast to C', which vanishes for any periodic system, the excess entropy E for a configuration of period P is $\log_2 P$. Only if the period is 1, indicating trivial ordering and predictability, does the excess entropy vanish for a periodic process. Thus, the $(h_\mu, \mathrm{E}) = (0, 0)$ points in Fig. (11.2) correspond to the system's ferromagnetic ground states of period 1 and the $(h_\mu, \mathrm{E}) = (0, 1)$ points correspond to the antiferromagnetic ground states of period 2 [38]. Statistical complexity measures such

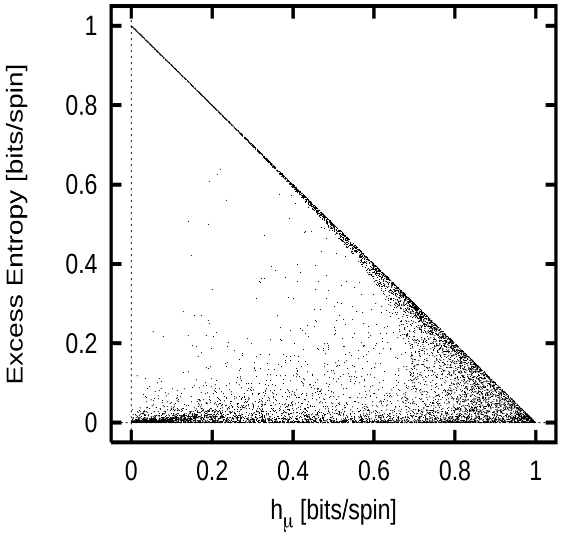


Figure 11.2: Excess entropy E, a statistical complexity measure, versus entropy density h_μ for a spin-1/2 one-dimensional Ising spin system: 10^4 (h_μ, E) points. The system parameters were randomly chosen from the following intervals: J (coupling constant) $\in [-3, 3]$; T (temperature) $\in [0.05, 4.05]$; and B (external field) $\in [0, 3]$.

as C' or, for that matter C_{LMC}, that are zero for all $h_\mu = 0$ configurations are very blunt implements with which to detect structure. A measure of complexity should be able to distinguish between structures of different periodicities.

For maximal randomness ($h_\mu = 1$), the excess entropy E vanishes, as expected. At $h_\mu = 1$, corresponding to infinite temperature, the spins decouple and there is no information shared between them. But the excess entropy does more than satisfy the "boundary conditions" of vanishing for $h_\mu = 0$ and 1. The interpretation of **E** as the memory stored in spatial configurations holds for intermediate values of h_μ as well. As a result, Fig. (11.2) lets us place an upper bound on the memory stored in spatial configurations for a spin-1/2 nearest neighbor Ising model: $\mathrm{E} \leq 1 - h_\mu$. This result, derived analytically in Ref. [38], applies to all one-step Markov chains over a binary alphabet.

We conclude this section by noting that there is a growing body of evidence indicating that, aside from the requirement of vanishing at the ordered and disordered extremes, entropy and "complexity" (defined in a number of different ways to reflect "structure") are more or less independent. That is, there is a vastly wider range of complexity versus entropy relationships than indicated by Eq. (11.15) [96, 33, 38]. Fig. (11.2) is just one example of many possible statistical complexity-entropy density relationships.

11.4 Conclusion

To summarize, we have shown that C_{LMC} vanishes in the thermodynamic limit for finite-memory regular Markov chains. This class of systems includes, at a minimum, all finite-range one-dimensional spin systems. We have also shown that C_{LMC} is not an extensive variable. We have proposed modifying C_{LMC}, replacing the "disequilibrium" of Ref. [106] with the relative entropy with respect to the uniform distribution. This results in an quantity C' that grows appropriately in the thermodynamic limit, making it possible to define a meaningful statistical complexity density that, nonetheless, retains the spirit of C_{LMC}. However, the product of this modification is a quantity that is a trivial, "over-universal" function of the entropy density h_μ. In short, based on the above observations, it seems to us that $C_{\rm LMC}$ and C' may be of little use in measuring the complexity of a statistical mechanical system.

We conclude by pointing out that the "boundary conditions" of vanishing in the extreme ordered and disordered limits do not uniquely specify a measure of complexity, an observation also made by Anteneodo and Plastino [2]. In fact, if this is the only feature one demands of a complexity measure, it's not clear to us why one would be motivated to devise a new statistic at all.

Statistical mechanics, for example, is replete with functions that vanish in the high and low temperature limits. Since thermodynamic entropy, a measure of randomness, is a monotonic function of temperature, high (low) temperature corresponds to high (low) randomness. Examples of quantities that vanish in these extremes (assuming there is not a critical point at $T = 0$) include the connected correlation functions, the correlation length, and magnetic susceptibility. These functions can be easily applied to any probability distribution describing a spatially or spatio-temporally extended collection of random variables.

Information theory also comes equipped with a function that vanishes for perfectly ordered and disordered systems: the mutual information I [142, 30]. If two random variables Z and Y are independently distributed, then the mutual information between them, $I[Z;Y]$ vanishes. At the other extreme, if Z and Y are both known with certainty—that is, $H[Z] = H[Y] = 0$—then $I[Z;Y]$ also vanishes. For statistical dependencies between these extremes, $I[Z;Y]$ is positive and measures the amount of information shared between Y and Z.

Given that there are many functions that vanish in the extreme ordered and disordered limits, it is clear that requiring this property does not sufficiently restrict a statistical complexity measure. What other criteria can we use, then, to guide us as we attempt to detect structures and patterns in nature? To this question we offer two suggestions.

First, it is helpful if the statistical complexity measure has a clear interpretation: *What exactly is the statistical complexity measuring?* The two English words "statistical complexity" do not sufficiently answer the question. Many of the statistical complexity measures proposed over the last decade or so do have clear interpretations. For example, the excess entropy may be interpreted as the mutual information between two halves of an infinite configuration [64, 96, 38]. Logical depth is the run time required by a universal Turing machine executing the minimal program to reproduce a given pattern. These unambiguous interpretations help put these statistical complexity measures on a solid footing.

Second, it is essential to consider the motivations behind a measure of statistical complexity: *How is the measure to be used? What questions might it help answer?* It is possible to meaningfully assess its utility only if the motivations and goals for defining a complexity measure are stated clearly. One set of issues is the detection and quantification of patterns produced by a process. It has been proposed that particular notions of structure adapted from computation theory capture the intrinsic "patterns" and information processing architecture embedded in a system. In this setting, one finds well-defined and easily interpreted measures of statistical complexity [33]. For other views on questions that a measure of statistical complexity might help answer, see Ref. [14].

Finally, Ref. [2] mentions several different notions of complexity and notes that "there is not yet a consensus on a precise definition." "Complexity" has accepted meanings in other fields—meanings established prior to the recent attempts to use it as a label for structure in natural systems. For example, *Kolmogorov-Chaitin complexity* in algorithmic information theory means something quite different from *computational complexity* in the analysis of algorithms. These, in turn, are each different from the *stochastic complexity* used in model-order estimation in statistics [135]. Though at a future date relationships may be found, at present all of these are different from the notions of statistical complexity discussed here.

Unfortunately, "complexity" has been used without qualification by so many authors, both scientific and non-scientific, that the term has been almost stripped of its meaning. Given this state of affairs, it is even more imperative to state clearly why one is defining a measure of complexity and what it is intended to capture.

Part IV

Preliminary Work in Two Dimensions

Chapter 12

Introduction: Memory and Computation in Two Dimensions and the Statistical Mechanics of Critical Phenomena

12.1 Pattern, Organization, and Computation in Two Dimensions

The central questions addressed in this dissertation concern pattern, organization, and intrinsic computation. What are patterns, and how do we discover them? What does it mean to say a system is organized? Is it possible to distinguish between different levels of organization? And, how can one infer a system's intrinsic computation—how does it store, transmit, and manipulate historical information to produce future behavior?

In Parts I–III we presented an approach to answering these questions in the context of one-dimensional spin systems. In particular, we showed that the excess entropy serves as an information theoretic measure of a system's apparent spatial memory and that ϵ-machines provide a minimal representation of all the regularities—approximate and exact—present in configurations. Although situated in a discussion of one-dimensional spin systems, the excess entropy and the ϵ-machine are defined for all one-dimensional stochastic processes with a stationary measure.

We feel that the approach developed thus far—a complementary application of information theory and computational mechanics—provides a concrete answer to the questions of memory, pattern, and organization posed above for one-dimensional systems. There are certainly important outstanding issues: most notably the question of how one can optimally infer the structures present in infinite memory processes. Nevertheless, despite the fact that there is work yet to be done to establish an optimal set of implementations and calculational methods for ϵ-machine reconstruction, we feel that the questions posed at the beginning of this chapter applied to one-dimensional systems have been satisfactorily answered.

In the following three chapters we turn our attention to extending the formalism discussed in parts I and II to apply to two-dimensional configurations. We shall see that the notion of pattern and organization is considerably more elusive in two dimensions than it is in one. Indeed, how to define pattern, organization, or memory in a multidimensional system is, by and large, an open question.

Here, in the final part of this dissertation before concluding, we present some preliminary work in this area and propose some possible directions for future research. The present chapter contains a range of introductory material. In Sec. 12.2 we present a very brief survey of some previous work that aims to address questions of structure, organization, and memory in more than one dimension. Then, in Sec. 12.3 we introduce the two-dimensional Ising model and state the exact results for the magnetization, nearest-neighbor two-spin correlation function, and the specific heat. A discussion of the behavior of the two-dimensional Ising model leads us to Secs. 12.4 and 12.5, where we discuss, in the context of the two-dimensional Ising model, the theory of critical phenomena, including a brief mention of the important notion of universality. Finally, in Sec. 12.6 we reexamine questions of memory, structure, and computation in the context of the two-dimensional Ising model.

In chapter 13 we give some exact results for the mutual information between a pair of nearest neighbor spins for the two-dimensional Ising model. This quantity, being a function of the distribution of only two spins, certainly fails to give a complete accounting of the memory stored across the entire, two-dimensional lattice. Nevertheless, the two-spin nearest-neighbor mutual information does provide a crude measure of the probabilistic structure of the system.

In chapter 14 we propose several ways to extend to two dimensions the computational mechanical approach to pattern and memory developed thus far. In Sec. 14.2 we review one possible generalization of ϵ-machines to two dimensions [69, 71]. We conclude chapter 14 and part IV by discussing directions for future research.

12.2 Brief Survey of Previous Work

In this section we provide a brief survey of work that has been put forth to address issues of pattern, memory, and computation in more than one dimension. Our purpose here is not to provide a complete, in-depth review. Rather, the goal of this section is to give the reader a sense of the ideas in this area and to provide references for those wishing to pursue in greater depth.

The question of how to describe spatial patterns was considered in the early and mid 80's by Wolfram in the context of cellular automata (CA) [166]. In particular, he examined the patterns produced by various one-dimensional CA as they evolved in time. The resulting "space-time" diagram, containing one spatial and one temporal dimension is a two-dimensional object.

In an influential paper [168], Wolfram proposed a qualitative classification scheme in which CAs are characterized by their asymptotic behavior. Specifically, Wolfram grouped CAs into four classes: those with homogeneous, periodic, "chaotic", and "complex" asymptotic behaviors. The "complex" group, known as class IV CA's, are those that exhibit "complex localized structures, sometimes long-lived" [168, p. 5].

Wolfram's distinctions between different classes of behaviors are largely qualitative and subjective. He does, however discuss the entropy density for one-dimensional strings extended either spatially or temporally. In Ref. [167] Wolfram explored making his classification scheme more quantitative by using a computation theoretic analysis of the sets of configurations produced by a CA, when the CA operates on *all* initial conditions. This work, while breaking significant new ground at the time, nevertheless is largely one-dimensional. A collection of Wolfram's papers from 1983–1985, as well as those of several other authors, has been reprinted in Ref. [169].

Motivated in part by Wolfram's work, Grassberger [64] and Lindgren and Nordahl [99, 100, 101, 104, 117] explored information and computation theoretic approaches to structure and complexity, mainly as applied to the configurations produced by one-dimensional CA. [99, 100, 101, 104, 117]. In Ref. [64], Grassberger briefly discusses some of the difficulties associated with extending a measure of memory to more than one dimensions. Later, in Refs. [100, 101] Lindgren put forth a method through which the information theoretic approach to structure discussed in Chapter 4 can be extended to more than one dimension. These ideas will be examined more closely in Chapter 14.

The excess entropy has been estimated for one-dimensional strings drawn from Monte Carlo-generated configurations of the two-dimensional Ising model [3]. (There the excess entropy is called the "complexity.") The analysis of Ref. [3] is, at root, one dimensional: the quantities calculated do not take into account the full, two-dimensional structure possessed by the configurations. Furthermore, we find the work presented in Ref. [3] to be less than thorough. For example, there is no attempt to investigate the nature of the cusp (as a function of temperature) in the excess entropy that is observed to occur at the critical temperature, and no mention is made of the crucial issue of disorder averaging when Ref. [3] reports results for a spin-glass version of the model.

Lempel and Ziv [91] and Sheinwald, Lempel, and Ziv [145] have discussed the compression of two-dimensional configurations. To do so, they parse the two-dimensional configuration by considering the variables seen as one moves through the lattice following a Peano-Hilbert space-filling curve. (For a discussion of the Peano-Hilbert curve, see, e.g., Ref. [127].) These authors then construct an algorithm for encoding the two-dimensional that is asympotically optimal in the sense that the size (in bits) of the code can be made arbitrarily close to the entropy density of the configuration. This result is analogous to Shannon's coding theorems for one-dimensional information sources, as discussed in Secs. 3.2.6 and 4.5

Despite the important results of Lempel, Ziv, and Sheinwald, at present there is not a fully-developed multidimensional information theory. In particular, little is known about how to quantify the memory shared across a multidimensional lattice. However, a considerable amount of recent effort has gone into the study of constrained two-dimensional codes; see, e.g., Refs. [77, 78, 120, 146, 163]. Constrained codes are those in which all configurations are not allowed. For example, Ref. [120] considers $M \times N$ binary arrays with the restriction that there must be the same number of 0's in each row and the same number of 1's in each column.

Constrained codes are of considerable practical importance. On magnetic coding materials, such as a computer's hard disk, all patterns of "up" and "down" magnetic domains can not be used, due to limits in the materials' properties and the ability of the read device to distinguish between certain patterns. For example, the mechanism through which the codes is read may not be able to distinguish between 4 and 5 "up" domains in a row. The results on constrained codes in the references cited above mainly involve (non-trivial) combinatorics; in all of the references a uniform measure is assumed over all allowed configurations.

Turning now from information theory to computation theory, we note that the computation theory employed in Chapter 5 is fundamental one dimensional. The main items of consideration in discrete computation theory are one-dimensional symbolic strings [21, 73]. For recent results on the "state of the art" in two-dimensional formal language theory, see Lindgren, Moore, and Nordahl [103].

Finally, we note that there has been some work done on shift spaces in more than one dimension. A brief discussion, complete with many references, may be found in Section 13.10 of Lind and Marcus [98].

12.3 Review of the two-dimensional Ising Model

In this section we introduce the two-dimensional Ising model. This model provides a concrete example for our discussion of pattern, organization, and computation in two dimensions. We also introduce the model since a discussion of its properties provides a natural setting for our review in Secs. 12.4 and 12.5 of the statistical mechanics of phase transitions.

The discussion of the Ising model and the exact results given below are standard; they can be found in many graduate-level statistical mechanics texts. See, for example, Refs. [130] and [136]. For a much more thorough treatment of the two-dimensional Ising model see Baxter [11] or the Ising model "bible" of McCoy and Wu [110].

The Hamiltonian for the two-dimensional Ising model is given by:

$$\mathcal{H} = -J \sum_{\langle ij,kl \rangle} s_{i,j} s_{k,l} \ , \tag{12.1}$$

where the variables $s_{i,j} \in +1, -1$ reside on a two-dimensional lattice, rather than a one-dimensional chain. We label the lattice sites by $i, j \in (0, 1, \cdots N-1)$. The angular brackets in Eq. (12.1) indicate that the sum is performed only over nearest neighbors. For example $s_{1,1}$ and $s_{2,1}$ are nearest neighbors, but $s_{1,1}$ and $s_{2,2}$ are not. Without loss of generality, we set $J = 1$ and use units such that $k_B = 1$.

Note that the Hamiltonian is rotationally and translationally invariant. That is, the Hamiltonian does not pick out a preferred direction or location. Also note that the Hamiltonian possesses a spin-flip symmetry; the value of the Hamiltonian is unchanged if all spins change sign.

While statistical mechanical functions can be calculated relatively easily for the one-dimensional model, this isn't the case for the two-dimensional model. The mathematics involved in an exact treatment of the two-dimensional model is rather formidable. The partition function can no longer be expressed as the dominant eigenvalue of a finite-dimensional transfer matrix; instead, an infinite-dimensional matrix, i.e., an operator, must be used [140]. Another technique for approaching the two-dimensional Ising model entails expressing the partition function as the Pfaffian of an antisymmetric matrix and determining its behavior in the thermodynamic limit where the size of the matrix diverges [110, 136]. (The Pfaffian is, like the determinant, a scalar function of a matrix.)

We shall not discuss either of these techniques here. Rather, we state without derivation some essential results. We begin with the magnetization. Recall that the magnetization per site m was defined in Eq. (2.11) as the expectation value of a single spin. For the two-dimensional Ising model with the Hamiltonian of Eq. (12.1), the

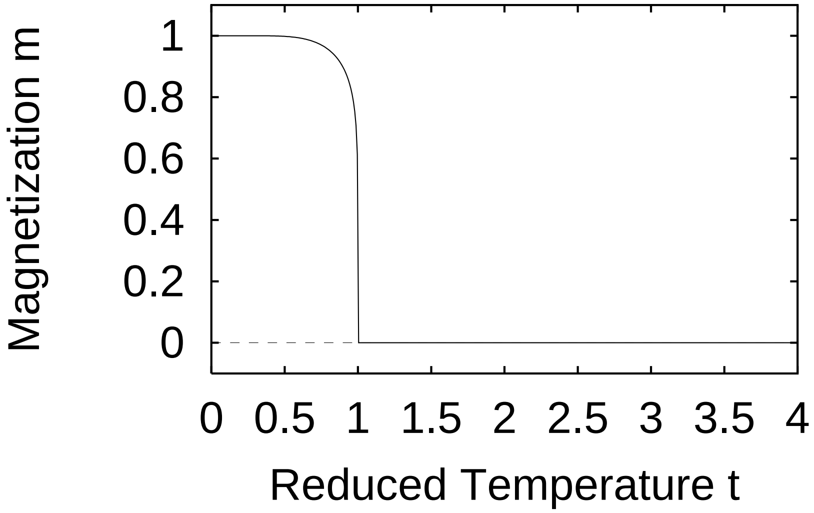


Figure 12.1: The magnetization m as a function of $t \equiv T/T_c$ for the two-dimensional, zero-field Ising Model. The critical temperature is $t = 1$.

magnetization per site is given by:

$$m = \left\{ \begin{array}{cc} \left(1 - \frac{\left[1-\tanh^2(1/T)\right]^4}{16\tanh^4(1/T)}\right)^{1/8} & T < T_c \\ \\ 0 & T > T_c \end{array} \right\} . \tag{12.2}$$

The critical temperature T_c is found numerically;

$$T_c \approx 2.269 \; . \tag{12.3}$$

The magnetization is plotted as as a function of the reduced temperature $t \equiv T/T_c$

in Fig. 12.1. Note that m is zero above T_c and non-zero below T_c.

The non-zero magnetization below T_c indicates that the spin-flip symmetry possessed by the Hamiltonian no longer holds. Below the critical temperature the spins tend to align (say) up; a situation that is clearly changed by a global spin flip. This phenomena is generally referred to as spontaneous symmetry breaking or ergodicity breaking. The adjective "spontaneous" is used to indicate that it is the system itself which is responsible for the broken symmetry, not some external influence. For the two-dimensional Ising model it is the interaction between spins that is responsible for the symmetry breaking, not the application of an external field.

The magnetization is an example of an *order parameter*, a quantity that is zero above the critical point and non-zero below. Note that this definition of the order parameter does not give any indication how to determine an appropriate order parameter for a given phenomenon. The selection of an order parameter is sometimes not obvious. This is the case, for example, in spin glass models where a useful order parameter is the *overlap*, a quantity that measures the degree to which a system fails to be self-averaging. For details, see, e.g., Ref. [57].

Another quantity that will be of interest in the following is the expectation value of the product of two spins:

$$\Gamma_u(m,n) \equiv \langle s_{i,j}\, s_{i+m,j+n} \rangle \ , \tag{12.4}$$

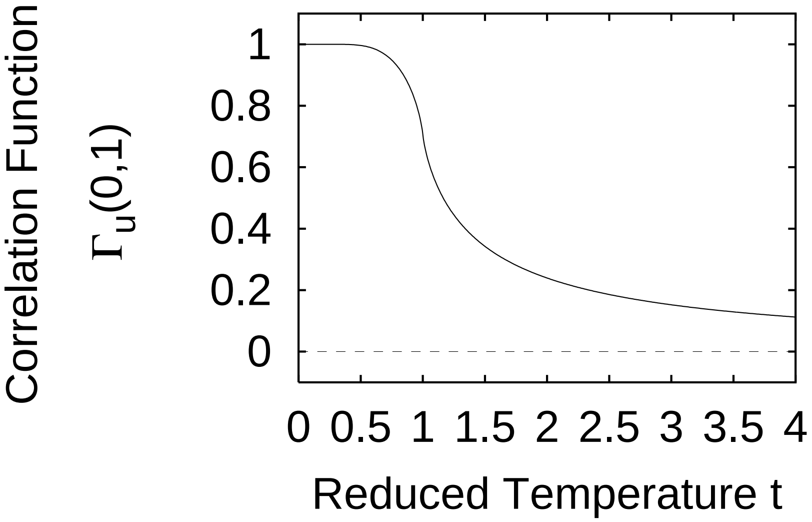


Figure 12.2: The nearest-neighbor two-spin correlation function $\Gamma_u(1,0) \equiv \langle s_{i,j} s_{i+1,j} \rangle$ as a function of $t \equiv T/T_c$ for the two-dimensional, zero-field Ising Model. The critical temperature is $t = 1$.

sometimes called the nearest-neighbor correlation function. Note that this correlation function is defined differently than the connected correlation function of Eq. (2.14). The quantity defined in Eq. (12.4) is distinguished from the connected correlation function through the use of the subscript "u". (The "u" stands for unconnected.) By rotational symmetry, it follows that:

$$\Gamma_u(m,n) \,=\, \Gamma_u(n,m) \,=\, \Gamma_u(|n|,|m|)\ . \tag{12.5}$$

Looking at the Hamiltonian, Eq. (12.1), observe that the nearest-neighbor two-spin correlation function is proportional to U, the expectation value of the Hamiltonian. That is:

$$\Gamma_u(1,0) \;=\; -U \;\equiv\; \frac{-1}{N^2}\langle\mathcal{H}\rangle \;, \tag{12.6}$$

for an $N \times N$ lattice.

The nearest-neighbor correlation function $\Gamma_u(1,0)$ is given by [136]:

$$\langle S_{i,j}\, S_{i+1,j}\rangle \;\equiv\; \Gamma_u(1,0) \;=\; \tanh(2/T) \;+\; \frac{1-\sinh^2(2/T)}{\sinh(2/T)\cosh(2/T)}\left[\frac{1}{2}-\frac{1}{\pi}K(\gamma)\right] \;, \tag{12.7}$$

where $K(\gamma)$ is the complete elliptic integral of the first kind:

$$K(\gamma) \;\equiv\; \int_0^{\pi/2} \frac{d\phi}{\sqrt{1-\gamma^2\sin^2\phi}} \;, \tag{12.8}$$

and,

$$\gamma \;\equiv\; \frac{2\sinh(2/T)}{\cosh^2(2/T)} \;. \tag{12.9}$$

The correlation function is plotted in Fig. 12.2 versus reduced temperature. Note the point of inflection at the critical reduced temperature $t = 1$.

The specific heat per site C, defined in Eq. (2.25), as $\partial U/\partial T$. Thus, by Eq. (12.6), the specific heat is equal to the negative slope of $\Gamma_u(1,0)$. We see in Fig. 12.2 that the slope of $\Gamma_u(1,0)$ diverges at the critical temperature $t = 1$. As a result, C is infinite at $t = 1$. The specific heat is plotted as a function of t in Fig. 12.3.

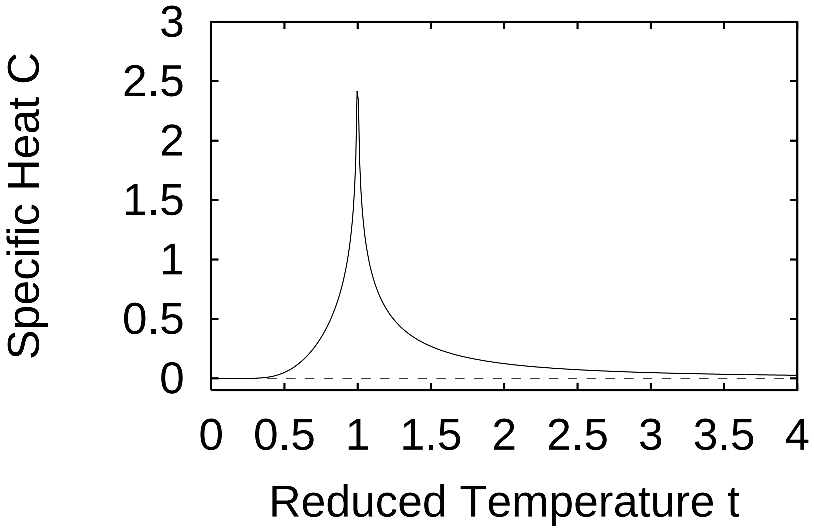


Figure 12.3: The specific heat C as a function of $t \equiv T/T_c$ for the two-dimensional, zero-field Ising Model. The critical temperature is $t = 1$.

Analytically, C is given by:

$$
\begin{aligned}
C \;=\; & 2\left(\frac{1}{T\sinh(2/T)}\right)^2 \times \\
& \left[\frac{2K(\gamma)}{\pi}\left(\frac{2+\sinh^2(2/T)+\sinh^4(2/T)}{\cosh^2(2/T)}\right) - \frac{2E(\gamma)}{\pi}\cosh^2(2/T) - 1\right] \;, \qquad (12.10)
\end{aligned}
$$

where $E(\gamma)$ is the complete elliptic integral of the second kind;

$$
E(\gamma) \;\equiv\; \int_0^{\pi/2} d\varphi \, \sqrt{1-\gamma^2\sin^2\varphi} \;. \qquad (12.11)
$$

12.4 Phase Transitions and the Critical Behavior of the 2D Ising Model

In this section and the next, we briefly discuss the modern theory of phase transitions. This field is quite rich and well developed; the review given here is, of necessity, rather brief. There is a range of sources the reader may consult to obtain a more thorough account. Ref. [164] by Wilson contains a very clear, non-technical discussion of critical phenomena. Yeomans has written an excellent, easy-to-read short text, "Statistical Mechanics of Phase Transitions" [170] that introduces all of the key ideas and techniques typically encountered in the study of critical phenomena. The text by Binney *et al*. [17] is recommended for those seeking a more mathematically advanced treatment.

We saw in Eq. (12.2) and Fig. 12.1 that the magnetization m suddenly assumes a non-zero value at the critical temperature T_c; the slope of m as a function of T is infinite. This behavior is an example of a *phase transition*, defined qualitatively as a sharp change in a thermodynamic function. The change of state between a liquid and a solid is a familiar example of a phase transition. At the critical temperature—zero degrees Celsius for water under normal atmospheric conditions—the properties of water change discontinuously.

This qualitative definition can be restated mathematically: a phase transition occurs at parameter settings where the partition function, Z_N of Eq. (2.7), is non-analytic in the thermodynamic limit. Recall that in Chapter 2 we saw that thermodynamic variables such as m, χ, and C, can be written as derivatives of the partition function. Thus, a non-analyticity in Z_N manifests itself as a discontinuity or a divergence in one or more thermodynamic functions.

Phase transitions are grouped into two classes. A phase transition is said to be *first-order* if the first derivative of the free energy per site $f = -\frac{T}{N} \log Z_N$ is discontinuous. First-order phase transitions that occur as the temperature T is varied are accompanied by discontinuity in the entropy. To see this, recall that the thermodynamic entropy density is given by:

$$\mathbf{S} = -\frac{\partial f}{\partial T} \ . \tag{12.12}$$

So, if the phase transition is first order, then it follows that $\mathbf{S}$ is discontinuous at the critical temperature. Thermodynamics then tells us that additional heat of, $\Delta Q = T \Delta S$, must be added to the system in order to cause the system to move from the low temperature to the high temperature state. This additional heat is known as the *latent heat*. For example, the latent heat associated with the solid-liquid phase change of water is 3.33×10^5 Joules per kilogram.

Of interest to us here are *continuous* phase transitions; which occur at parameter settings where the first derivative of f is continuous but higher derivatives of f are discontinuous or infinite. Continuous phase transitions are also known as *critical* phase transitions; a system that exhibits a continuous phase transition is an example of a *critical phenomenon*.

The behavior of the magnetization m in Fig. 12.1 is an example of a continuous phase transition. The magnetization may be written as a first derivative of f; $m = -\partial f/\partial B$. Note that $\partial m/\partial t$ diverges at the critical temperature $t = 1$. Hence, the second derivative of f diverges and this phase transition is classified as as continuous.

The magnetization m is not the only thermodynamic function that is non-analytic at the critical temperature T_c. For example, we saw in Eq. (12.10) and Fig. 12.3 that the specific heat diverges at T_c. The magnetic susceptibility χ also diverges at T_c. (No exact, closed form expression is known for χ for the two-dimensional Ising model.)

At T_c the two-spin connected correlation function $\Gamma(r)$ decays algebraically as a function of r, where $r \equiv \sqrt{m^2 + n^2}$ is the distance between the two spins. As a result, the correlation length ξ diverges. This divergence is not unexpected, given that the susceptibility χ also diverges. Recall that χ may be written as a sum over two-spin connected correlation functions as in Eq. (2.21). If this sum is divergent, then it follows that $\Gamma(r)$ must decay slower than exponentially and so ξ must diverge. We shall return to discuss the implications of a diverging correlation length below.

12.5 Critical Exponents and Universality

Summarizing the above comments, we saw that at T_c the linear response functions C and χ diverge; the magnetization m is non-analytic; the two-spin correlation function decays algebraically; and the correlation length ξ diverges. In this section we examine the nature of these divergences. Our motivation for doing so will become clear in the following section on universality.

We ask: What is the asymptotic form of these thermodynamic functions near T_c? To answer this question is given by the *critical exponents.* First, define:

$$\mathfrak{t} \equiv \frac{T - T_c}{T_c} \; . \tag{12.13}$$

Note that when $T = T_c$, $\mathfrak{t} = 0$. Then, we say that [170]:

$$F(\mathfrak{t}) \sim |\mathfrak{t}|^{\alpha} \; , \tag{12.14}$$

if

$$\lim_{\mathfrak{t} \to \infty} \frac{\log |F(\mathfrak{t})|}{\log \mathfrak{t}} = \alpha \; . \tag{12.15}$$

The constant α is known as the critical exponent.

For example, near T_c, one finds from Eq. (12.10), after some non-trivial analysis, that:

$$C \sim \log|\mathfrak{t}| \,, \tag{12.16}$$

and thus the specific heat critical exponent is zero. Other critical exponents and their definitions are given in Table 12.1. The critical isotherm exponent δ will not be discussed here, but is given in the table for completeness. For a discussion of its definition see Refs. [17, 170].

Why focus on critical exponents? The answer lies in their *universality*: the critical exponents defined in Table 12.1 are the same for many different systems. Remarkably, it turns out that the values of the critical exponents depend only on the symmetry of the order parameter and the dimensionality of the system. For example, the order parameter for the two-dimensional Ising model—in fact, for all Ising models—is the magnetization, a scalar. Hence, all two-dimensional systems with a scalar order parameter that undergo a continuous phase transition have the same critical exponents as the two-dimensional Ising model. All of these systems are said to belong to the two-dimensional Ising *universality class.*

For example, a two-dimensional Ising model with a coupling constant $J = J' > 0$ will have a different critical temperature than the system with $J = 1$. Nevertheless, the critical exponents for the systems will be the same. As another example, suppose we add to the Hamiltonian of Eq. (12.1) an interaction between next-nearest vertical neighbors so that the new Hamiltonian is:

$$\widetilde{\mathcal{H}} \equiv -J \sum_{\langle ij,kl \rangle} s_{i,j} s_{k,l} - J'' \sum_{ij} s_{i,j} s_{i+2,j} \ . \tag{12.17}$$

Even this system will have the same critical exponents as those given in Table 12.1.

Qualitatively, universality can be explained as follows. At the critical point of a continuous phase transition symmetry is spontaneously broken as a result of the correlations between degrees of freedom. Since the symmetry-breaking is a macroscopic effect, the correlations must be significant enough to exert an influence on a macroscopic scale. Thus, the two-spin correlation function must decay more slowly than exponentially and the correlation length must diverge. The central idea is that at the critical point all length scales are important; correlations and fluctuations at all scales contribute to the system's behavior. Given this, it is not surprising, that the details of the microscopic interactions do not effect the critical behavior of the system as measured by the critical exponents.

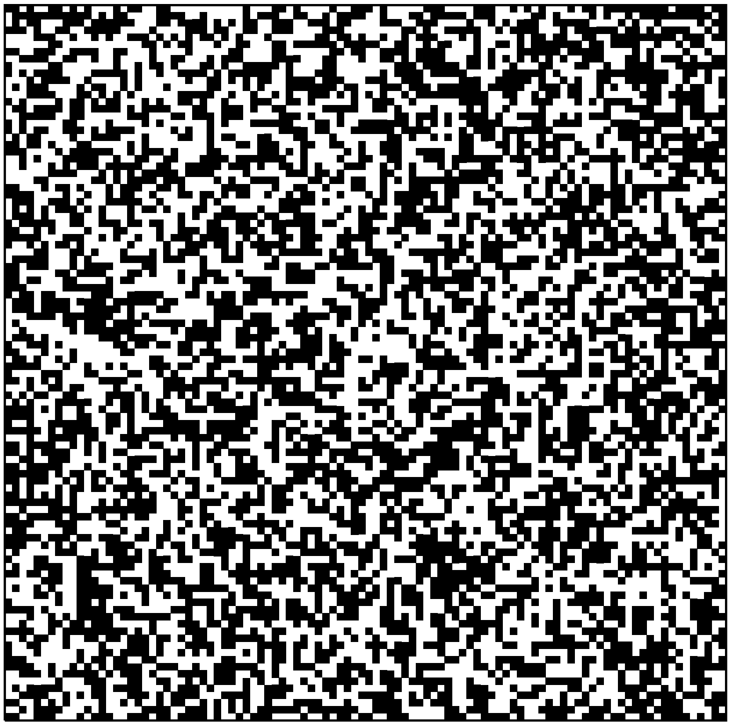
A rigorous, theoretical justification for the phenomenon of universality is found by examining how the configurations change under a *renormalization group* operation that removes or "traces out" degrees of freedom. A discussion of the renormalization group is beyond the scope of this review. The interested reader is urged to consult Refs. [17, 164, 170] for a more thorough explication.

In summary, the notion of universality is one of the cornerstones of modern

Critical Exponents						
Definition				Numerical Value for 2D Ising Universality Class		
Specific heat ($B = 0$)	C	$\sim$	$\lvert\mathsf{t}\rvert^{-\alpha}$	α	$=$	0
Magnetization ($B = 0$)	M	$\sim$	$(-\mathsf{t})^{\beta}$	β	$=$	1/8
Susceptibility ($B = 0$)	χ	$\sim$	$\lvert\mathsf{t}\rvert^{-\gamma}$	γ	$=$	7/4
Critical isotherm (t = 0)	H	$\sim$	$\lvert M\rvert^{\delta}\mathrm{sgn}(M)$	δ	$=$	15
Correlation length	ξ	$\sim$	$\lvert\mathsf{t}\rvert^{-\nu}$	ν	$=$	1
Two-Spin Connected Corr. Fun.	$\Gamma(r)$	$\sim$	$1/r^{d-2+\eta}$	η	$=$	1/4

Table 12.1: Definitions of critical exponents and their numerical values for the two-dimensional Ising model university class. The quantity d is the dimension of the system; $\mathsf{t} \equiv (T - T_c)/T_c$. (After Tables 2.3 and 3.1 of Ref. [170].)

statistical physics. Universality is well-justified theoretically by the renormalization group and has been thoroughly verified experimentally. Universality tells us that despite the enormous range of systems that undergo critical phenomena, nature is somehow limited in the manner by which symmetry can be spontaneously broken. There are certain features of continuous phase transitions—namely the critical exponents—that are the same for many microscopically different systems.



(12.18)

Figure 12.4: A typical configuration of the Ising model for $T \gg T_c$. There are $10,000$ lattice sites, 5010 of which are up (black).

12.6 Pattern, Organization, and Computation in Two Dimensions Revisited

Having discussed the mathematical framework used to analyze critical phenomena, we now return to the physical picture presented by the two-dimensional Ising model. To so do, we examine the typical configurations above, below, and at the critical temperature. Typical configurations for these temperature regimes are plotted in Figs. 12.4-12.6. Up $(+1)$ spins are colored black, and down spins (-1) were left

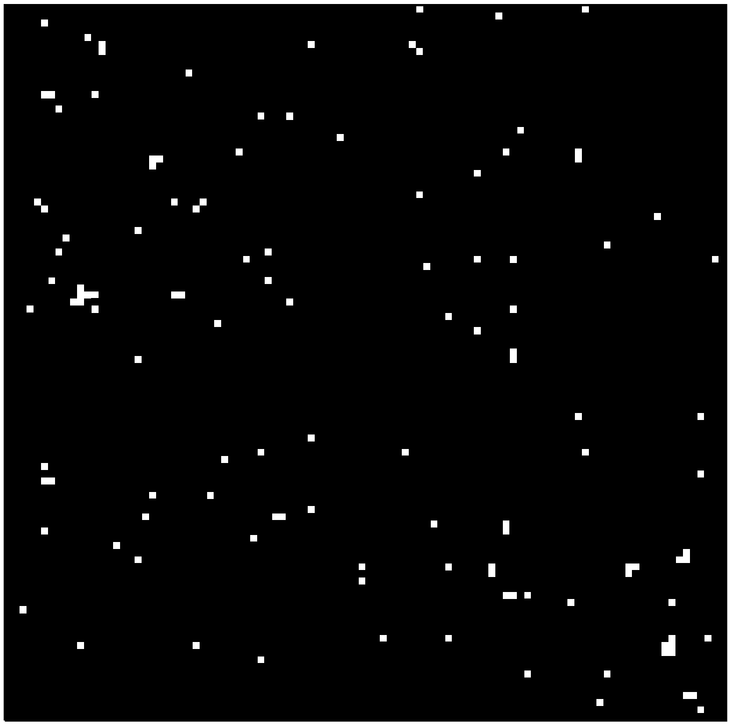


Figure 12.5: A typical configuration of the Ising model for $T \ll T_c$. There are 10,000 lattice sites, 9880 of which are up (black).

blank. The configurations were generated by a Monte Carlo simulation using standard methods; see, e.g., Ref. [170, Chpt. 7].

First, consider the situation in the high temperature limit. Well above T_c the magnetization m is zero, indicating that on average there are as many up (+1) spins as down (−1) spins. The thermal fluctuations (controlled by T) dominate, and the tendency for neighboring spins to align (controlled by J) is negligible. The result is a random arrangement of up and down spins. This can be seen in Fig. 12.4 where a

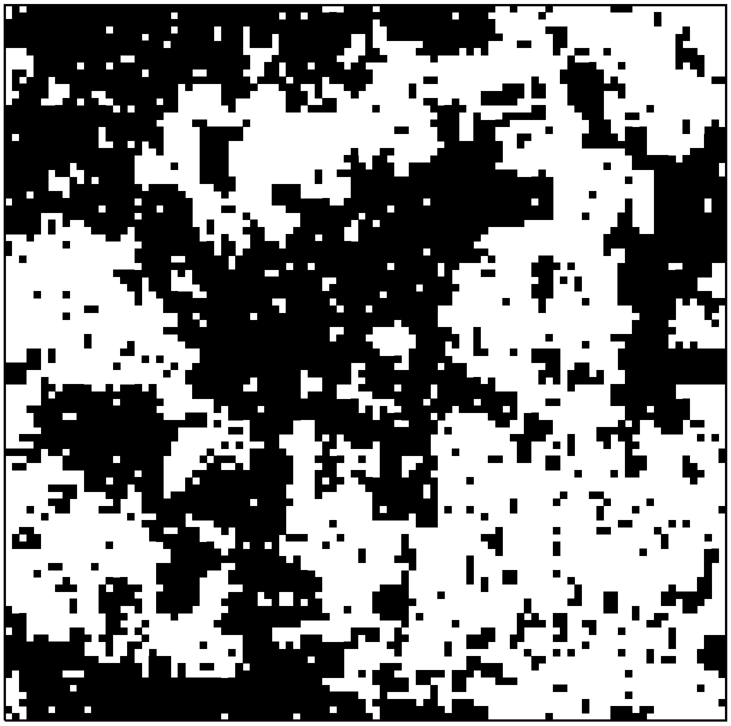


Figure 12.6: A typical configuration of the Ising model for $T \approx T_c$. There are $10,000$ lattice sites, 4838 of which are up (black).

typical configuration for $T \gg T_c$ is shown.

Well below the critical temperature, $m \approx 1$; almost all the spins are up. The thermal fluctuations are weak compared to the tendency for spins to align. A typical configuration for low temperature is shown in Fig. 12.5.

A typical configuration for $T \approx T_c$ is shown in Fig. 12.6. Like the high temperature configuration of Fig. 12.4, the magnetization is zero. However, the two configurations certainly are structured differently. In the critical configuration, Fig. 12.6

clusters of up spins of many different sizes can be seen, consonant with our statement above that at a critical point fluctuations at all length scales contribute to the observed phenomena.

As discussed in the above sections, at T_c the two-spin connected correlation function $\Gamma(r)$ decays algebraically with r and the correlation length diverges. Nevertheless, the lattice manages to balance up and down spins so that on average the magnetization is zero. The theory of critical phenomena describes this critical state by using the critical exponents to answer the question: What is the asymptotic form of the diverging thermodynamic functions at $T = T_c$?

In keeping with the theme of this dissertation, we ask a different set of questions. *How* does the lattice maintain $m = 0$ and have a diverging correlation length? How much information must be shared across the lattice to achieve and maintain this critical state? And what must be done with this information—how is information stored and manipulated to produce clusters of spins at all lengths scales? Are the critical configurations of Fig. 12.6 qualitatively different than the low and high temperature configurations of Figs. 12.5 and 12.4? If so, how to we express this difference?

To answer these questions, one could examine entropy convergence and apply computation mechanics as done in Part II when analyzing the structures exhibited by one-dimensional spin systems. However, a complete theoretical framework to do so is not yet in place. Preliminary work towards such a framework is discussed in

Chapter 14.

Chapter 13

Quasi-Two-Dimensional Results for the Two-Dimensional Ising System

13.1 Introduction

The two spin mutual information $I(r)$ defined in Eq. (4.14) of Chapter 4 in the context of a one-dimensional spin system can easily be extended to measure the information shared between spins on a two-dimensional lattice. To this end, we define:

$$I(n,m) \equiv I[S_{i,j}; S_{i+n,j+m}] \, , \tag{13.1}$$

where the mutual information I is given by Eq. (3.24). The quantity $I(n,m)$ measures the mutual information between spins separated by n lattice sites vertically and m

lattices sites horizontally. Note that by rotational symmetry $I(n,m) = I(m,n) = I(|n|,|m|)$.

The nearest-neighbor two-spin mutual information $I(1,0)$ of the two-dimensional Ising model has been estimated numerically via a Monte Carlo simulation by Solé *et al.* in Ref. [149]. The two-spin mutual information (not just the nearest neighbor case) has also been numerically estimated by Matsuda *et al.* in Ref. [109]. In this short chapter we present exact results for the nearest-neighbor two-spin mutual information for the two-dimensional Ising model.

13.2 Exact Results for Two-Spin Nearest-Neighbor Mutual Information

We begin our calculation of the nearest neighbor two-spin mutual information $I(1,0)$ by recalling that the mutual information can be written as the difference between the marginal and the conditional entropies;

$$I[S_{i,j}\,;\,S_{i+1,j}] \;=\; H[S_{i,j}] \;-\; H[S_{i,j}\;S_{i+1,j}]\;. \tag{13.2}$$

From Eq. (13.2) and the definition of conditional entropy, Eq. (3.20), it is clear that to calculate $I(1,0)$ we will need expressions for $\Pr(S_{i,j})$ and $\Pr(S_{i,j}, S_{i+1,j})$. One can obtain formulae for these in terms of the magnetization m and the nearest-neighbor correlation function $\Gamma_u(1,0)$. By exploiting the rotational and translational symmetries of the lattice and the normalization constraints on the various probabilities, we find:

$$\begin{aligned}
\Pr(\pm 1) &= \frac{1 \pm m}{2}\ , && (13.3)\\
\Pr(-1|1) &= \frac{1-m}{1+m}\left[\frac{1-\Gamma_u(1,0)}{2-m}\right]\ , && (13.4)\\
\Pr(-1|-1) &= \frac{1+\Gamma_u(1,0)-m}{2-m}\ , && (13.5)\\
\Pr(1|1) &= 1\ -\ \frac{1-m}{1+m}\left[\frac{1-\Gamma_u(1,0)}{2-m}\right]\ , && (13.6)\\
\Pr(1|-1) &= \frac{1-\Gamma_u(1,0)}{2-m}\ . && (13.7)
\end{aligned}$$

Similar results have been derived independently by Matsuda *et al.* in Ref. [109].

The details of the manipulations are shown in the Appendix F. We sketch the derivation here. First, for the single spin distribution $\Pr(S_{i,j})$ there is only one independent quantity since there are two possible values of $S_{i,j}$ and the probabilities must sum to one. Thus, $\Pr(S_{i,j})$ can be expressed solely as a function of m as is done in Eq. (13.3).

To determine the joint distribution of nearest-neighbor spins, we begin by factoring the joint distribution into the product of a marginal and a conditional distribution:

$$\Pr(S_{i,j}, S_{i+1,j}) = \Pr(S_{i,j}|S_{i+1,j})\Pr(S_{i+1,j})\ . \qquad (13.8)$$

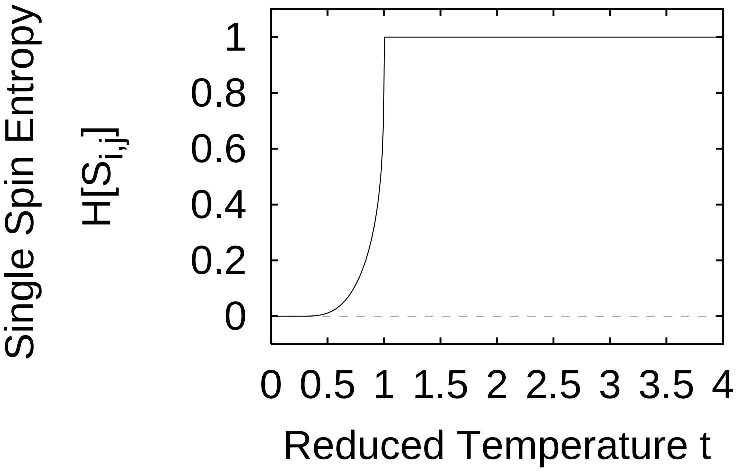


Figure 13.1: The single-spin entropy $H[S_{i,j}]$ versus the reduced temperature $t \equiv T/T_c$.

The marginal distribution $\Pr(S_{i+1,j})$ is known; $\Pr(S_{i+1,j}) = \Pr(S_{i,j})$ by translation symmetry. Thus, it suffices to determine the conditional distribution $\Pr(S_{i,j}|S_{i+1,j})$. There are four possibilities for the two spins, but two normalization conditions reduce the number of independent quantities to two. The rotational symmetry of the lattice implies that $\Pr(1,-1) = \Pr(-1,1)$. From this we deduce a relation between $\Pr(+1|-1)$ and $\Pr(-1|+1)$ in terms of m. We are then able to obtain Eqs. (13.4)–(13.7) which express $\Pr(S_{i,j}|S_{i+1,j})$ as a function of m and $\Gamma_u(1,0)$.

As mentioned above, note that the conditional probabilities exhibit a spin-flip symmetry above T_c but not below T_c. For example, one might expect that $\Pr(-1|1) = \Pr(1|-1)$. However this is not the case below $T = T_c$ where $m \neq 0$ and the the spin-flip symmetry is spontaneously broken.

We now have all the pieces needed to calculate the mutual information between

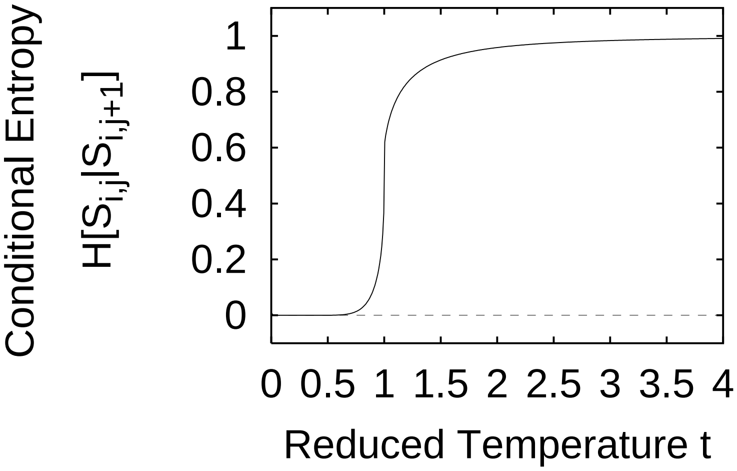


Figure 13.2: The conditional entropy $H[S_{i,j}|S_{i+1,j}]$ as a function of reduced temperature $t \equiv T/T_c$.

nearest-neighbor spins using the definition of Eq. (13.2). The probability distributions are given in Eqs. (13.3)–(13.7) and m and $\Gamma_u(0,1)$ are given by Eq. (12.2) and (12.7), respectively. The single spin entropy $H[S_{i,j}]$, the conditional entropy $H[S_{i,j}|S_{i,j+1}]$, and the nearest-neighbor two-spin mutual information $I(1,0)$ are plotted versus reduced temperature $t \equiv T/T_c$ in Figs. 13.1–13.3.

13.3 Discussion of Results

We first examine the behavior of the single-spin entropy $H[S_{i,j}]$, plotted in Fig. 13.1. Above T_c, the magnetization m is zero, indicating that, on average, equal numbers of spins are are up and down. As a result, the single-spin entropy is just that of a fair coin—1 bit—and $H[S_{i,j}] = 1$ for all $T > T_C$. As the temperature is lowered

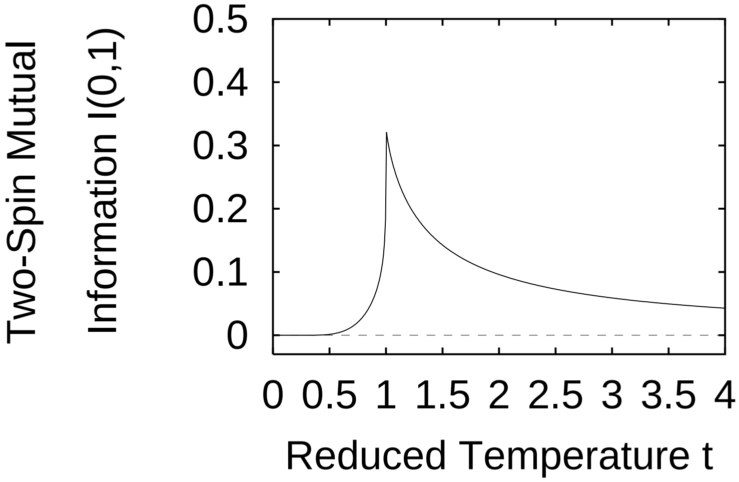


Figure 13.3: The two-point nearest-neighbor mutual information $I(1,0)$ as a function of reduced temperature. $t \equiv T/T_c$.

from T_c, a bias in the spins is introduced and m approaches 1. Accordingly, the single spin entropy decreases monotonically until $T = 0$, where the magnetization is 1 indicating that all spins are up. Hence, there is no uncertainty at zero temperature and $H[S_{i,j}] = 1$.

Next, consider the conditional entropy $H[S_{i,j}|S_{i,j+1}]$ shown in Fig. 13.2. At high temperatures, thermal fluctuations dominate and the spins are distributed almost independently. Thus, the conditional entropy is almost one bit. Below T_c, the conditional entropy is seen to vanish. To understand this, note that:

$$H[S_{i,j}|S_{i,j+1}] \;=\; H[S_{i,j}, S_{i,j+1}] - H[S_{i,j}] \ . \tag{13.9}$$

Below the critical temperature the spin-flip symmetry has been broken and almost all the spins are up. Both terms on the right-hand side of the above equation approach zero since there is little uncertainty associated with the value of any number of spins. As a result, the left-hand side of Eq. (13.9) goes to zero as T goes zero.

Finally, we consider the difference between the single-spin and the conditional entropies: two-spin nearest-neighbor mutual information:

$$I(0,1) \equiv I[S_{i,j}\,;\,S_{i+1,j}] \;=\; H[S_{i,j}] \;-\; H[S_{i,j}\;S_{i+1,j}]\;. \tag{13.10}$$

The quantity $I(0,1)$ is plotted as a function of t in Fig. (13.3). Note the sharp maximum in the mutual information at the critical temperature $t = 1$.

Well above the critical temperature, the thermal fluctuations overpower the coupling and the system is random. There are almost no correlations between neighboring spins. Thus, knowledge of one spin does not reduce our uncertainty of the value of the neighboring spins at all and the mutual information $I(0,1)$ vanishes. Said another way, there is a lot of Shannon information at each site, but this information is not shared.

Below the critical temperature, the spin-flip symmetry has been broken and almost all the spins point in the same direction. As a result of this bias, there is little uncertainty about the value of a single spin. Thus, knowledge of one spin does not lead to a large reduction in the uncertainty of the value of its neighbors—there was very little uncertainty in the first place. In short, there is little information at site to share.

At the critical temperature, the two-spin correlation function decays algebraically with distance, whereas it decays exponentially at all other temperatures. Thus, the two-spin mutual information achieves a maximum at T_c.

Note the very rapid fall of the mutual information below T_c. This is due to the spontaneous symmetry breaking; a bias is introduced (measured by m), breaking the symmetry between up and down spins. Thus, below T_c there is less information for the two neighboring spins to share, leading to a sharply falling mutual information.

13.4 Summary

In summary, we have presented exact results for the two-spin nearest-neighbor mutual information. This quantity has be estimated in Ref. [149] and, independently in Ref. [109], via a Monte Carlo simulation. (A related two-spin statistic, “Distance to Independence” was put forth and estimated in Ref. [148].) Since our results are exact, they are a clear improvement over these numerical estimates. For example, practically speaking, the mutual information, like all average quantities, requires a relatively large amount of computer time to estimate with a Monte Carlo algorithm due to critical slowing down near. For example, in both Refs. [109] and [149], the error bars on $I(1, 0)$ are quite large near T_c.

The two-spin mutual information is a function of the distribution of only two spins. As such, it does not provide a measure of global memory in the manner that the excess entropy does for a one-dimensional systems. Nor is it sufficiently "global" to be used as a general indicator of structure, pattern, or memory as the excess entropy is. How can one define an analogue of the excess entropy in two dimensions that explicitly accounts for the manner in which information is shared across space? Can the procedure through which ϵ-machines are defined be generalized so that the ϵ-machine can applied to multidimensional configurations? These questions will be addressed in the next chapter.

Chapter 14

Toward a Fully Two-Dimensional Generalization of Entropy Convergence and Computational Mechanics

In this chapter we discuss several ideas for extending our computation and information theoretic analysis of pattern, organization, and structure to more than one dimension. For simplicity, we shall focus on two-dimensional systems; our discussion can easily be extended to higher dimensions if needed. The work presented here is quite preliminary; it is not put forth as a final solution. Rather, our main goal is to state clearly the issues and difficulties that arise when trying to extend our results to more than one dimension. In so doing, we suggest several directions for future research.

We begin in Sec. 14.1 by discussing a number of different, complementary approaches to defining an information theoretic measure of memory for two-dimensional systems. Then, in Sec. 14.2 we review and extend an approach to two-dimensional computational mechanics introduced by Crutchfield and Hanson [69, 71]. In Sec. 14.3 several simple two-dimensional patterns are analyzed using the techniques of the previous two sections; the results of this analysis are discussed in Sec. 14.4. Finally, in Sec. 14.5 we mention some open questions and discuss possibilities for future work.

14.1 Information Theoretic Measures of Memory in Two Dimensions

The excess entropy $\mathbf{E}$ of a one-dimensional stationary chain of discrete random variables was defined in Chapter 4. There we saw that $\mathbf{E}$ can be interpreted as the apparent spatial memory of the system. There are several formulae for the excess entropy, each of which emphasizes a different point of view; $\mathbf{E}$ can be viewed as a mutual information, a sum of the finite-L overestimates of the entropy density h_μ, or as the constant term in the scaling form for the total entropy. Each of these views is conceptually distinct, but all three formulations can be shown to be equivalent.

In two dimensions, these different points of view lead to different possible generalizations of the excess entropy. In this section, we explore different ways of extending our information theoretic analysis of structure to two dimensions. Before discussing excess entropy, we first fix notation and state a few definitions.

Consider a two-dimensional array of discrete random variables $S_{i,j}$ where the variables are drawn from a finite set: $s_{i,j} \in \mathcal{A}$. Let $\mathcal{B}_{i,j}^{(M,N)}$ denote a rectangular $M \times N$ configuration whose upper left variable is $S_{i,j}$. The block has N rows and M columns. That is,

$$\mathcal{B}_{i,j}^{(M,N)} \equiv \begin{array}{ccccc} S_{i,j} & S_{i,j+1} & S_{i,j+2} & \cdots & S_{i,j+M-1} \\ S_{i+1,j} & S_{i+1,j+1} & S_{i+1,j+2} & \cdots & S_{i+1,j+M-1} \\ S_{i+2,j} & S_{i+2,j+1} & S_{i+2,j+2} & \cdots & S_{i+2,j+M-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ S_{i+N-1,j} & S_{i+N-1,j+1} & S_{i+N-1,j+2} & \cdots & S_{i+N-1,j+M-1} \end{array} . \qquad (14.1)$$

We assume that the configuration is translationally invariant, and thus

$$\Pr(\mathcal{B}_{i,j}^{(M,N)}) = \Pr(\mathcal{B}_{i+K,j+L}^{(M,N)}) \equiv \Pr(\mathcal{B}^{(M,N)}) \qquad (14.2)$$

for all K, L.

We define the block entropy $H(N.M)$ as the entropy of an $M \times N$ rectangle

$\mathcal{B}_{i,j}^{(M,N)}$ of adjoining variables:

$$H(N,M) \equiv H[\,\mathcal{B}^{(M,N)}\,] = -\sum_{\{\mathcal{B}\}} \Pr(\mathcal{B}^{(M,N)}) \log_2 \Pr(\mathcal{B}^{(M,N)}) \;. \qquad (14.3)$$

We then define the entropy density in the natural way:

$$h_\mu \equiv \lim_{M,N\to\infty} \frac{H(N,M)}{NM} \;. \qquad (14.4)$$

Lempel and Ziv show in Ref. [91] that this limit exists. We now proceed to explore several ways to define a two-dimensional analogue of the excess entropy. We shall apply these different definitions to some simple two-dimensional patterns in Sec. 14.3.

14.1.1 E as the intensive part(s) of the total entropy

We now consider our first candidate for a two-dimensional generalization of the excess entropy. Recall that we saw in Eq. (4.9) that $\mathbf{E}$ is equal to the $\mathcal{O}(1)$ term in the asymptotic expression for the block entropy for large L:

$$H(L) \sim \mathbf{E} + h_\mu L \;, \qquad (14.5)$$

where $H(L)$ is the entropy of a block of L consecutive spins and h_μ is the entropy density. This relationship was illustrated in Fig. 4.1.

Eq. (14.5) suggests the following scaling form for the two-dimensional block entropy:

$$H(M,N) \sim \mathbf{E}_s + h_\mu^{(y)} M + h_\mu^{(x)} N + h_\mu M N \ . \tag{14.6}$$

The subscript s denotes that this form of the excess entropy is arrived at by considering entropy scaling. The quantities $h_\mu^{(y)}$ and $h_\mu^{(y)}$ measure the entropy growth rates in the x and y directions. We shall see in Sec. 14.3 that $h_\mu^{(y)}$ and $h_\mu^{(y)}$ do not function as a measure of memory or structure.

We emphasize that Eq. (14.6) is, like Eq. (14.5), an *ansatz*; these asymptotic forms need not hold for all systems. For example, for the class of aperiodic one-dimensional sequences considered in Ref. [15], one can show that $H(L) \sim k_1 + k_2 \log L$, where k_1 and k_2 are constants that are independent of L. Eq. (14.6) does hold for the simple examples considered below in Sec. 14.3.

14.1.2 E as Mutual Information

In one dimension, the excess entropy can also be written as the mutual information between the left and right halves of the string as was done in Eq. (4.10):

$$\mathbf{E} \, = \, I[\overleftarrow{S};\overrightarrow{S}] \ . \tag{14.7}$$

This form can easily be generalized to two dimensions. However, what sets of variables should be used in place of the one-dimensional semi-infinite halves in Eq. (14.7)? One possibility is to consider the mutual information between half planes. Then $\mathbf{E}$ would be the information shared across a line, whereas in one dimension the excess entropy was the information shared across a point. Another, less general approach is to consider the mutual information between, say, upper left and lower right quadrants or "past" and "future" light-cones. The latter alternative could be appropriate if the two-dimensional block of variables is the space-time diagram of a one-dimensional CA. *A priori* there seems to be no way to choose one of these mutual information forms as being more natural than any other.

For use later on, we shall introduce notation for two of these mutual information forms of the excess entropy. First, we define $\mathbf{E}_I^{(\mathrm{TB})}$ as the mutual information between semi-infinite "top" and "bottom" half planes:

$$\mathbf{E}_I^{(\mathrm{TB})} \equiv \lim_{M,N\to\infty} I[\mathcal{B}_{i,j}^{(M,N)};\mathcal{B}_{i-N,j}^{(M,N)}] \ . \qquad (14.8)$$

Similarly, we define $\mathbf{E}_I^{(\mathrm{LR})}$ as the mutual information between semi-infinite "left" and "right" half planes:

$$\mathbf{E}_I^{(\mathrm{LR})} \equiv \lim_{M,N\to\infty} I[\mathcal{B}_{i,j}^{(M,N)};\mathcal{B}_{i,j+M}^{(M,N)}] \ . \qquad (14.9)$$

The subscript I on these two forms of the excess entropy indicates that they are defined via the mutual information. The superscripts make reference to the block variables between which we measure mutual information. By translational invariance, Eq. (14.2), the right-hand sides of Eqs. (14.8) and (14.9) do not depend on i, j.

14.1.3 E as the sum of finite-L h_μ overestimates

Our last formula for **E** in one dimension expresses the excess entropy as the sum of the finite-L h_μ overestimates. Recall that in Eq. (11.4) we defined:

$$h_\mu(L) \equiv H(L) - H(L-1) \, . \tag{14.10}$$

We then defined **E** by summing over L:

$$\mathbf{E} = \sum_{L=1}^{\infty} [\, h_\mu(L) - h_\mu \,] \, . \tag{14.11}$$

Note that in one dimension $h_\mu(L)$ may also be written as a entropy of a single spin conditioned on $L-1$ adjoining spins:

$$h_\mu(L) = H[S_L | S_1 \ldots S_{L-1}] \, . \tag{14.12}$$

This suggests defining a two-dimensional analogue of $h_\mu(L)$ as the entropy of a single spin conditioned on a neighborhood of adjoining spins. But what shape neighborhood is appropriate? Lindgren [53, 100, 101, 102] suggests the following approach.

Let $\mathcal{L}^M_{i,j}$ be an $M{+}1 \times 2M{+}1$ block of variables with the $M{+}1$ variables deleted from right side of the top row. That is:

$$\mathcal{L}^M_{i,j} \equiv \begin{array}{ccccccc} S_{i,j-M} & \cdots & S_{i,j-1} & & & \\ S_{i-1,j-M} & \cdots & S_{i-1,j-1} & S_{i-1,j} & \cdots & S_{i-1,j+M} \\ S_{i-2,j-M} & \cdots & S_{i-2,j-1} & S_{i-2,j} & \cdots & S_{i-2,j+M} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ S_{i-M,j-M} & \cdots & S_{i-M,j-1} & S_{i-M,j} & \cdots & S_{i-M,j+M} \end{array} . \qquad (14.13)$$

We then define the two-dimensional analogue of $h_\mu(L)$ as:

$$h^{\mathcal{L}}_\mu(M) \equiv H[\, S_{i,j} \,|\, \mathcal{L}^M_{i,j} \,] \, . \qquad (14.14)$$

This formula is illustrated in Fig 14.1. For each M, $h^{\mathcal{L}}_\mu(M)$ is the entropy of the shaded block conditioned on all the other blocks in the figure.

Assuming translation invariance, Eq. (14.2, it turns out that [100, 101]:

$$\lim_{M\to\infty} h^{\mathcal{L}}_\mu(M) = h_\mu \, , \qquad (14.15)$$

where the entropy density h_μ is defined by Eq. (14.4). The L-shaped blocks defined in Eq. (14.13) are not the only shape blocks for which Eq. (14.15) holds [100, 101, 102]. Consider a sequence of blocks $\widetilde{\mathcal{L}}^M$ with the following properties: the blocks are increasing in size; $\widetilde{\mathcal{L}}^{M-1}$ is included in $\widetilde{\mathcal{L}}^M$ for all M; and $\lim_{M\to\infty} \widetilde{\mathcal{L}}^M = \lim_{M\to\infty} \mathcal{L}^M$.

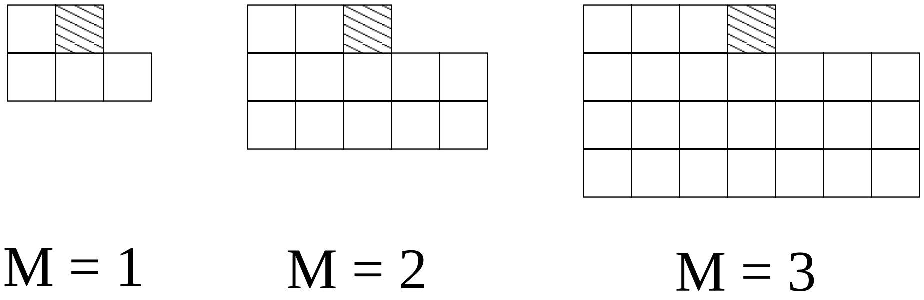


Figure 14.1: The construction of $h_\mu^{\mathcal{L}}(M)$. At each M, $h_\mu^{\mathcal{L}}(M)$ is given by the entropy of the shaded block conditioned on all the unshaded blocks.

For such a $\widetilde{\mathcal{L}}^M$ if follows that $\lim_{L\to\infty} h_\mu^{\widetilde{\mathcal{L}}}(M) = h_\mu$,[100, 101, 102].

Having defined the finite-size entropy density estimate $h_\mu^{\mathcal{L}}(M)$ we now define another form for the excess entropy by analogy with Eq. (14.11):

$$\mathbf{E}_r^{\mathcal{L}} \equiv \sum_{M=1}^{\infty} \left(h_\mu^{\mathcal{L}}(M) - h_\mu \right) \ , \tag{14.16}$$

The subscript r indicates that this form for the excess entropy is defined by summing up the successive *redundancies* $h_\mu^{\mathcal{L}}(M) - h_\mu$. The superscript refers to the shape of the neighborhood—the $\mathcal{L}$ of Eq. (14.13)—upon which we condition the entropy of a single spin to obtain a finite-size entropy density estimate.

14.2 Two-Dimensional Computational Mechanics

14.2.1 Issues Concerning Templates and Lattice Partitioning

A question implicitly arose as we considered all of the above proposals for the excess entropy in two dimensions: What sort of “template” should one use? That is, when building up larger and larger pieces of the lattice, as one would need to do for the entropy convergence and entropy scaling pictures, what shapes should the pieces be? The mutual information between what composite variables (e.g., blocks, quadrants) will yield a measure of the information shared across the lattice? Said another way, how should one parsimoniously group the individual random variables on the lattice to reduce unnecessary redundancy and focus on intrinsic features.

In one dimension, these questions concerning how to best group configurations are answered by the causal states. Recall that the causal states are defined as the minimal set of aggregate variables needed to perform optimal prediction. What is the appropriate two-dimensional analogue of causal states? In our view, this unanswered questions lies at the heart of the subtleties in generalizing the excess entropy and computational mechanics to more than one dimension.

One approach to two-dimensional computational mechanics considers the one-dimensional strings obtained when parsing the two-dimensional lattice along different paths. But which path should one choose and why? It seems clear that the view of the underlying two-dimensional structure will depend strongly on the path chosen.

This difficulty has also been articulated by Grassberger [64]. Lempel and Ziv [91] and Sheinwald, Lempel, Ziv [145] have employed a path approach as part of an encoding scheme for two-dimensional patterns. Specifically, they parse the two-dimensional pattern by observing the variables seen along a Peano-Hilbert space-filling curve. They find that this path is optimal, in that they prove a constructive coding theorem [91, 145] in which the length of the code (in bits) is arbitrarily close to the entropy density h_μ. However, they note that the Peano-Hilbert curve is not the only path for which their result holds [91].

In the following section we review one approach, suggested by Crutchfield [36] for parsing the configurations into one-dimensional strings: a stochastic generalization of the "space-time" machines introduced by Crutchfield and Hanson in Refs. [69] and [71].

14.2.2 One Approach to Two-Dimensional Computational Mechanics: Path ϵ-Machines

Our starting point is again a two-dimensional infinite square lattice of discrete random variables $S_{i,j}$ with values drawn from a finite set $\mathcal{A}$. The subscripts i, j denote lattice site. Imagine recording the values of the variables observed as one moves through the lattice along some particular path, recording at each step not only the variable seen, but also the direction of the move that takes one to that variable. We shall impose the significant restriction that only south (S) and east (E) moves are allowed. A south move corresponds to moving "down": from $S_{i,j}$ to $S_{i,j-1}$. An east move corresponds to moving "right": from $S_{i,j}$ to $S_{i+1,j}$. As we step through the lattice, we will indicate south moves with the symbol S and east moves with E.

The result of stepping through the lattice and recording the direction of the steps and the variables seen is a one-dimensional string over an expanded alphabet that is the Cartesian product of $\{S, E\}$ with the original alphabet $\mathcal{A}$. Denote this alphabet with the added spatial information by

$$\mathcal{A}_S \equiv \mathcal{A} \times \{S, E\} \ . \tag{14.17}$$

For example, if the two-dimensional configuration is binary, $\mathcal{A} = \{0, 1\}$, then $\mathcal{A}_S = \{0_S, 0_E, 1_S, 1_E\}$. Note that $|\mathcal{A}_S| = 2|\mathcal{A}|$.

As mentioned above, when parsing the two-dimensional lattice we restrict ourselves to only down (south) and right (east) moves. In so doing, we disallow any loops in our path. There is, of course, not a unique way to step through the lattice in this manner. For an L-step journey there are 2^L possible paths; at each step one is free to move either south S or E.

We determine a probability distribution over strings in the expanded alphabet as follows. We choose different lattice sites as our starting point, and from each consider all the length-L, no-loop paths. We consider each possible path through the patch only once; i.e., we assume each path has equal probability. We then take as the probability of an expanded-alphabet string its relative frequency in the limit that we consider an infinite number of starting points in the two-dimensional lattice. This method for generating a distribution over patches of two-dimensional variables has also been used by Lempel and Ziv [91].

We have now reduced the two-dimensional configuration to a ensemble of one-dimensional strings and thus can reconstruct (analytically or empirically) an ϵ-machine using this new, expanded-alphabet, one-dimensional ensemble. The result is a *path ϵ-machine* that describes the non-looping paths through the two-dimensional configuration. Examples of such ϵ-machines will be given in Sec. 14.3. The limitations of the path ϵ-machine approach are discussed in Sec. 14.4.3. In the following, we shall ignore the admittedly important issue of synchronization—how one determines by observing a system what causal state a process is in (see Appendix D)—and will focus only on the recurrent portions of path ϵ-machines.

14.2.3 Properties of Path ϵ-Machines

As noted above, the ensemble of sequences obtained by considering non-looping paths through the lattice can be viewed as a one-dimensional stationary stochastic process. As a result, we can proceed to calculate $\mathbf{E}$ using any of the one-dimensional formulae; all the one-dimensional formulae for $\mathbf{E}$ from Chapter 4 will yield identical numerical values in the limit that the path length goes to infinity. We denote the excess entropy calculated from the ensemble of one-dimensional non-looping paths by $\mathbf{E}^{\mathcal{P}}$. We denote by $C_\mu^{\mathcal{P}}$ the statistical complexity of the path ϵ-machine—i.e., the Shannon entropy of asymptotic probability of the causal states.

We denote the entropy density of the one-dimensional expanded-alphabet strings by $h_\mu^{\mathcal{P}}$. Note, however, that the random walk through the lattice introduces entropy above and beyond that intrinsic to the patch, and thus $h_\mu^{\mathcal{P}} \neq h_\mu$. This entropy arises as a result of the entropy of the random walk itself. Thus, the entropy density of our one-dimensional expanded-alphabet strings is larger than the entropy density h_μ of the two-dimensional system as defined in Eq. (14.4). Since the entropy density of the random walk is 1 bit, we might expect that $h_\mu^{\mathcal{P}} = 1 + h_\mu$. However, the examples treated in the following section will show that this is not always correct. The relationship between $h_\mu^{\mathcal{P}}$ and h_μ will be discussed again in Sec. 14.4.2.

However, the additional randomness introduce by the random walk will not affect $\mathbf{E}^{\mathcal{P}}$. This can be seen most clearly by writing $\mathbf{E}^{\mathcal{P}}$ as the mutual information between semi-infinite "past" and "future" sequences over the expanded alphabet strings:

$$\mathbf{E}^{\mathcal{P}} \equiv \lim_{L\to\infty} I[S_{-L}^L ; S_0^L] = \lim_{L\to\infty} \left(H[S_0^L] - H[S_0^L | S_{-L}^L] \right) , \tag{14.18}$$

where $S^L \in \mathcal{A}_S^L$. The move through the lattice taken at a given step is independent of the sequence of east or south moves taken previously. As a result, the entropy of the random walk will contribute equally to both terms on the right-hand side of

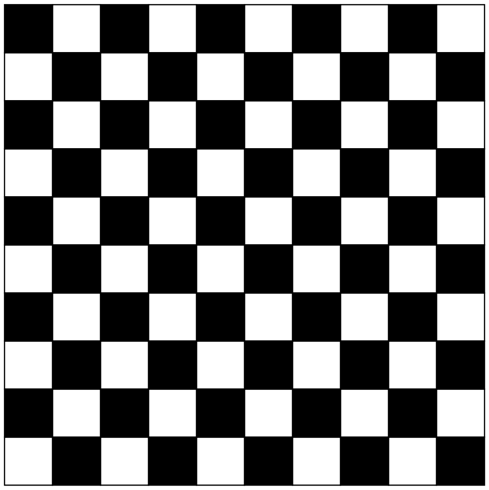


Figure 14.2: A two-dimensional checkerboard pattern.

Eq. (14.18). Thus, the difference of these two terms, $\mathbf{E}^{\mathcal{P}}$, will not be effected and will still serve as a measure of memory.

14.3 Examples

14.3.1 Checkerboard

In this section we calculate the path ϵ-machines and the different versions of the excess entropy for four simple configurations. We begin by considering a checkerboard pattern, shown in Fig. 14.2.

To determine the path ϵ-machine we must form the ensemble of non-looping paths through square patches. We begin considering all length 3 paths. If we choose as our starting point a black square, then, parsing Fig. 14.2 we see that we will always see the sequence 010 regardless of the spatial moves made while moving through the lattice. (We denote black squares with 1 and empty square with 0.) Thus, we find the following $2^3 = 8$ paths:

$$\begin{array}{c} \{0_E1_E0_E,\ 0_E1_E0_S,\ 0_E1_S0_E,\ 0_E1_S0_S, \\ 0_S1_E0_E,\ 0_S1_E0_S,\ 0_S1_S0_E,\ 0_S1_S0_S\} \end{array} . \tag{14.19}$$

We denote black squares with 1 and empty square with 0.

If we start our path on a white square, then we obtain these paths:

$$\begin{array}{c} \{1_E0_E1_E,\ 1_E0_E1_S,\ 1_E0_S1_E,\ 1_E0_S1_S, \\ 1_S0_E1_E,\ 1_S0_E1_S,\ 1_S0_S1_E,\ 1_S0_S1_S\} \end{array} . \tag{14.20}$$

The expanded-alphabet strings of Eqs. (14.19) and (14.20) are the only possible length $L = 3$ south-east paths through the configuration. Note that it is not always the case that all paths can be obtained by considering only two starting locations; this holds only for particularly simple patterns. Here, that only two starting points need to be considered to generate all possible strings is due to the fact that the pattern is periodic with period 2.

It is not hard to see that all of the paths shown above are equally likely. Using the methods of Chapter 5, we can proceed to determine the path ϵ-machine using these one-dimensional expanded-alphabet strings. For this simple example, it turns

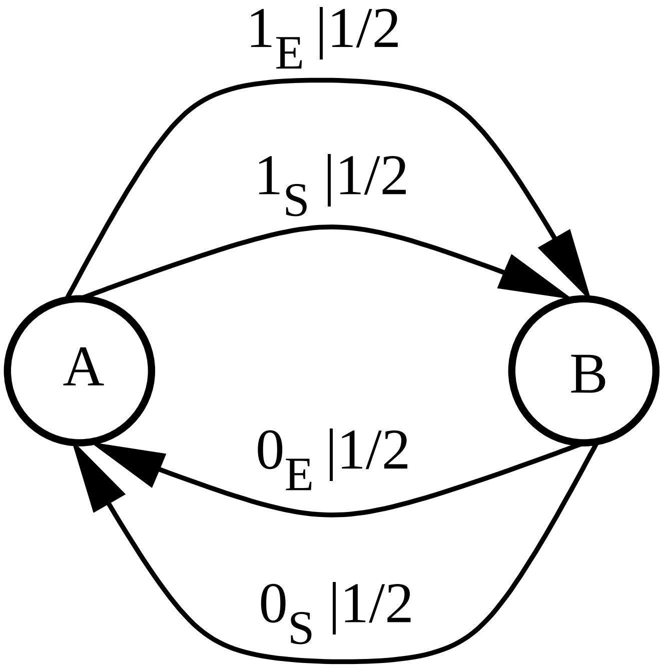


Figure 14.3: The path ϵ-machine for the checkerboard configuration of Fig. 14.2.

out that all one needs to determine the causal states are length-2 paths. No new causal states arise if longer paths are considered.

The recurrent portion of the resulting path ϵ-machine is shown in Fig. 14.3. This machine has a statistical complexity of 1; there are two causal states, each visited equally often. The properties of the path ϵ-machines for the checkerboard pattern and for all other examples in this section are summarized in Table 14.1.

The excess entropy of the path ϵ-machine is determined by investigating the behavior of $H(L)$, the total entropy of length-L paths. There are 4 length 1 paths: $0_E, 1_E, 0_S$, and 1_S, all equally likely. Thus, $H(L) = 2$. There are 8 length 2 paths, 0_E1_E, 0_E1_S, 0_S1_E, 0_S1_S, 1_E0_E, 1_E0_S, 1_S0_E, and 1_S0_S, and again, all are equally likely

Configuration	Path ϵ-Machine Properties			Entropy Density
	$h_\mu^\mathcal{P}$	$C_\mu^\mathcal{P}$	$\mathbf{E}^\mathcal{P}$	h_μ
Checkerboard	1	1	1	0
Coin Tosses	2	0	0	1
Period Two Stripes	1	1	1	0
Random Stripes	1.5	1	0.5	0.5

Table 14.1: Summary of the path ϵ-machine properties for the four example configurations of Figs. 14.2, 14.4, 14.6, and 14.8.

so $H(2) = 3$. In general,

$$\text{Number of paths of length } L \;=\; 2 \times 2^L \;, \tag{14.21}$$

since there are two phases of the pattern, $1010\cdots$ or $0101\cdots$, and 2^L paths for a given L. All paths are equally likely. Thus,

$$H(L) \;=\; 1 + L \;. \tag{14.22}$$

Thus, from the scaling form for $H(L)$, Eq. (14.5), we see that $h_\mu^{\mathcal{P}} = 1$ and $E^{\mathcal{P}} = 1$.

We now calculate the forms of the excess entropy defined in Sec.14.1. The results of these calculations are summarized in Fig. 14.2. For the configuration of Fig. 14.2 it is not hard to see that:

$$H(M,N) \;=\; 1 + 0M + 0N + 0MN \;. \tag{14.23}$$

As a result, it follows from Eq. (14.6) that $\mathbf{E}_s = 1$, ${h_\mu}^{(x)} = {h_\mu}^{(y)} = 0$, and $h_\mu = 0$. The top-bottom and left-right mutual information forms for $\mathbf{E}$ each yield a value of 1. To see this observe that there are two possible values of (say) the top half, corresponding to the two different phases of the checkerboard pattern. Thus, the entropy of the top half is 1 bit. If the bottom half is known, then this determines the value of the top half as well. Hence, knowledge of the bottom half reduces the entropy of the top half by 1 bit, so the mutual information between the two halves is 1 bit.

The finite-size entropy density estimates are easily found: $h_\mu^{\mathcal{L}}(0) \equiv H[S_{i,j}] = 1$, and $h_\mu^{\mathcal{L}}(M) = 0$ for all $M \geq 1$. As a result, $\mathbf{E}_r^{\mathcal{L}} = 1$. In summary, these values for the excess entropy seem reasonable. All forms of yield $\mathbf{E} = 1$, consistent with our discussion in Chapters 5 and 8 where we saw that for one-dimensional processes of period k, $\mathbf{E} = log_2 k$.

Configuration	Excess Entropy Form					Entropy Density		
	$\mathbf{E}_s$	$\mathbf{E}_I^{(TB)}$	$\mathbf{E}_I^{(LR)}$	$\mathbf{E}_r^{\mathcal{L}}$	$\mathbf{E}^{\mathcal{P}}$	${h_\mu}^{(x)}$	${h_\mu}^{(y)}$	h_μ
Checkerboard	1	1	1	1	1	0	0	0
Coin Tosses	0	0	0	0	0	0	0	1
Period Two Stripes	1	1	1	1	1	0	0	0
Random Stripes	0	∞	0	1	0.5	1	0	0

Table 14.2: Different forms of the excess entropy **E** calculated for the four example configurations of Figs. 14.2, 14.4, 14.6 and 14.8.

14.3.2 Coin Flips

As our next example, consider Fig. 14.4. Here, the variables are independently and identically distributed (iid); each variable is 1 with a probability of one half. It is easy to see that the path ϵ-machine for this configuration has only one causal state; there are no correlations between variables. The path ϵ-machine is shown in Fig. 14.5. It follows from the machine that $h_\mu^{\mathcal{P}} = 2$ and $\mathbf{E}^{\mathcal{P}} = C_\mu^{\mathcal{P}} = 0$.

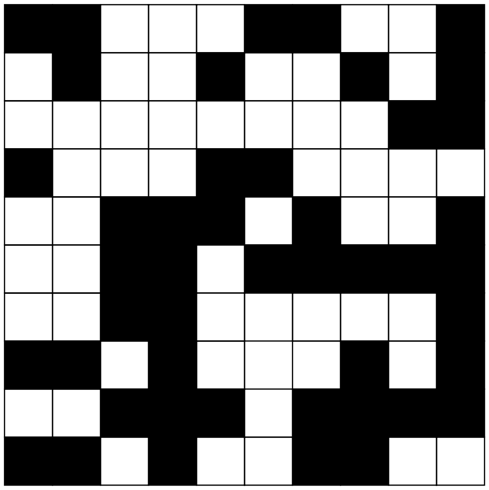


Figure 14.4: A two-dimensional random configuration. Each variable is independently distributed and equally likely to be 1 (black) or 0 (white).

The block entropy for the configuration of Fig. 14.4 is given by:

$$H(M, N) \;=\; 0 + 0M + 0N + 1MN \;, \tag{14.24}$$

and thus $\mathbf{E}_s = \mathbf{E}_s^{(M)} = \mathbf{E}_s^{(N)} = 0$ and $h_\mu = 1$. Since the variables are identically distributed no variables contain information about any others and $\mathbf{E}_I^{(TB)} = \mathbf{E}_I^{(LR)} = 0$. The finite-size approximation to the entropy density $H_\mu^{\mathcal{L}}(M)$ is 1 for all M. Thus, $\mathbf{E}_r^{\mathcal{L}} = 0$.

As expected, all forms of the excess entropy vanish for this system. The symbols are independently distributed; no information whatsoever is shared across the lattice.

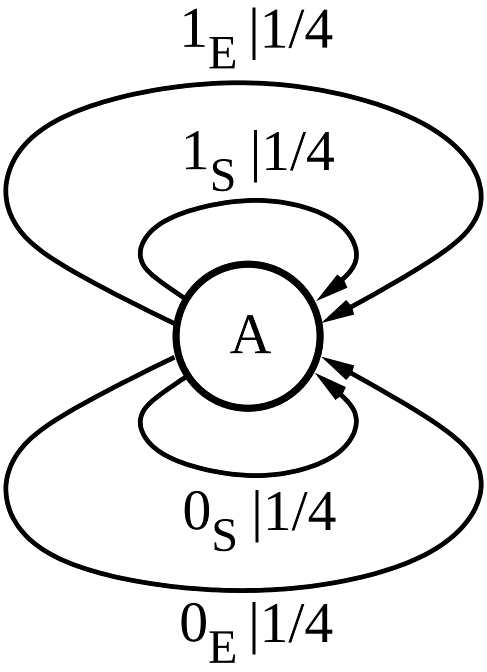


Figure 14.5: The path ϵ-machine for the completely random pattern of Fig. 14.4.

14.3.3 Period-2 Stripes

As our next example, we consider a configuration of alternating all-black and all-white vertical stripes as shown in Fig. 14.6. Following an analysis similar to that used for the checkerboard pattern in Sec. 14.3.1, we find that the path ϵ-machine for Fig. 14.6 is given by the machine shown in Fig. 14.7. This path ϵ-machine has a statistical complexity of 1. There are two causal states, each visited equally often as the ϵ-machine is run arbitrarily long.

The entropy $H(L)$ of the length L expanded-alphabet strings produced by the

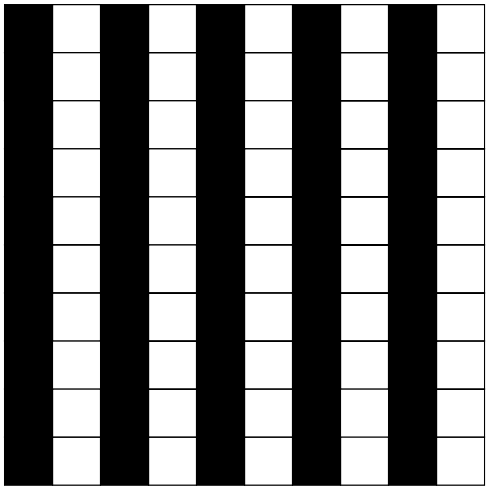


Figure 14.6: Periodic Stripes: alternating columns of all up (black) or down (white).

machine is identical to that found for the checkerboard pattern, Eq. (14.22):

$$H(L) \;=\; 1 + L \; . \tag{14.25}$$

Thus, $h_\mu^{\mathcal{P}} = 1$ and $E^{\mathcal{P}} = 1$.

The block entropy for the period-2 stripes is given by:

$$H(M,N) \;= 1 + 0M + 0N + 0MN \tag{14.26}$$

from which we conclude that $\mathbf{E}_s = 1$, ${h_\mu}^{(M)} = {h_\mu}^{(N)} = 0$, and $h_\mu = 0$.

The mutual information between the left and right halves is 1 bit. There are two possible right halves corresponding to the two phases of the pattern. As a result, the marginal entropy of the right half is 1. Once the left half is known, the right half

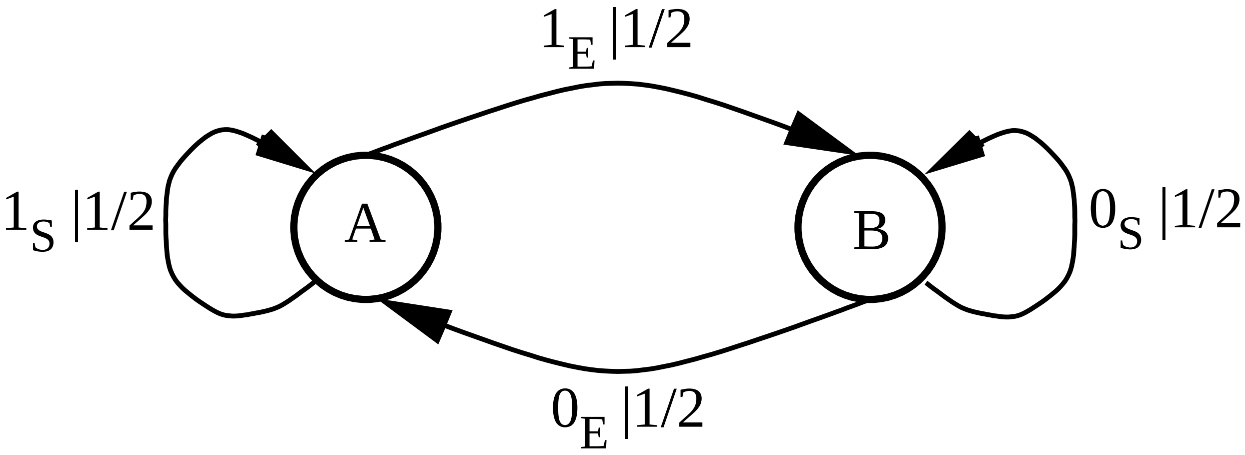


Figure 14.7: The path ϵ-machine for the period-2 striped configuration of Fig. 14.6.

is determined. Thus the entropy of the left half conditioned on the right half is zero. Subtracting the marginal and conditional entropies to obtain the mutual information, we see that $\mathbf{E}_I^{(LR)} = 1$. By a similar line of reasoning, one finds the same numerical value for the mutual information between the top and bottom half planes; $\mathbf{E}_I^{(TB)} = 1$.

The conditional entropies $h_\mu^{\mathcal{L}}(M)$ are seen to be identical to those of the checkerboard pattern: $h_\mu^{\mathcal{L}}(0) = 1$, and $h_\mu^{\mathcal{L}}(M) = 0$ for all $M \geq 1$. Thus, $\mathbf{E}_r^{\mathcal{L}} = 1$. As was the case for the checkerboard pattern, we find that all forms of the excess entropy are 1. This is not a surprise; like the checkerboard, the striped pattern of Fig. 14.6 is periodic with period 2.

14.3.4 Random Stripes

Lastly, we consider the "random stripes" pattern of Fig. 14.8. Here the configuration consists of columns of either all-black or all-white; the columns themselves are iid

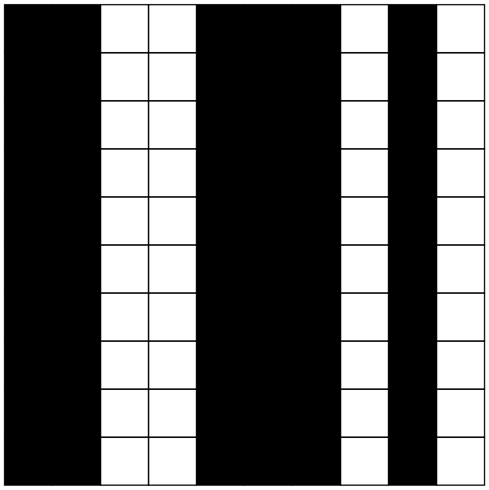


Figure 14.8: A two-dimensional "random stripes" pattern. Columns of all black or white occur independently with probability 1/2.

with probability 1/2 of being all black. The path ϵ-machine for this pattern is shown in Fig. 14.8. This machine has a statistical complexity of 1, since, like the machines of Figs. 14.3 and 14.7, there are two causal states, each visited equally frequently.

The entropy density $h_\mu^{\mathcal{P}}$ is 1.5. To see this, recall that the entropy density is given by the average per-step uncertainty of the machine, as discussed in Sec. 5.6.4. From each causal state, A and B, the uncertainty is the same; there is one transition with probability 1/2 and two transitions with probability 1/4. Thus,

$$h_\mu^{\mathcal{P}} = -\frac{1}{2}\log_2\frac{1}{2} - 2\frac{1}{4}\log_2\frac{1}{4} = \frac{3}{2} \,. \tag{14.27}$$

The excess entropy can be found by using Eq. (5.28), which expresses the probability of a particular sequence of observed variables in terms of the ϵ-machine.

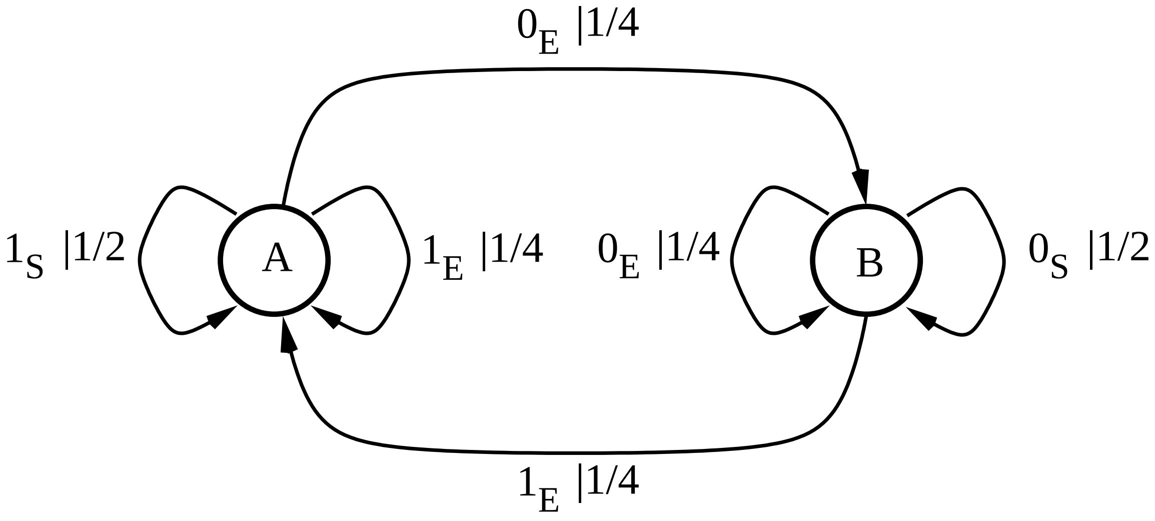


Figure 14.9: The path ϵ-machine for the random stripes configuration of Fig. 14.8.

Using this, we find $H(1) = 2$ and,

$$H(L) = \frac{1}{2} + \frac{3}{2}L \ , \quad L \geq 2 \ . \tag{14.28}$$

Hence, $\mathbf{E}^{\mathcal{P}} = 0.5$ and $h_\mu^{\mathcal{P}} = 1.5$, in agreement with Eq. (14.27).

We now calculate the other forms for the excess entropy. The block entropy is given by:

$$H(M,N) = 0 + 1M + 0N + 0MN \ . \tag{14.29}$$

Thus, $\mathbf{E}_s = {h_\mu}^{(y)} = 0$ while ${h_\mu}^{(x)} = 1$ and $h_\mu = 0$. This equation indicates, as expected, that the configuration is random in the x (horizontal) direction and ordered in the y (vertical) direction. Note that despite the fact that $H(N,M)$ diverges as M and N go to infinity, h_μ is zero.

The behavior of the mutual information forms of the excess entropy is rather interesting. Since the columns are independently distributed, the left half of the configuration contains no information about the right half and $\mathbf{E}_I^{(LR)} = 0$. This certainly does not hold for the mutual information between the top and bottom half configurations. The entropy of a particular $M \times N$ block is N bits. However, if the variables in the block immediately below (or above) this block are known, then the entropy of the $M \times N$ block is now zero, since knowing the value of one row determines all the other variables in the lattice. Thus, the mutual information between two $M \times N$ blocks positioned on top of another is equal to N bits. In the limit that the block size diverges, the mutual information diverges as well, and $\mathbf{E}_I^{(TB)} = \infty$.

14.4 Discussion of Examples

14.4.1 Different forms of the excess entropy

Due to the simplicity of the examples just considered, it is difficult to draw conclusions about interpretation of the different excess entropy forms. Nevertheless, some tentative statements can be made. First, as anticipated, the linear entropy scaling terms, ${h_\mu}^{(x)}$ and ${h_\mu}^{(y)}$ in the expression for the total block entropy, Eq. (14.6), do appear to be sensitive to randomness, not memory or structure. A non-zero value for ${h_\mu}^{(x)}$ or ${h_\mu}^{(y)}$ indicates that the entropy grows linearly in, respectively, the horizontal or vertical direction. Accordingly, ${h_\mu}^{(x)}$ and ${h_\mu}^{(y)}$ are zero for all the examples we considered, except for the random stripes configuration of Fig. 14.8. (See Table 14.2.) The stripes are distributed independently in the horizontal direction, leading to a ${h_\mu}^{(x)}$ of 1. Thus, we conclude that ${h_\mu}^{(x)}$ and ${h_\mu}^{(y)}$ measure randomness, not memory.

Second, the examples indicate that the mutual information form of the excess entropy is very sensitive to the shape of the regions considered: top and bottom half-planes, left and right half-planes, etc. This is seen in the random stripes example, where $\mathbf{E}_I^{(TB)} = \infty$ and $\mathbf{E}_I^{(LR)} = 0$. It seems unreasonable to assign the random stripes pattern an infinite memory; it possesses very little structure, as reflected in its two-state path ϵ-machine, Fig. 14.9. A possible remedy for this divergence to normalize the mutual information forms of the excess entropy by dividing by the linear dimension of the interface between the two blocks and then examining the limit in which the block size goes to infinity. The result is a mutual information density. Note, however, that this approach will yield a zero mutual information density for the checkerboard and period-2 striped patterns. Thus, the mutual information densities complement, rather than replace, the mutual information forms for the excess entropy, $\mathbf{E}_I^{(TB)}$ and $\mathbf{E}_I^{(LR)}$.

Based on the simple examples we've considered here, $\mathbf{E}_s$, $\mathbf{E}_I^{\mathcal{L}}$, and $\mathbf{E}^{\mathcal{P}}$ appear to be the most promising as a global measure of a two-dimensional configuration's memory. However, it is quite clear that a single scalar function cannot adequately capture all aspects of a system's memory. To some extent, this bluntness of the excess entropy was the case in one dimension, where, for example, all periodic systems of period p have an excess entropy of $\log_2 p$.

But the situation is even more subtle in two dimensions. In one dimension, it is natural to inquire about the memory shared between the left and the right halves of the string. In two dimensions, there is not a unique natural way to divide up the configuration. As a result, in analogy with the directional derivative of a function of two variables, it might be best to formulate the excess entropy as a directional quantity.

14.4.2 The entropy density of the path ϵ-machine

As noted above, the random non-looping walk through the lattice produces randomness above h_μ, the randomness intrinsic to the system as defined in Eq. (14.4). The entropy density associated with this random walk is 1 bit; for each path of length L there are 2^L possible sequences of south and east moves. Thus, it seems reasonable to suppose that $h_\mu^{\mathcal{P}}$, the entropy density of the path ϵ-machine is given by $1 + h_\mu$. However, the examples of the previous section show that this is not the case. For the random stripes pattern of Fig. 14.8, we found that $h_\mu = 0$ while $h_\mu^{\mathcal{P}} = 1.5$. Given this, it is unclear how to interpret $h_\mu^{\mathcal{P}}$.

14.4.3 Path ϵ-machine limitations

Note that the checkerboard and the period-2 striped patterns have identical values for all the quantities we calculated in the above section: all the different forms of $\mathbf{E}$, the entropy density h_μ, and the path ϵ-machine properties $h_\mu^{\mathcal{P}}$, $C_\mu^{\mathcal{P}}$ and $\mathbf{E}^{\mathcal{P}}$. Nevertheless, the two configurations are clearly different.

The situation is similar in one dimension. Two different period 4 patterns, $\cdots 110011001100 \cdots$ and $\cdots 111011101110 \cdots$, will both have $\mathbf{E} = C_\mu = 2$ and $h_\mu = 0$. To distinguish between these patterns, we need to look at their ϵ-machines. As discussed in Parts I and II, the ϵ-machine is a minimal representation of all the regularities present in the system. Said another way, knowledge of the ϵ-machine is sufficient to reproduce the joint distribution of any finite-sized block of variables.

In two dimensions, the path ϵ-machine does distinguish between the checkerboard and striped patterns. The machines of Figs. 14.3 and 14.7 are clearly structured differently. Can we conclude, as we did in one dimension for ϵ-machines, that the path ϵ-machine is a complete representation of the exact and approximate symmetries of the two-dimensional configuration? Can the path ϵ-machine be used to determine the joint distribution over any finite block of variables?

The answer to these two questions is no. When constructing the path ϵ-machine, one effectively forms a histogram of the same variables seen along different paths. This means that each path is treated independently. As a result, one needs more than the probabilities the sequences to construct the joint distribution over clusters of variables.

For example, suppose we wish to determine the joint distribution $\Pr(s_{i,j}, s_{i,j+1}, s_{i+1,j})$ over the adjoining variables $S_{i,j}, S_{i,j+1}, S_{i+1,j}$. Without loss of generality, we assume a binary alphabet. There are then $2^3 = 8$ possible values the 3 variables can assume. Since the probability distribution of the 3-spin cluster must be normalized, there are thus $8 - 1 = 7$ independent degrees of freedom in $\Pr(s_{i,j}, s_{i,j+1}, s_{i+1,j})$. However, as we will now show, the path ϵ-machine alone is not sufficient to deduce $\Pr(s_{i,j}, s_{i,j+1}, s_{i+1,j})$.

The path ϵ-machine tells the probability distribution of the sequences seen by making "east" and "south" moves: $\Pr(s_{i,j}, si, j+1)$ and $\Pr(s_{i,j}, si+1, j)$, respectively. We factor these probability distributions as follows:

$$\Pr(s_{i,j}, s_{i,j+1}) \;=\; \Pr(s_{i,j+1}|s_{i,j})\Pr(s_{i,j}) \; . \tag{14.30}$$

and

$$\Pr(s_{i,j}, s_{i+1,j}) \;=\; \Pr(s_{i+1,j}|s_{i,j})\Pr(s_{i,j}) \; . \tag{14.31}$$

Consider, say, Eq. (14.31). For each of the two values of $S_{i,j}$ there are $2^1 = 2$ possible values of $S_{i+1,j}$. Normalization demands that the distribution sums to one. As a result, there are $(2 - 1) \times 2 = 2$ independent degree of freedom in the conditional distribution $\Pr(s_{i+1,j}|s_{i,j})$ in Eq. (14.31)—one degree of freedom for each of the two possible values of $S_{i,j}$. Similarly, there are 2 degrees of freedom in Eq. (14.30)'s $\Pr(s_{i,j+1}|s_{i,j})$. There is one additional degree of freedom in $\Pr(s_{i,j})$; there are two values of $S_{i,j}$ minus one normalization constraint. Note that both of the distributions depend on $\Pr(s_{i,j})$.

Thus, there are $(2 \times 2) + 1 = 5$ independent degrees of freedom that we can obtain from the path ϵ-machine via the joint probability distributions of Eqs. (14.30) and (14.31). Yet there are 7 independent quantities in $\Pr(s_{i,j}, s_{i,j+1}, s_{i+1,j}, s_{i+1,j+1})$. We conclude, then, that it is not possible to determine the joint probability distribution of the cluster of variables $S_{i,j}, S_{i,j+1}, S_{i+1,j}$ given the path ϵ-machine; some additional information is required.

Given this, we see that the path ϵ-machine cannot reproduce the joint distribution over patches of random variables. As a result, we cannot interpret the path ϵ-machine as the minimal model capable of reproducing the original configuration. All the path ϵ-machine can do is reproduce paths—not at all surprising given how the machine was constructed. Thus, although the path ϵ-machine does provide useful structural information, as the examples of the previous section illustrate, the particular form of the path ϵ-machine introduced here does not provide a complete description of the regularities of the configuration. The Construction of an ϵ-machine (path or otherwise) that is capable of reproducing the distribution over any finite cluster of variables from a two-dimensional configuration remains an open challenge.

14.5 Future Work

As the above discussion indicates, there is clearly much work to be done to obtain a satisfactory set of tools to discover and quantify pattern, organization, and memory in more than one dimension. The examples worked out in Sec. 14.3 are all quite simple. Each of the four patterns may be viewed as the “direct product” of two one-dimensional patterns. For example, the random stripes pattern of Fig. 14.8 is the product of a coin flip in the x-direction and a period-1 pattern in the y-direction. An important (and probably straightforward) next step is to develop a general theory that relates the two-dimensional properties of such “direct product” patterns to the properties of the one-dimensional patterns of which they are composed.

However, to develop an appropriate information-theoretic measure of memory, it will be necessary to consider richer examples than those considered thus far. In a recent paper, Lindgren, Moore and Nordahl [103] explore formal language theory as applied to two-dimensional systems. It would be of interest to determine the different excess entropies for some of the systems they consider. (Most of their examples are not decomposable into two one-dimensional patterns.)

As mentioned above, we feel that a single scalar measure of memory will not adequately capture the manner in which information is shared across a two-dimensional lattice. To formulate a directional excess entropy, it will be necessary to understand more clearly the behavior of the derivatives of $H(L)$ in one dimension, building on the analysis of Ref. [104]. An examination along these lines will be published elsewhere [39].

While the path ϵ-machines defined in Sec. 14.2 do capture many of a two-dimensional configuration, we showed in Sec. 14.4.3 that they fail to provide a complete description of the joint distribution over clusters of adjoining spins. Thus, a pressing open problem is to develop a generalization of ϵ-machines that is able to reproduce the joint distribution of *all* finite patches of variables. Unlike the above path ϵ-machines, such a machine would be a representation of all the regularities and symmetries of the two-dimensional configuration.

Further down the road, it will be of great interest to apply the techniques developed here to configurations generated by the two-dimensional Ising model. Do the measures of memory diverge at the critical temperature, as one would expect? What critical exponents describe this divergence? Is an infinite number of causal states needed to describe the system at all temperatures or just at the critical temperature? How are these infinite causal states organized? Are the causal state transitions structured similarly to those found from the symbolic dynamics of the logistic map at the period-doubling accumulation point [44, 45]?

Is the intrinsic computation of the two-dimensional Ising system universal in the way that the critical exponents of Sec. 12.5 are? If they are universal, do they scale with different exponents that those defined in Table 12.1? Or do computational mechanical and information theoretic quantities fail to be universal? Memory and intrinsic computation reflect rather more detailed features of the underlying distributions than typical universal quantities, such as the two-spin correlation function or the specific heat. Thus, it is likely that computational mechanical and information theoretic quantities are not universal. If so, this indicates that focusing on critical exponents—as is widely done in statistical mechanics—is not sufficient to capture information processing and intrinsic computation.

Part V

Conclusion

Chapter 15

Conclusion

15.1 Summary

We began in Part I with a review of three complementary approaches to correlational structure. This review commenced in chapter 2 where we briefly recounted statistical mechanical measures of structure: the correlation length ξ, the two-spin correlation function $\Gamma(r)$, and the structure factors $S(q)$. In chapters 3 and 4 we discussed an information theoretic approach to memory and structure, first reviewing different forms of the Shannon entropy H and then focusing on entropy density convergence and the excess entropy $\mathbf{E}$. Lastly, in Chapter 5 we reviewed computational mechanics, a computation theoretic approach to memory and structure, and introduced the ϵ-machine, a minimal representation of the deterministic and statistical regularities of a system.

In Part II we turned our attention to one-dimensional finite-range classical spin systems. The statistical mechanics of such spin models was reviewed in Chapter 6. In Chapter 7 we gave exact analytic results for the excess entropy $\mathbf{E}$, ϵ-machines, and related quantities for this class of spin systems. The next three chapters developed a direct comparison of statistical mechanical, information theoretic, and computational mechanical approaches to structural complexity. There were three main conclusions that emerged as a result of these comparisons.

First, in Chapter 8 we saw that the excess entropy $\mathbf{E}$ serves as a wavelength-independent measure of periodic structure. This is an entirely general result, true for any stationary stochastic process—not just finite-range spin systems. In particular, if a system is periodic with period $\mathcal{P}$ then $h_\mu = 0$ and $\mathbf{E} = C_\mu = \log_2 \mathcal{P}$. We also found that, for a spin system with range-R where the R-spin blocks are in a one-to-one relation with the causal states, $C_\mu = \mathbf{E} + Rh_\mu$. Thus, for any such system with a positive entropy density h_μ, C_μ, the memory needed to statistically reproduce the configuration, is greater than $\mathbf{E}$, the system's apparent spatial memory.

Second, in Chapter 9 we showed that to fully capture the structure in entropic systems, one must examine ϵ-machines. An ϵ-machine reveals how the memory is organized and gives all of the system's measure semigroup theoretic properties. Said somewhat informally, the ϵ-machine is an irreducible representation of all the regularities—approximate and exact—possessed by the configuration.

Third, in Chapter 10 we explicitly compared the excess entropy to the specific heat, correlation length, nearest-neighbor correlation function, and the ferromagnetic structure factor. We saw that these statistical mechanical functions behave similarly, but not identically to **E**. More importantly, none of these functions has a numerical value that can be directly interpreted as memory, as **E** can be. In short, then, our comparison of different approaches to structure has shown that information theory and computational mechanics capture important properties of a system that statistical mechanics misses.

In Part III we presented a critical analysis of a statistical complexity measure introduced in Ref. [106] and discussed further in Ref. [2]. In so doing, we gave a brief historical review of a variety of statistical complexity measures and argued that it is essential that a statistical complexity measure have a clear, unambiguous interpretation.

Finally, in Part IV we explored extensions of our approach to more than one dimension. In Chapter 12 we gave a brief sketch of some of the extant work on quantifying patterns in more than one dimension. We also introduced the two-dimensional Ising model and reviewed the modern theory of critical phenomena. An exact, analytic result for the nearest-neighbor, two-spin mutual information for the two-dimensional Ising model was given in Chapter 13. We consider this result to be, at best, quasi two-dimensional in character, since the two-spin mutual information, a function of the distribution of only two variables, certainly fails to fully account for the spatial structure and memory shared across the two-dimensional lattice.

In Chapter 14 we presented several ways to extend the definition of the excess entropy so that it may be applied to two-dimensional systems. We also introduced the path ϵ-machine, one possible procedure for adapting computational mechanics to apply to multi-dimensional configurations. We calculated the path ϵ-machine and the different forms of the excess entropy for several simple two-dimensional patterns.

15.2 Concluding Remarks

15.2.1 Measure versus Support

Several ancillary observations, based on the foregoing results, are now in order. First, we have seen that for one-dimensional spin systems the number of causal states and their connectivity typically does not change as spin system parameters are varied. What does change, however, are the probabilities of the causal states and their transitions. In contrast, for deterministic dynamical systems it is typically the number of causal states that change as the system parameters are varied [44, 45]. Thus, it is our belief that "topological" measures of structure or complexity such as those of Refs. [6] and [103]—i.e., those measures that account for configurations only in terms of whether they are allowed or disallowed, and so ignore their probabilities—will not adequately capture important structural changes in statistical mechanical systems.

15.2.2 Limitations of Universal Turing Machine-Based Approaches to Structural Complexity

Second, approaches to structural complexity, such as those of Refs. [4], [13], [61], and [88], that are based on the Kolmogorov-Chaitin (KC) complexity, strike us as being of little use for addressing the questions of pattern and organization posed here. Our concerns about these KC complexity-based approaches are three-fold.

First, by adopting a UTM, the most powerful discrete computational model, one loses the ability to distinguish between systems that can be described by different computational models less powerful than a UTM [33, 44].

Second, and perhaps more importantly, the KC complexity is uncomputable: there exists no general algorithm for the calculation of the KC complexity. Thus, approaches focusing on KC complexity, including logical depth [13], tend to be non-constructive. In contrast, in the mathematical domain there are broad classes of processes for which the excess entropy and ϵ-machines can be determined [33, 38, 154]. In the empirical domain, moreover, there exist algorithms for estimating the excess entropy and determining an ϵ-machine [45, 37, 69]. The computational complexity of these algorithms is determined by the class of processes analyzed. Indeed, for the one-dimensional spin systems studied here, we gave closed-form expressions for various complexity measures of interest.

Third, KC complexity-based approaches inherit a fundamental relativity—a relativity that is built into how regularity and structure are accounted for and that derives from the UTM's lack of uniqueness and minimality. Computational mechanics takes a completely different approach and makes a specific commitment to causal states and ϵ-machines as a fundamental representation for the intrinsic computation embedded in a process. It also associates this computation, via the algebraic structure of ϵ-machines, with a system's internal organization and the patterns the system produces.

Thus, given the problems arising from KC complexity being based on UTMs, it seems to us that these approaches to structure and pattern will continue to find few empirical applications. Significant supplemental assumptions would have to be introduced to make these approaches viable. In contrast, due to its specificity of representation, computational mechanics is testable and its hypotheses—e.g., linking pattern, organization, and computation—are refutable.

15.3 Open Questions

We conclude by discussing some open questions and possible directions for future work. It remains an interesting open question as to how the excess entropy $\mathbf{E}$ characterizes quasiperiodic or more general $h_\mu = 0$ aperiodic configurations. Unfortunately, the simple spin systems analyzed in Part II are not rich enough to address this question. Some significant progress has been made in calculating the finite-L entropy density $h_\mu(L)$ for some aperiodic sequences [15, 58, 151]. One can deduce from this work that the excess entropy diverges for these systems. However, the nature of this divergence has not been fully examined. In Refs. [44] and [45], ϵ-machines been determined for some aperiodic systems. Nevertheless, there remains some work to be done in this area.

As discussed at length in Chapters 12 and 14, another crucial set of issues concerns extending the information theoretic and computational mechanical approaches to more than one spatial dimension. As mentioned above there has been some preliminary work in these areas, (e.g., Refs. [64, 69, 71, 91, 100, 103, 167]), but much remains to be done. In our view, a careful, genuinely two-dimensional computational mechanical treatment of a non-trivial two-dimensional system is still lacking.

Finally, there is a need to develop better techniques for estimating the excess entropy and reconstructing ϵ-machines in experimental settings. At present, there is no complete theory of statistical error estimation for inferring ϵ-machines from finite data. Some the the difficulties inherent in this task were discussed in Sec. 5.3. Nevertheless, some promising work has been done in this direction [37, 69].

In summary, there are two, broad open questions: How can an information theoretic measure of memory and, more importantly, computational mechanics, be extended to apply to multi-dimensional configurations? And how can ϵ-machines be optimally estimated from finite data? We believe that answering these questions is an essential step toward understanding how nature organizes and how its emergent structures take on functionality.

15.4 A New Walk is a New Walk

Where does the analysis and discussion of these many pages leave us? What might the future hold for computational mechanics? Where might computational mechanics find fruitful application? We believe that computational mechanics provides a powerful, broadly-applicable set of tools for discovering organization and structure in natural phenomena. We are confident that computational mechanics is an appropriate mathematical language to use when discussing how nature organizes and forms patterns. However, this optimism and excitement must be tempered by the realization that nature is complicated, complex, and diverse.

There has been much recent interest in the study of "complex systems"—see, e.g., Refs. [10, 114] as well as several popular books [59, 79, 93]. In these contexts the phrase "complex systems" typically is understood to refer to some or all of the following: ecosystems, economies, evolutionary systems such as population genetics and genetic algorithms, social insects and other animal groups, human languages, protein folding, nervous systems and the brain, and even human societies. Will this research lead to a "theory of complex systems"? Will we find "laws of complexity" as suggested in Ref. [79]? Is it the case "that the dynamics of complex systems are founded on universal principles" [10, p. xi]?

It is our belief that a "theory of complex systems" —if such a thing is found at all—will be very different than many existing physical theories. Mendeleev's periodic table successfully explained many chemical phenomena that were previously seen as unrelated; the periodic table was subsequently explained by quantum theory. Maxwell showed that electricity and magnetism were different manifestations of the same phenomena. And the standard model of particle physics succeeded in unifying phenomena previously viewed as unrelated.

We suspect that this trend toward unification of phenomena and explanation by appeal to uniform principles will not continue as the scope of problems considered widens. The disparate phenomena that fall under the umbrella of "complex systems" most likely will not be explained by a simple, uniform law.

On this subject Chomsky remarked, [28, p. 34]:

> Collapse of the traditional uniformity hypothesis should not come as a surprise. We find nothing like it in the study of other complex systems: the visual cortex, the kidney, the circulatory system, and others. Each of these "organs of the body" has its properties. They fall together, presumably, at the level of cellular biology, but no "organ theory" deals with the properties of organs in general.

To continue with Chomsky's analogy, there is an "organ theory" but in different sense than the word theory is often used in physics. There is not a uniform principle of organs, but there does exist a set of tools developed to probe the structure of organs: X-rays, magnetic resonance imaging, positron emission tomography, etc. An "organ theory," to the extent that such a theory exists, is a theory that concerns methods.

Likewise, we feel that a useful "complex systems theory" will primarily be concerned with developing theoretical and experimental tools for inquiring about the structures, patterns, and degrees of organization exhibited by complex systems. We have presented in this dissertation one such set of tools: the use of computational mechanics to determine a system's intrinsic computational abilities.

We close by turning again to the words of A. R. Ammons. We began this dissertation with a passage from the beginning of Ammons' poem *Corson's Inlet* [1]. In the poem Ammons records his thoughts while observing the ecology of the inlet on the New Jersey coast for which the poem is named. Ammons sees regularities and chaos combining to produce patterns and "disorderly orders" in a range of phenomena: a flock of swallows, patches of bayberries, the changing shapes of the dunes. He concludes the poem by resisting the urge to draw final conclusions, to bundle his thoughts into a simple package. In so doing, he cautions us against describing complex systems in too simple a way. Ammons is mindful of and comfortable with the complexities and uncertainties of nature [1]:

```
   I see narrow orders, limited tightness, but will
not run to that easy victory:
        still around the looser, wider forces work:
        I will try
    to fasten into order enlarging grasps of disorder, widening
scope, but enjoying the freedom that
Scope eludes my grasp, that there is no finality of vision,
that I have perceived nothing complete,
        that tomorrow a new walk is a new walk.
```

Part VI

Appendices

Appendix A

Summary of Information Theoretic Definitions

A.1 Notation

Let X and Y be random variables. The variable X may take on the values $x \in \mathcal{X}$. Here $\mathcal{X}$ is the (finite) set of all possible values for X and is referred to as the *alphabet* of X. Likewise, $Y = y \in \mathcal{Y}$.

The probability that X takes on the particular value x is written $\Pr(X = x)$, or just $\Pr(x)$. We may also form joint and conditional probabilities. The probability that $X = x$ and $Y = y$ is written $\Pr(X = x, Y = y)$, or $\Pr(x, y)$ and is referred to as a joint probability. The conditional probability that $X = x$ given $Y = y$ is written $\Pr(X = x|Y = y)$ or simply $\Pr(x|y)$.

Let $\overset{\leftrightarrow}{S}$ be a bi-infinite chain of random variables:

$$\overset{\leftrightarrow}{S} \equiv \ldots S_{-2} S_{-1} S_0 S_1 \ldots \ , \tag{A.1}$$

where the S_i's that range over a finite set $\mathcal{A}$.

We divide the chain into two semi-infinite halves by choosing a site i as the dividing point. Denote the left half by

$$\overset{\leftarrow}{S}_i \equiv \ldots S_{i-3} S_{i-2} S_{i-1} \tag{A.2}$$

and the right half by

$$\vec{S}_i \equiv S_i S_{i+1} S_{i+2} S_{i+3} \ldots \ . \tag{A.3}$$

We will assume that a spin system is described by a spatial shift-invariant measure μ that induces a family of distributions.

Let $\Pr(s_i)$ denote the probability that the i^{th} random variable S_i takes on the particular value $s_i \in \mathcal{A}$ and $\Pr(s_{i+1}, \ldots, s_{i+L})$ the joint probability over blocks of L consecutive spins. Assuming spatial translation symmetry, $\Pr(s_{i+1}, \ldots, s_{i+L}) = \Pr(s_1, \ldots, s_L)$. We denote a block of L consecutive variables by $S^L \equiv S_1 \ldots S_L$.

A.2 Fundamental Definitions

The *Shannon entropy* of the random variable X by:

$$H[X] \equiv -\sum_{x \in \mathcal{X}} \Pr(x) \log_2 (\Pr(x) \ . \tag{A.4}$$

The *joint entropy* of X and Y is defined:

$$H[X,Y] \equiv -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \Pr(x,y) \log_2 \Pr(x,y) \ . \tag{A.5}$$

The *conditional entropy* is defined by:

$$H[X|Y] \equiv -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \Pr(x,y) \log_2 \Pr(x|y) \ . \tag{A.6}$$

The Shannon entropy obeys the following *chain rule:*

$$H[X,Y] \ = \ H[X] + H[Y|X] \ . \tag{A.7}$$

The *mutual information* $I[X;Y]$ between X and Y:

$$I[X;Y] \ \equiv \ \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \Pr(x,y) \log_2 \frac{\Pr(x,y)}{\Pr(x)\Pr(y)} \ . \tag{A.8}$$

Properties of I:

$$
\begin{aligned}
I[X;Y] &= H[Y] - H[Y|X] && \text{(A.9)} \\
&= H[Y] + H[X] - H[X,Y] \; . && \text{(A.10)}
\end{aligned}
$$

The *information gain between* two distributions P and Q is defined by

$$
\mathcal{D}(P|Q) \equiv \sum_{x \in \mathcal{A} : \mathrm{Q}(x) > 0} \mathrm{P}(x) \log_2 [\mathrm{P}(x)/\mathrm{Q}(x)] \; , \tag{A.11}
$$

where we assume that if $\mathrm{Q}(x) = 0$, then $\mathrm{P}(x) = 0$.

A.3 Definitions of Information Theoretic Quantities Applied to a Stationary Stochastic Process

The entropy of a block of L consecutive variables is defined by:

$$
H(L) \equiv - \sum_{s^L \in \mathcal{A}^L} \Pr(s^L) \log_2 \Pr(s^L) \; . \tag{A.12}
$$

The *entropy density* h_μ is defined as follows. For a stationary process, the following three limits exist, and are equal [30]:

$$\begin{aligned} h_\mu &\equiv \lim_{L\to\infty} \frac{H(L)}{L} && \text{(A.13)} \\ &= \lim_{L\to\infty} \left[H(L+1) - H(L)\right] && \text{(A.14)} \\ &= \lim_{L\to\infty} H[S_L | S_1 \ldots S_{L-1}] \ . && \text{(A.15)} \end{aligned}$$

The finite-L approximation of the entropy density is defined by

$$h_\mu(L) \equiv H(L) - H(L-1), \ L = 1, 2, \ldots \ , \qquad \text{(A.16)}$$

The *excess entropy* $\mathbf{E}$ is defined by

$$\mathbf{E} \equiv \sum_{L=1}^{\infty} [h_\mu(L) - h_\mu] \ . \qquad \text{(A.17)}$$

For a stationary stochastic process, the following two expressions are equal to the definition of Eq. (A.17):

$$\begin{aligned} \mathbf{E} &= \lim_{L\to\infty} \left[H(L) - L h_\mu\right] && \text{(A.18)} \\ &= \lim_{L\to\infty} I[\overleftarrow{S}_i ; \overrightarrow{S}_i] \ . && \text{(A.19)} \end{aligned}$$

Appendix B

ϵ-Machine Entropy Density

We derive Eq. (5.34), an expression for the entropy density h_μ in terms of the probability of the causal states and their transitions. We begin with the expression for the entropy density, Eq. (4.4):

$$h_\mu = \lim_{L \to \infty} H[S_L | S_1 \cdots S_{L-2} S_{L-1}] \ . \tag{B.1}$$

Using the definition of the conditional entropy, Eq. (3.20), this may be rewritten as:

$$h_\mu = \lim_{L \to \infty} - \sum_{s_L, s^{L-1}} \Pr(s_L, s^{L-1}) \log_2 \Pr(s_L | s^{L-1}), \tag{B.2}$$

where s_L denotes the single spin variable at site L and s^{L-1} denotes the block of $L-1$ spins from sites 1 to $L-1$.

The causal states $\mathcal{S}_\alpha$ partition the set $\{s^{L-1}\}$; each s^{L-1} belongs to one and only one causal state equivalence class. (Cf. App. C.) As a result we may reexpress the sum as follows:

$$h_\mu = \lim_{L\to\infty} -\sum_{s_L,\alpha}\sum_{s^{L-1}\in\mathcal{S}_\alpha} \Pr(s_L, s^{L-1}) \log_2 \Pr(s_L|s^{L-1}) \ . \tag{B.3}$$

Causal states were defined in Eq. (5.7) such that two blocks of spins s_i^{L-1} and s_j^{L-1} belong to the same causal state if and only if $\Pr(\overrightarrow{s}\ |s_i^{L-1}) = \Pr(\overrightarrow{s}\ |s_j^{L-1})$, $\forall\ \overrightarrow{s}$. This observation enables us to perform the inner sum in Eq. (B.3). Each term in the argument of the logarithm is identical, since all the s^{L-1}'s belong to the same causal state. As a result, we can pull this term outside the sum:

$$h_\mu = \lim_{L\to\infty} -\sum_{s_L,\alpha}\Big[\log_2 \Pr(s_L|\mathcal{S}_\alpha) \sum_{s^{L-1}\in\mathcal{S}_\alpha} \Pr(s_L, s^{L-1})\Big] \ . \tag{B.4}$$

Note that since we are interested in the $L \to \infty$ limit, we need only concern ourselves with recurrent causal states. The inner summation has the effect of adding up the probabilities of all the s^{L-1}'s in the $\alpha^{\rm th}$ causal state:

$$\sum_{s^{L-1}\in\mathcal{S}_\alpha} \Pr(s_L, s^{L-1}) = \Pr(s_L, \mathcal{S}_\alpha) \ . \tag{B.5}$$

Inserting this into Eq. (B.4), we immediately obtain

$$h_\mu = -\sum_\alpha \sum_{s \in \mathcal{A}} \Pr(s, \mathcal{S}_\alpha) \log_2 \Pr(s|\mathcal{S}_\alpha) \ , \tag{B.6}$$

where $s \in \mathcal{A}$ are the spin values that can follow $\mathcal{S}_\alpha$. This result is Eq. (5.34). A little more explicitly we have

$$h_\mu = -\sum_\alpha \Pr(\mathcal{S}_\alpha) \sum_{s \in \mathcal{A}} \Pr(s|\mathcal{S}_\alpha) \log_2 \Pr(s|\mathcal{S}_\alpha) \ , \tag{B.7}$$

where $\Pr(\mathcal{S}_\alpha)$ is the left eigenvector of the stochastic connection matrix T, normalized in probability, and the second sum is seen to be the single-spin uncertainty at each causal state.

Appendix C

On the Equivalence Relation $\sim$ that Induces Causal States

Consider the set $\overleftarrow{\mathbf{S}}$ of all left-half configurations, of any length:

$$\overleftarrow{\mathbf{S}} = \{ \overleftarrow{s}^{L} = s_{L-1} \cdots s_{-1} : \; s_i \in \mathcal{A}, \; L = 0, 1, \ldots \} \; . \tag{C.1}$$

Recall that $\overleftarrow{s}^{0} = \lambda$, the empty string. It was claimed in Eq. (5.7) that

$$\overleftarrow{s}_i^{M} \sim \overleftarrow{s}_j^{L} \; \Leftrightarrow \; \Pr(\overrightarrow{s} \, | \overleftarrow{s}_i^{K}) = \Pr(\overrightarrow{s} \, | \overleftarrow{s}_j^{L}) \; , \tag{C.2}$$

for all semi-infinite $\overrightarrow{s} = s_0 s_1 s_2 \cdots$, where $M, L = 0, 1, 2, \ldots$, defined an equivalence relation $\sim$ over $\overleftarrow{\mathbf{S}}$. Here we show that this is indeed the case by reviewing the basic properties of relations, equivalence classes, and partitions. (The proof details are straightforward and are not included. See Ref. [97].) We will drop the length variables M and L and denote by $\overleftarrow{s}, \overleftarrow{s}', \overleftarrow{s}'' \in \overleftarrow{\mathbf{S}}$ members of any length in the set of Eq. (C.1).

First, $\sim$ is a relation on $\overleftarrow{\mathbf{S}}$ since we can represent it as a subset of the Cartesian product

$$\overleftarrow{\mathbf{S}} \times \overleftarrow{\mathbf{S}} = \{(\overleftarrow{s}, \overleftarrow{s}') : \overleftarrow{s}, \overleftarrow{s}' \in \overleftarrow{\mathbf{S}}\} . \tag{C.3}$$

Second, the relation $\sim$ is an equivalence relation on $\overleftarrow{\mathbf{S}}$ since it is

1. reflexive: $\overleftarrow{s} \sim \overleftarrow{s}$, $\forall \; \overleftarrow{s} \in \overleftarrow{\mathbf{S}}$;
2. symmetric: $\overleftarrow{s} \sim \overleftarrow{s}' \Rightarrow \overleftarrow{s}' \sim \overleftarrow{s}$; and
3. transitive: $\overleftarrow{s} \sim \overleftarrow{s}'$ and $\overleftarrow{s}' \sim \overleftarrow{s}'' \Rightarrow \overleftarrow{s} \sim \overleftarrow{s}''$.

Third, if $\overleftarrow{s} \in \overleftarrow{\mathbf{S}}$, the equivalence class of $\overleftarrow{s}$ is

$$[\overleftarrow{s}] = \{\overleftarrow{s}' \in \overleftarrow{\mathbf{S}} : \overleftarrow{s}' \sim \overleftarrow{s}\} . \tag{C.4}$$

The set of all equivalence classes in $\overleftarrow{\mathbf{S}}$ is denoted $\overleftarrow{\mathbf{S}} / \sim$ and is called the *factor set* of $\overleftarrow{\mathbf{S}}$ with respect to $\sim$. In Sec. 5.3 we called the individual equivalence classes causal states $\mathcal{S}_\alpha$ and denoted the set of causal states $\mathcal{S} = \{\mathcal{S}_\alpha : \alpha = 0, 1, \ldots, k-1\}$. That is, $\mathcal{S} = \overleftarrow{\mathbf{S}} / \sim$. (We noted in the main text that $k = |\mathcal{S}|$ may or may not be infinite.)

Finally, we list several basic properties of the causal state equivalence classes.

1. $\bigcup_{\overleftarrow{s} \in \overleftarrow{\mathbf{S}}} [\overleftarrow{s}] = \overleftarrow{\mathbf{S}}$.

2. $\bigcup_{\alpha=0}^{k-1} \mathcal{S}_\alpha = \overleftarrow{\mathbf{S}}$.

3. $[\overleftarrow{s}] = [\overleftarrow{s}'] \Leftrightarrow \overleftarrow{s} \sim \overleftarrow{s}'$.

4. If $\overleftarrow{s}, \overleftarrow{s}' \in \overleftarrow{\mathbf{S}}$, either

   (a) $[\overleftarrow{s}] \bigcap [\overleftarrow{s}'] = \emptyset$ or

   (b) $[\overleftarrow{s}] = [\overleftarrow{s}']$.

5. The causal states $\mathcal{S}$ are a partition of $\overleftarrow{\mathbf{S}}$. That is,

   (a) $\mathcal{S}_\alpha \neq \emptyset$ for each α,

   (b) $\bigcup_{\alpha=0}^{k-1} \mathcal{S}_\alpha = \overleftarrow{\mathbf{S}}$, and

   (c) $\mathcal{S}_\alpha \cap \mathcal{S}_\beta = \emptyset$ for all $\alpha \neq \beta$.

We denote the start state with $\mathcal{S}_0$. The start state is the causal state associated with $\overleftarrow{s} = \lambda$. That is, $\mathcal{S}_0 = [\lambda]$.

Each causal state equivalence class $\mathcal{S}_\alpha$ thus has several structures attached:

1. The index α—the state's "name".

2. The set of left-half configurations, of various lengths, comprising the equivalence class: $[\overleftarrow{s}] = \{\overleftarrow{s} \in \mathcal{S}_\alpha\}$.

3. A conditional distribution over right-half configurations: $\Pr(\overrightarrow{s} \mid \overleftarrow{s}), \overleftarrow{s} \in \mathcal{S}_\alpha$. We denote this distribution more concisely by $\Pr(\overrightarrow{s} | \mathcal{S}_\alpha)$.

As noted in the text, the definitions and properties of the causal states obtained by scanning in the opposite direction, i. e., the causal states $\overrightarrow{\mathbf{S}} / \sim$, follow similarly. For general processes, $\overleftarrow{\mathbf{S}} / \sim \neq \overrightarrow{\mathbf{S}} / \sim$.

For completeness, we note that this construction of causal states is analogous to Nerode equivalence, used to determine the minimal number of states for a finite-state machine representation of a regular language [73, 153]. It is also somewhat similar to the states estimated in Rissanen's "context" algorithm [134]. Despite these similarities, there are important differences. With Nerode equivalence infinite strings and probability measures over them are not considered. For a random source—for which there is a single causal state—the context algorithm estimates a number of states that diverges (logarithmically) with the length of the data stream.

Appendix D

Transient Structure from the Recurrent ϵ-Machine

In this appendix we show how transient states can be constructed from the recurrent portion of an ϵ-machine. The latter, denoted $\mathbf{M}^{(R)}$, consists of the recurrent causal states $\mathcal{S}^{(R)}$ and their transitions $T^{(s)}_{\alpha\beta}$, where the indices α and β run over only the recurrent causal states. That the transient states can be constructed in this manner is a direct consequence of the equivalence relation $\sim$, defined in Eq. (5.7), that induces the causal states.

The basic idea of the construction procedure, detailed below, is as follows. First, we assume the generating process has been operating sufficiently long so that it is in equilibrium in the sense that it is being controlled by its recurrent causal states. We also assume we have a model of the process, namely $\mathbf{M}^{(R)}$, in hand. Then we begin making measurements—reading in spin values from a configuration—$s_0 s_1 s_2 \ldots$. With each measurement, we ask: In which recurrent causal state is the process? Initially, while making measurements and, of course, even before making the first, we are uncertain about which recurrent causal state the process is in. Thus, we describe our state of knowledge by a distribution over $\mathcal{S}^{(R)}$, denoted $\Pr(\mathcal{S}^{(R)}|s_0 s_1 s_2 \cdots s_{L-1})$. As we observe successive spins, this distribution changes. The structure of the machine $\mathbf{M}^{(R)}$ determines the change when we observe an individual spin. Presumably, with a sufficient number of spin measurements we become *synchronized* with the process: That is, we know with certainty in which recurrent state the process is. This procedure of tracking how our state of knowledge $\Pr(\mathcal{S}^{(R)}|s_0 s_1 s_2 \cdots s_{L-1})$ refines and focuses on smaller and smaller subsets of $\mathcal{S}^{(R)}$ determines the transient causal states. From here on we will drop the superscript (R) on $\mathcal{S}$, when no ambiguity arises.

Since we assume the process is in equilibrium, we take the initial probability distribution—that associated with having made no measurements—to be the asymptotic distribution over the recurrent causal states, which is given by Eq. (5.17). This initial distribution is denoted $\Pr(\mathcal{S}|\lambda)$ to indicate that it is the distribution before any spins have been observed. (Recall that λ denotes the empty string.) After observing spin s_0 our state of knowledge about the recurrent causal state of the process has improved and is now described by the distribution $\Pr(\mathcal{S}|s_0)$. Upon observing the next spin s_1, our state of knowledge becomes $\Pr(\mathcal{S}|s_0 s_1)$.

We may associate the recurrent causal state distributions with the causal states

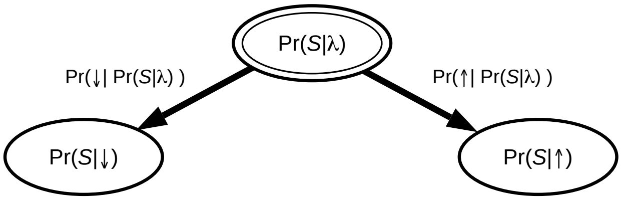


Figure D.1: First steps in performing transient state construction for the case of a spin-1/2 system where $s_i \in \{\uparrow, \downarrow\}$. The double oval indicates the start node n_0.

themselves. For example, a distribution that specifies a recurrent causal state with certainty—i. e., probability 1 of being in that state—can be taken to be that recurrent state. The transient states, in contrast, are those distributions in which the recurrent causal state is not known with certainty.

The procedure through which the transient states are deduced proceeds by constructing a tree $\mathcal{T} = \{\mathcal{N}, \mathcal{L}\}$ consisting of a set of nodes $\mathcal{N}$ and a set of links $\mathcal{L}$ connecting the nodes. A node $n \in \mathcal{N}$ in the tree corresponds to a recurrent causal state distribution $\Pr(\mathcal{S}|s_0 s_1 \cdots s_{L-1})$. A link $\ell_s \in \mathcal{L}$ corresponds to a transition between successive causal state distributions that occurs upon observing a particular spin value s. We call a tree node for which all the outgoing links have yet to be determined a *leaf*. We denote the set of leaves by $\hat{\mathcal{N}}$.

The tree is constructed recursively via the following steps:

1. **Initialize:** Given a recurrent ϵ-machine $\mathbf{M}^{(R)} = \{\mathcal{S}^{(R)}, T^{(s)}_{\alpha\beta}\}$ determine the asymptotic probability of the recurrent causal states via Eq. (5.17). This distribution is the starting node for the tree $n_0 = \Pr(\mathcal{S}|\lambda)$ and is indicated in

Fig. D.1 by the node with the double oval. At this stage n_0 is a leaf, since we have not yet determined all the links (transitions) that leave it. Thus, $\mathcal{N} = \emptyset$ and $\hat{\mathcal{N}} = \{n_0\}$.

2. **Build Transient Tree:** While $\hat{\mathcal{N}}$ is nonempty:

   (a) **Determine Links:** For each leaf $n = \Pr(\mathcal{S}|s^L) \in \hat{\mathcal{N}}$ draw a link $\ell_s \in \mathcal{L}$ (an outgoing transition) for each spin value s. Label the link with the transition probability $\Pr(s|n)$ that starting in node n spin value s is seen:

   $$\Pr(s|n) = \sum_{\mathcal{S}' \in \mathcal{S}^{(R)}} \Pr(\mathcal{S}')\Pr(s|\mathcal{S}') \ . \tag{D.1}$$

   If the transition has zero probability, ignore ℓ_s.

   (b) **Form Node Distributions:** For each link ℓ_s determine the probability distribution $\Pr(\mathcal{S}|s^L s)$ to which it leads using:

   $$\Pr(\mathcal{S}|s^L s) = \frac{\sum_{\mathcal{S}' \in \mathcal{S}^{(R)}} \Pr(\mathcal{S}'|s^L)\Pr(\mathcal{S}|s, \mathcal{S}')}{\sum_{\mathcal{S}', \mathcal{S} \in \mathcal{S}^{(R)}} \Pr(\mathcal{S}'|s^L)\Pr(\mathcal{S}|s, \mathcal{S}')} \ . \tag{D.2}$$

   Note that the term in the denominator is simply a normalization. The quantity $\Pr(\mathcal{S}|s^L s)$ gives the updated distribution over recurrent causal states after having observed the particular spin sequence $s^L s$. Recall that, since the ϵ-machine is deterministic, $\Pr(\mathcal{S}|s, \mathcal{S}') = 1$ if the transition is allowed and 0 otherwise.

(c) **Merge Duplicate Nodes:** Now consider, in turn, the probability distributions just formed: $n = \Pr(\mathcal{S}|s^L s)$. Is n identical to another node distribution $n' \in \mathcal{N}$?

i. If yes, then connect ℓ_s to node n'.

ii. If no, add n to the set $\hat{\mathcal{N}}$ of tree leaves.

3. **Minimize**: The resulting machine has a recurrent part that is identical to $\mathbf{M}^{(R)}$, but it may not be minimal. Merge nodes pairwise under the equivalence relation $\sim$ of Eq. (5.7). The result is the complete ϵ-machine, with all transient and recurrent states.

We illustrate the above procedure by considering a period-4 process. The recurrent portion of its ϵ-machine $\mathbf{M}^{(R)}$ is shown in Fig. D.2. The result of the first several steps of the transient state construction procedure is illustrated in Fig. D.3. The asymptotic probability of each recurrent causal state is 1/4. Thus, as per step 1 above, the start node is labeled $\Pr(\mathcal{S}|\lambda) = (1/4, 1/4, 1/4, 1/4)$; it is shown as the double oval of Fig. D.3.

From the start node, two transitions are possible—one for $s = \uparrow$ and one for $s = \downarrow$. In the figure they are labeled with the transition probabilities as given by

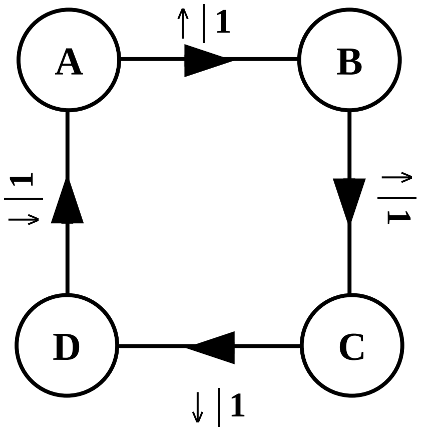


Figure D.2: Recurrent portion of the period-4 ϵ-machine.

Eq. (D.1). For example, from n_0 the probability of seeing $s = \uparrow$ is given by:

$$\begin{aligned} \Pr(\uparrow | n_0) &= \Pr(\mathbf{A})\Pr(\uparrow | \mathbf{A}) + \Pr(\mathbf{B})\Pr(\uparrow | \mathbf{B}) + \\ & \quad \Pr(\mathbf{C})\Pr(\uparrow | \mathbf{C}) + \Pr(\mathbf{D})\Pr(\uparrow | \mathbf{D}) \qquad \text{(D.3)} \\ &= (1/4)(1) + (1/4)(1) + \\ & \quad (1/4)(0) + (1/4)(0) = 1/2 \; . \qquad \text{(D.4)} \end{aligned}$$

The leaves (causal state distributions) to which the links (transitions) lead are determined by Eq. (D.2). For example, consider the $\downarrow$ transition from n_0. The normalization factor, the denominator in Eq. (D.2), is 1/2 and the probability of being in causal state $\mathbf{A}$ is given by:

$$\Pr(\mathcal{S} = \mathbf{A} | \downarrow) = 2 \sum_{\mathcal{S}' \in \boldsymbol{\mathcal{S}}^{(R)}} \Pr(\mathcal{S}')\Pr(\mathbf{A} | \downarrow, \mathcal{S}') \qquad \text{(D.5)}$$

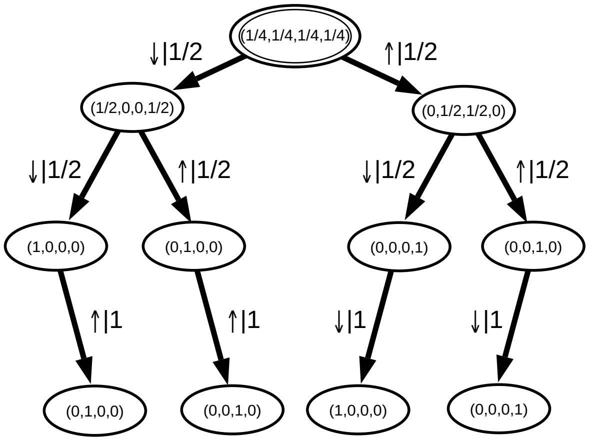


Figure D.3: The tree $\mathcal{T}$ part of the way through transient state construction for the period-4 process. To construct this tree spin blocks up to length 3 were examined.

$$\begin{aligned} &= 2\Big[\Pr(\mathbf{A})\Pr(\mathbf{A}|\downarrow,\mathbf{A}) + \Pr(\mathbf{B})\Pr(\mathbf{A}|\downarrow,\mathbf{B}) \\ &+ \Pr(\mathbf{C})\Pr(\mathbf{A}|\downarrow,\mathbf{C}) + \Pr(\mathbf{D})\Pr(\mathbf{A}|\downarrow,\mathbf{D})\Big] \qquad \text{(D.6)} \\ &= 2\Big[(1/4)(0) + (1/4)(0) + \\ &\quad (1/4)(0) + (1/4)(1)\Big] \ = \ 1/2\ . \qquad \text{(D.7)} \end{aligned}$$

Similarly, one finds that $\Pr(\mathbf{D}|\downarrow) = 1/2$ and $\Pr(\mathbf{B}|\downarrow) = \Pr(\mathbf{C}|\downarrow) = 0$. As a result, this node is associated with the distribution $\Pr(\mathcal{S}|\downarrow) = (1/2, 0, 0, 1/2)$.

The process of adding links and nodes to the tree $\mathcal{T}$ is shown repeated up to length-3 spin blocks in Fig. D.3. At the last level of the tree, transitions that occur with probability 0 are not drawn. At this point, notice that the leaves at the bottom level have already appeared as nodes above in the tree. For example, the lower left

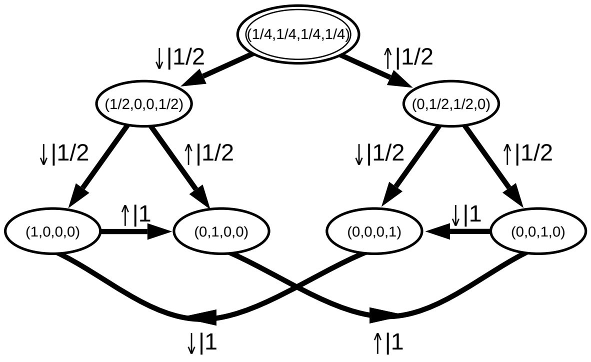


Figure D.4: The end result of transient state construction for the period-4 process after duplicate leaves are removed.

leaf is identical to the node that is second from the left, one level above. Thus, as per step 2(c)i, the link pointing to the leaf node is directed to the pre-existing node. This reconnection step is repeated until there are no leaves left. The result is illustrated in Fig. D.4.

The final result of the procedure, shown in Fig. D.4, is the complete ϵ-machine with all recurrent and transient states. The recurrent states—distributions over recurrent causal states that determine individual causal states—are the four nodes along the bottom of the figure. They are identical to those we started out with in Fig. D.2. The three transient states that have been constructed are the three nodes whose distributions correspond to some uncertainty about in which recurrent causal state the process is. In this way, the structure of the transient portion of an ϵ-machine shows how successive measurements refine an observer's knowledge about in which causal state a process is.

Appendix E

D vanishes exponentially fast for Regular Markov Chains

We will show that D goes to zero exponentially fast with increasing system size for a Markov chain governed by a regular transition matrix $T_{ab} = \Pr(b|a)$, where $0 \leq T_{ab} \leq 1$ and there exists a $K < \infty$ such that $(T^K)_{ab} > 0$ for all a and b. Since the conditional probabilities are normalized, the matrix T is stochastic: $\sum_{b=1}^{k} T_{ab} = 1$.

The probability of a block of L consecutive variables taking on the values $x_1, x_2, \ldots, x_L$ is given by

$$\Pr(x_1, x_2, \ldots, x_L) = p_{x_1} T_{x_1 x_2} T_{x_2 x_3} \cdots T_{x_{L-1} x_L} \ , \tag{E.1}$$

where p is the stationary distribution of a single variable, as given by the left eigenvector of T with eigenvalue 1. The eigenvector p is chosen so as to be normalized in probability, $\sum_{a=1}^{k} p_a = 1$.

Let $\tilde{T}$ be a matrix whose components $\tilde{T}_{ab}$ are given by $(T_{ab})^2$. Note that $\tilde{T} \neq T^2$. Similarly, let $\tilde{p}$ be a vector whose components $\tilde{p}_a$ are given by $(p_a)^2$. Eq. (11.6) indicates that we are interested in

$$\sum_{\{x^L\}} \Pr(x^L)^2 = \sum_{\{x^L\}} \tilde{p}_{x_1} \tilde{T}_{x_1 x_2} \tilde{T}_{x_2 x_3} \cdots \tilde{T}_{x_{L-1} x_L} . \tag{E.2}$$

The sum runs over all configurations of length L. The effect of the sum is to multiply the matrices together;

$$\sum_{\{x^L\}} \Pr(x^L)^2 \; = \sum_{x_1} \sum_{x_L} \tilde{p}_{x_1} (\tilde{T}^{L-1})_{x_1 x_L} \; . \tag{E.3}$$

We shall show that in the $L \to \infty$ limit the above expression goes to zero exponentially fast.

We begin by considering the vector $V \equiv \tilde{p}\tilde{T}^{L-1}$ and its L_∞ norm, $\|V\| \equiv \max\{|V_1|, |V_2|, \ldots\}$. Eq. (E.3) may be rewritten in terms of V,

$$\sum_{\{x^L\}} \Pr(x^L)^2 \; = \sum_{i=1}^{k} V_i . \tag{E.4}$$

Since V is finite dimensional and all elements are nonnegative, if $\|V\|$ goes to zero exponentially fast, $\sum_{\{x^L\}} \Pr(x^L)^2$ must also go to zero exponentially fast. To show the former we use some well-known properties of vector and matrix norms [20].

Consider the matrix norm induced by the L_∞ vector norm:

$$\|\tilde{T}\| \equiv \max_a \{ \sum_b \tilde{T}_{ab} \} \ . \tag{E.5}$$

Any matrix norm is compatible with its associated vector norm:

$$\|V\| = \|\tilde{p}\,\tilde{T}^{L-1}\| \leq \|\tilde{p}\|\,\|\tilde{T}^{L-1}\| \ . \tag{E.6}$$

Recall that the components of $\tilde{p}$ are the square of the components of the stationary probability p of the Markov process. Except for the trivial case in which there is only one symbol in our chain and T is a one-by-one matrix, the maximum component of p is less than one. Thus, $0 < \|\tilde{p}\| < 1$ and we have

$$\|V\| \leq \|\tilde{p}\|\,\|\tilde{T}^{L-1}\| \leq \|\tilde{T}^{L-1}\| \ . \tag{E.7}$$

By assumption, there exists a K such that $0 < (T^K)_{ab} < 1$ for all a and b. Each element of T^K is a sum of terms that are products of T's elements. Likewise, each element of $\tilde{T}^K$ is a sum of terms that are products of $\tilde{T}$'s elements. However, since $\tilde{T}_{ab} = (T^2)_{ab}$ and $0 \leq T_{ab} \leq 1$, it follows that each component of $\tilde{T}^K$ is strictly less than the corresponding component of T^K. The product of stochastic matrices is itself a stochastic matrix, so $\sum_{b=1}^{k} (T^K)_{ab} = 1$. Thus, since each component of $\tilde{T}$ is less than the corresponding component of T, $\sum_{b=1}^{k}(\tilde{T}^K)_{ab} < 1$. As a result, $||\tilde{T}|| < 1$.

Now, by the consistency condition obeyed by all matrix norms, $||\tilde{T}^2|| \leq ||\tilde{T}|| \, ||\tilde{T}||$. Rewriting eq. (E.7), we have:

$$||V|| \leq ||\tilde{T}^{L-1}|| = ||\tilde{T}^{K(L-1)/K}|| \, . \tag{E.8}$$

In the $L/K \to \infty$ limit—equivalent to the $L \to \infty$ limit since K is finite—we then have by the consistency condition:

$$||V|| \leq ||\tilde{T}^K||^{L-1} \, . \tag{E.9}$$

Since $||\tilde{T}^K|| < 1$ we see that $||V||$ is bounded above by a function that decreases exponentially in L. Hence $||V||$ itself also decreases exponentially in L.

Appendix F

Details of some Probability Manipulations for 2D Ising Two-spin Distributions

The goal of this appendix is to determine the joint distribution over nearest-neighbor spins as a function of the magnetization m and the expectation value of the product of the two spins, $\Gamma_u(1,0)$ This will be quite straightforward, although somewhat messy.

First, we note that the magnetization is just the expectation value of a single spin S. (Since the system is translationally invariant, it doesn't matter which spin.) Thus,

$$m \equiv \langle S \rangle \equiv (1)\mathrm{Pr}(1) + (-1)\mathrm{Pr}(-1) \; , \tag{F.1}$$

where $\mathrm{Pr}(1) \equiv \mathrm{Pr}(\mathrm{s}=1)$. As the probabilities must be normalized,

$$1 = \mathrm{Pr}(1) + \mathrm{Pr}(-1)\ . \tag{F.2}$$

Adding and subtracting the above two equations yields:

$$\mathrm{Pr}(\pm 1) = \frac{1 \pm \mathrm{m}}{2}\ . \tag{F.3}$$

We now seek the joint distribution, $\mathrm{Pr}(s_{i,j}, s_{i+1,j})$. We may factor this into the product of a conditional and a marginal distribution:

$$\mathrm{Pr}(s_{i,j}, s_{i+1,j}) = \mathrm{Pr}(s_{i,j}|s_{i+1,j})\mathrm{Pr}(s_{i+1,j})\ . \tag{F.4}$$

As the marginal distribution has already been determined in Eq. (F.3), our goal will be to obtain expressions for the conditional probabilities.

Configurations posses a rotational symmetry; it is unchanged by rotations of 90 degrees. Hence,

$$\mathrm{Pr}(s_{i,j}, s_{i+1,j}) = \mathrm{Pr}(s_{i+1,j}, s_{i,j})\ . \tag{F.5}$$

Note that the probabilities do not necessarily possess a spin flip symmetry; this symmetry is spontaneously broken below T_c.

As a particular case of Eq. (F.5), consider:

$$\mathrm{Pr}(1,-1) \;=\; \mathrm{Pr}(-1,1)\;. \tag{F.6}$$

We may factor this joint distribution into the product of a marginal and a conditional distribution two different ways to obtain:

$$\mathrm{Pr}(1|-1)\mathrm{Pr}(-1) \;=\; \mathrm{Pr}(-1|1)\mathrm{Pr}(1)\;. \tag{F.7}$$

Rearranging this, and using Eq. (F.3), we obtain:

$$\mathrm{Pr}(1|-1) \;=\; \frac{1+\mathrm{m}}{1-\mathrm{m}}\,\mathrm{Pr}(-1|1)\;. \tag{F.8}$$

We now have an equation relating two of the four conditional probability distributions we seek. To proceed further, consider the expectation value of the product of two neighboring spins:

$$\begin{aligned}\Gamma_u(1,0) \;\equiv \langle\, S_{i,j}S_{i+1,j}\,\rangle \equiv\;& (1)\mathrm{Pr}(1,1) + (-1)\mathrm{Pr}(-1,1) + \\ & (-1)\mathrm{Pr}(1,-1) + (1)\mathrm{Pr}(-1,-1)\;. \end{aligned} \tag{F.9}$$

Rewriting the joint probabilities, this becomes:

$$\begin{aligned}\Gamma_u(1,0) \;&=\; \mathrm{Pr}(1|1)\mathrm{Pr}(1) \;-\; \mathrm{Pr}(-1|1)\mathrm{Pr}(1) \;- \\ &\quad \mathrm{Pr}(1|-1)\mathrm{Pr}(-1) \;+\; \mathrm{Pr}(-1|-1)\mathrm{Pr}(-1)\;. \end{aligned} \tag{F.10}$$

Plugging Eq. (F.8) into the above equation, we obtain:

$$\begin{aligned}\Gamma_u(1,0) \;&=\; \mathrm{Pr}(1|1)\mathrm{Pr}(1) + \mathrm{Pr}(-1|-1)\mathrm{Pr}(-1) \;- \\ &\quad \mathrm{Pr}(-1|1)\left[\mathrm{Pr}(1) + \eta\mathrm{Pr}(-1)\right]\;, \end{aligned} \tag{F.11}$$

where,

$$\eta \equiv \frac{1+m}{1-m}\;. \tag{F.12}$$

The conditional probabilities must be normalized. Thus,

$$\mathrm{Pr}(1|1) \;=\; 1 - \mathrm{Pr}(-1|1)\;. \tag{F.13}$$

Similarly,

$$\mathrm{Pr}(-1|-1) \;=\; 1 - \mathrm{Pr}(1|-1)\;. \tag{F.14}$$

Using Eq. (F.8), this may be rewritten as

$$\mathrm{Pr}(-1|1) \;=\; \eta^{-1}\left[1 - \mathrm{Pr}(-1|-1)\right]\;. \tag{F.15}$$

Combining Eqs. (F.11), (F.13), and (F.15) and simplifying, one finds:

$$\begin{aligned}\Gamma_u(1,0) \quad &= \quad \mathrm{Pr}(1) - \eta^{-1}\left[1 - \mathrm{Pr}(-1|-1)\right]\left\{2\mathrm{Pr}(1) + \eta\mathrm{Pr}(-1)\right\} \; + \\ &\qquad \mathrm{Pr}(-1|-1)\mathrm{Pr}(-1)\ . \end{aligned} \tag{F.16}$$

We solve for Pr(−1| − 1) using Eqs. (F.3) and (F.12), and obtain:

$$\mathrm{Pr}(-1|-1) \;=\; \frac{1 + \Gamma_{\mathrm{u}}(1,0) - \mathrm{m}}{2 - \mathrm{m}}\ . \tag{F.17}$$

Eq. (F.15) relates Pr(−1| − 1) to Pr(−1|1), from which it follows that:

$$\mathrm{Pr}(-1|1) \;=\; \frac{1 - \mathrm{m}}{1 + \mathrm{m}}\left[\frac{1 - \Gamma_{\mathrm{u}}(1,0)}{2 - \mathrm{m}}\right]\ . \tag{F.18}$$

Using eq. (F.13), I can then solve for Pr(1|1);

$$\mathrm{Pr}(1|1) \;=\; 1 \;-\; \frac{1 - \mathrm{m}}{1 + \mathrm{m}}\left[\frac{1 - \Gamma_{\mathrm{u}}(1,0)}{2 - \mathrm{m}}\right]\ . \tag{F.19}$$

Finally, we use Eq. (F.8) to solve for Pr(1| − 1), obtaining:

$$\mathrm{Pr}(1|-1) \;=\; \frac{1 - \Gamma_{\mathrm{u}}(1,0)}{2 - \mathrm{m}}\ . \tag{F.20}$$

Equations (F.17-F.20), together with Eq. (F.3) complete the task of this appendix.

Bibliography

[1] A. R. Ammons. Corsons Inlet. In *Selected Poems.* Cornell University Press, 1968. Reprinted in The Norton Anthology of Poetry, Revised Shorter Edition, W. W. Norton & Co., A. W. Allison H. Barrows, C. R. Blake, A. J. Carr A. M. Eastman and H. M. English, Jr., editors.

[2] C. Anteneodo and A.R. Plastino. Some features of the Lòpez-Ruiz–Mancini–Calbet statistical measure of complexity. *Physics Letters A*, 223:348–354, 1996.

[3] D. Arnold. Information-theoretic analysis of phase transitions. *Complex Systems*, 10:143–155, 1996.

[4] H. Atmanspacher, C. Räth, and G. Weidenmann. Statistics and meta-statistics in the concept of complexity. *Physica*, A243:819–829, 1997.

[5] R. Badii and A. Politi. *Complexity: Hierarchical structures and scaling in physics.* Cambridge University Press, Cambridge, 1997.

[6] R. Badii and A. Politi. Thermodynamics and complexity of cellular automata. *Phys. Rev. Lett.*, 78(3):444–447, 1997.

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

[8] P. Bak, C. Tang, and K. Weisenfield. Self-organized criticality: An explanation of 1/f noise. *Phys. Rev. Lett.*, 59(4):381–384, 1987.

[9] G. A. Baker. Markov-property Monte Carlo method: One-dimensional Ising model. *J. Stat. Phys.*, 72:621–640, 1993.

[10] Y. Bar-Yam. *Dynamics of Complex Systems.* Addison-Wesley, 1997.

[11] R. J. Baxter. *Exactly Solved Models in Statistical Mechanics.* Academic Press, 1982.

[12] C. Beck and F. Schlögl. *Thermodynamics of Chaotic Systems.* Cambridge University Press, 1993.

[13] C. H. Bennett. On the nature and origin of complexity in discrete, homogeneous locally-interacting systems. *Found. Phys.*, 16:585–592, 1986.

[14] C. H. Bennett. How to define complexity in physics, and why. In W. H. Zurek, editor, *Complexity, Entropy, and the Physics of Information*, volume VIII of *Santa Fe Institute Studies in the Sciences of Complexity*, pages 137–148. Addison-Wesley, 1990.

[15] V. Berthé. Conditional entropy of some automatic sequences. *Journal of Physics*, A27:7993–8006, 1994.

[16] W. Bialek, C. G. Callan, and S. P. Strong. Field theories for learning probability distributions. *Phys. Rev. Lett.*, 77:4693–4697, 1996.

[17] J. J. Binney, N. J. Dowrick, A. J. Fisher, and M. E. J. Newman. *The Theory of Critical Phenomena: An Introduction to the Renormalization Group.* Oxford Science Publications, 1992.

[18] D. Blackwell and L. Koopmans. On the identifiability problem for functions of Markov chains. *Ann. Math. Statist.*, 28:1011–1015, 1957.

[19] L. Boltzmann. *Lectures on gas theory.* University of California Press, Berkeley, 1964.

[20] R. Bronson. *Matrix Operations.* McGraw-Hill, 1989.

[21] J. G. Brookshear. *Theory of Computation: Formal Languages, Automata, and Complexity.* Benjamin/Cummings, 1989.

[22] A. A. Brudno. Entropy and the complexity of the trajectories of a dynamical system. *Trans. Moscow Math. Soc.*, 44:127, 1983.

[23] S. G. Brush. History of the Lenz-Ising model. *Rev. Mod. Phys.*, 39(1):883–893, 1967.

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

[25] G. Chaitin. On the length of programs for computing finite binary sequences. *J. ACM*, 13:145, 1966.

[26] G. Chaitin. *Information, Randomness and Incompleteness.* World Scientific, Singapore, 1987.

[27] D. Chandler. *Introduction to Modern Statistical Mechanics.* Oxford University Press, 1987.

[28] N. Chomsky. *Language and Thought*, volume Three of *The Frick Collection: Anshen Transdisciplinary Lectureships in Art, Science and the Philosophy of Culture*. Moyer Bell, 1993.

[29] S. N. Coppersmith, T. C. Jones, L. P. Kadanoff, A. Levine, J. P. McCarten, S. R. Nagel, S. C. Venkataramani, and X. Wu. Self-organized short-term memories. *Phys. Rev. Lett.*, 78(21):3983–3986, 1997.

[30] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley & Sons, Inc., 1991.

[31] R. T. Cox. Probability, frequency, and reasonable expectation. *Am. J. Phys.*, 14:1–13, 1946.

[32] J. P. Crutchfield. Semantics and thermodynamics. In M. Casdagli and S. Eubank, editors, *Nonlinear Modeling and Forecasting*, volume XII of *Santa Fe Institute Studies in the Sciences of Complexity*, pages 317–359, Reading, Massachusetts, 1992. Addison-Wesley.

[33] J. P. Crutchfield. The calculi of emergence: Computation, dynamics, and induction. *Physica D*, 75:11–54, 1994.

[34] J. P. Crutchfield. Critical computation, phase transitions, and hierarchical learning. In M. Yamaguti, editor, *Towards the Harnessing of Chaos*, pages 29–46, Amsterdam, 1994. Elsevier Science.

[35] J. P. Crutchfield. Is anything ever new? Considering emergence. In G. Cowan, D. Pines, and D. Melzner, editors, *Complexity: Metaphors, Models, and Reality*, volume XIX of *Santa Fe Institute Studies in the Sciences of Complexity*, pages 479–497, Reading, MA, 1994. Addison-Wesley.

[36] J. P. Crutchfield, 1998. Personal Communication.

[37] J. P. Crutchfield and C. Douglas. 1998. In Preparation.

[38] J. P. Crutchfield and D. P. Feldman. Statistical complexity of simple one-dimensional spin systems. *Phys. Rev. E*, 55(2):1239R–1243R, 1997.

[39] J. P. Crutchfield and D. P. Feldman. Regularities unseen, randomness observed: The entropy convergence hierarchy. 1998. In Preparation.

[40] J. P. Crutchfield and J. E. Hanson. Turbulent pattern bases for cellular automata. *Physica D*, 69:279–301, 1993.

[41] J. P. Crutchfield and N. H. Packard. Noise scaling of symbolic dynamics entropies. In H. Haken, editor, *Evolution of Order and Chaos*, pages 215–227, Berlin, 1982. Springer-Verlag.

[42] J. P. Crutchfield and N. H. Packard. Symbolic dynamics of one-dimensional maps: Entropies, finite precision, and noise. *Intl. J. Theo. Phys.*, 21:433–466, 1982.

[43] J. P. Crutchfield and N. H. Packard. Symbolic dynamics of noisy chaos. *Physica D*, 7:201–223, 1983.

[44] J. P. Crutchfield and K. Young. Inferring statistical complexity. *Phys. Rev. Lett.*, 63:105–108., 1989.

[45] J. P. Crutchfield and K. Young. Computation at the onset of chaos. In W. H. Zurek, editor, *Complexity, Entropy and the Physics of Information*, volume VIII of *Santa Fe Institute Studies in the Sciences of Compexity*, pages 223–269. Addison-Wesley, 1990.

[46] A. Csordás and P. Szépfalusy. Singularities in Rényi information as phase transitions in chaotic states. *Phys. Rev. A.*, 39(9):4767–4777, 1989.

[47] P. Cvitanović. Invariant measurement of strange sets in terms of cycles. *Phys. Rev. Lett.*, 61:2729–2732, 1988.

[48] J. Delgado and R. V. Solé. Collective-induced computation. *Phys. Rev. E*, 55(3):2338–2344, 1997.

[49] J. F. Dobson. Many-neighbored Ising chain. *J. Math. Phys.*, 10(1):40–45, January 1969.

[50] K. E. Drexler. *Nanosystems: molecular machinery, manufacturing, and computation.* Wiley, New York, 1992.

[51] R. J. Elliot, L. Aggoun, and J. B. Moore. *Hidden Markov Models: Estimation and Control*, volume 29 of *Applications of Mathematics.* Springer, New York, 1995.

[52] K-E. Eriksson and K. Lindgren. Structural information in self-organizing systems. *Physica Scripta*, 1987.

[53] K.-E. Eriksson and K. Lindgren. Entropy and correlations in lattice systems. Technical report, Physical Resource Theory Group, Chalmers University of Technology, Göteborg, Sweden, 1989.

[54] D. P. Feldman and J. P. Crutchfield. Discovering non-critical organization: Statistical mechanical, information theoretic, and computational views of patterns in simple one-dimensional spin systems. 1998. Submitted to the Journal of Statistical Physics.

[55] D. P. Feldman and J. P. Crutchfield. Measures of statistical complexity: Why? *Phys. Lett. A*, 238:244–252, 1998.

[56] R. P. Feynman. *Feynman Lectures on Computation*. Addison-Wesley, 1996. A. J. G. Hey and R. W. Allen, eds.

[57] K. H. Fischer and J. A. Hertz. *Spin Glasses*. Cambridge Studies in Magnetism. Cambridge University Press, Cambridge, 1988.

[58] J. Freund, W. Ebeling, and K. Rateitschak. Self-similar sequences and universal scaling of dynamical entropies. *Phys. Rev. E*, 54:5561–5566, 1996.

[59] M. Gell-Mann. *The quark and the jaguar: adventures in the simple and the complex*. W. H. Freeman, 1994.

[60] M. Gell-Mann. What is complexity? *Complexity*, 1(1):16–19, 1995.

[61] M. Gell-Mann and S. Lloyd. Information measures, effective complexity, and total information. *Complexity*, 2(1):44–52, 1996.

[62] W. M. Goncalves, R. D. Pinto, J. C. Sartorelli, and M. J. de Oliveira. Inferring statistical complexity in the dripping faucet experiment. *Physica A*, 1998. To appear.

[63] W. T. Grandy. *Foundations of Statistical Mechanics*. Fundamental Theories of Physics. D. Reidel, Dordrecht, 1988.

[64] P. Grassberger. Toward a quantitative theory of self-generated complexity. *Intl. J. Theo. Phys.*, 25(9):907–938, 1986.

[65] R. M. Gray. *Entropy and Information Theory*. Springer-Verlag, New York, 1990.

[66] I. Guyon and P. S. P. Wang, editors. *Advances in Pattern Recognition systems Using Neural Network Technologies*. World Scientific, 1993.

[67] H. Haken. *Synergetics*. Springer-Verlag, third edition, 1983.

[68] 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*, 33:1141–1151, 1986.

[69] J. E. Hanson. *Computational Mechanics of Cellular Automata.* PhD thesis, University of California, Berkeley, 1993.

[70] J. E. Hanson and J. P. Crutchfield. The attractor-basin portrait of a cellular automaton. *J. Stat. Phys.*, 66:1415–1462, 1992.

[71] J. E. Hanson and J. P. Crutchfield. Computational mechanics of cellular automata: An example. *Physica D*, 103(1-4):169–189, 1997.

[72] R. V. L. Hartley. Transmission of information. *Bell Sys. Tech. J.*, July:535–563, 1928.

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

[74] B. A. Huberman and T. Hogg. Complexity and adaptation. *Physica D*, 22:376–384, 1986.

[75] E. Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift f. Physik*, 31:253, 1925.

[76] E. T. Jaynes. *Essays on Probability, Statistics, and Statistical Physics.* Reidel, London, 1983.

[77] J. Justesen and Y. M. Shtarkov. Simple models of two-dimensional information sources and codes. To be presented at the Information Theory Symposium, Boston, MA, August 1998.

[78] A. Kato and K. Zeger. On the capacity of two-dimensional run length limited codes. To be presented at the Information Theory Symposium, Boston, MA, August 1998.

[79] S. Kauffman. *At Home in the Universe: The Search for Laws of Complexity.* Oxford University Press, 1995.

[80] Z. Kaufmann. Characteristic quantities of multifractals—application to the Feigenbaum attractor. *Physica D*, 54:75–84, 1991.

[81] A. I. Khinchin. *Mathematical foundations of information theory.* Dover, New York, 1957.

[82] B. Kitchens and S. Tuncel. Finitary measures for subshifts of finite type and sofic systems. *Memoirs of the AMS*, 58(338):1–68, 1985.

[83] W. R. Knorr. *The Ancient Tradition of Geometric Problems.* Birkhauser, Boston, 1986.

[84] S. Kobe. Ernst Ising—physicist and teacher. *Actas: Noveno Taller Sur de Fisica del Solido, Misión*, page 1, 1995. cond-mat/9605174.

[85] A. N. Kolmogorov. A new metric invariant of transient dynamical systems and automorphisms in Lebesgue spaces. *Dokl. Akad. Nauk. SSSR*, 119:861–864, 1958. (Russian) Math. Rev. vol. 21, no. 2035a.

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

[87] A. N. Kolmogorov. Combinatorial foundations of information theory and the calculus of probabilities. *Russ. Math. Surveys*, 38:29, 1983.

[88] M. Koppel. Complexity, depth, and sophistication. *Complex Systems*, 1:1087–1091, 1987.

[89] H. A. Kramers and G. H. Wannier. Statistics of the two-dimensional ferromagnet: Part I. *Phys. Rev.*, 60:252–263, 1941.

[90] R. Landauer. A simple measure of complexity. *Nature*, 336(6197):306–307, 1988.

[91] A. Lempel and J. Ziv. Compression of two-dimensional data. *IEEE Transactions on Information Theory*, 32(1):2–8, 1986.

[92] W. Lenz. *Phys. Zeitschrift*, 21:613, 1920.

[93] R. Lewin. *Complexity: life at the edge of chaos.* Macmillian Publishing Company, 1992.

[94] M. Li and P. M. B. Vitanyi. *An Introduction to Kolmogorov Complexity and its Applications.* Springer-Verlag, New York, 1993.

[95] W. Li. Mutual information functions versus correlation functions. *J. Stat. Phys.*, 60(5/6):823–837, 1990.

[96] W. Li. On the relationship between complexity and entropy for Markov chains and regular languages. *Complex Systems*, 5(4):381–399, 1991.

[97] R. Lidl and G. Pilz. *Applied Abstract Algebra.* Springer, New York, 1984.

[98] D. Lind and B. Marcus. *An Introduction to Symbolic Dynamics and Coding.* Cambridge University Press, 1995.

[99] K. Lindgren. Microscopic and macroscopic entropy. *Phys. Rev. A*, 38(9):4794–4798, 1988.

[100] K. Lindgren. Entropy and correlations in dynamical lattice systems. In P. Manneville, N. Boccara, G. Y. Vichniac, and R. Bidaux, editors, *Cellular Automata and Modeling of Complex Systems*, volume 46 of *Springer Proceedings in Physics*, pages 27–40, Berlin, 1990. Springer-Verlag.

[101] K. Lindgren. Entropy and correlations in discrete dynamical systems. In J. L. Casti and A. Karlqvist, editors, *Beyond Belief: Randomness, Prediction and Explanation in Science*, pages 88–109. CRC Press, 1991.

[102] K. Lindgren, 1998. Personal Communication.

[103] K. Lindgren, C. Moore, and M. G. Nordahl. Complexity of two-dimensional patterns. *Journal of Statistical Physics*, 1998. To appear.

[104] K. Lindgren and M. G. Nordhal. Complexity measures and cellular automata. *Complex Systems*, 2(4):409–440, 1988.

[105] S. Lloyd and H. Pagels. Complexity as thermodynamic depth. *Annals of Physics*, 188:186–213, 1988.

[106] R. Lopez-Ruiz, H. L. Mancini, and X. Calbet. A statistical measure of complexity. *Physics Letters A*, 209:321–326, December 1995.

[107] R. Mainieri. Cycle expansion for the Lyapunov exponent of a product of random matrices. *Chaos*, 2(1):91–97, 1992.

[108] R. Mainieri. Thermodynamic Zeta functions for Ising models with long-range interactions. *Phys. Rev. A*, 45(6):3580–3591, 1992.

[109] H. Matsuda, K. Kudo, R. Nakamura, O. Yamakawa, and T. Murata. Mutual information of Ising systems. *International Journal of Theoretical Physics*, 35(4):839–845, 1996.

[110] B. M. McCoy and T. T. Wu. *The Two-Dimensional Ising Model*. Harvard, 1993.

[111] B. McMillan. The basic theorems of information theory. *Ann. Math. Stat.*, 24:196–219, 1953.

[112] D. Michie, D. Spiegelhalter, and C. C. Taylor, editors. *Machine Learning, Neural and Statistical Classification*, Series in Artificial Intelligence, New York, 1994. E. Horwood.

[113] R. K. Mishra, D. Maass, and E. Zwierlein, editors. *On Self-Organization: An Interdisciplinary Search for a Unifying Principle*. Springer-Verlag, 1994.

[114] L. Nadel and D. Stein, editors. *1993 Lectures in Complex Systems.* Addison-Wesley, 1995. See also other volumes in the series.

[115] K. Nagel. Particle hopping models and traffic flow theory. *Phys. Rev. E*, 53:4655–4672, 1996.

[116] K. Nagel and M. Paczuski. Emergent traffic jams. *Phys. Rev. E*, 51:2909–2918, 1995.

[117] M. G. Nordahl. Cellular automata probability measures. In P. Manneville, N. Boccara, G. Y. Vichniac, and R. Bidaux, editors, *Cellular Automata and Modeling of Complex Systems*, volume 46 of *Springer Proceedings in Physics.* Springer-Verlag, Berlin, 1990.

[118] Hero of Alexandria. *Opera*, volume III: Metrica. B. G. Teubner, Leipzig, 1903.

[119] Y. Oono. Large deviation and statistical physics. *Prog. Theo. Phys.*, 99:165–205, 1989.

[120] E. Ordentlich and R. M. Roth. On the redundancy of two-dimensional balanced codes. To be presented at the Information Theory Symposium, Boston, MA, August 1998.

[121] E. Ott. *Chaos in Dynamical Systems.* Cambridge University Press, 1993.

[122] N. H. Packard. *Measurements of Chaos in the Presence of Noise.* PhD thesis, University of California, Santa Cruz, 1982.

[123] C. H. Papadimitriou. *Computational Complexity.* Addison-Wesley, Reading, Massachusetts, 1994.

[124] G. Parisi. *Statistical Field Theory*, volume 66 of *Frontiers in Physics.* Addison-Wesley, 1988.

[125] A. Paz. *Introduction to Probabilistic Automata.* Academic Press, New York, 1971.

[126] R. Peierls. *Proc. Cambridge Phil. Soc.*, 32:477, 1936.

[127] H.-O. Peitgen, H. Jürgens, and D. Saupe. *Chaos and Fractals: New Frontiers of Science.* Springer-Verlag, 1992.

[128] L. Peliti and A. Vulpiani, editors. *Measures of Complexity.* Springer-Verlag, 1988.

[129] P. Perner, P. Wang, and A. Rosenfeld, editors. *Advances in Structural and Syntactical Pattern Recognition.* Springer, 1996.

[130] M. Plischke and B. Bergensen. *Equilibrium Statistical Physics.* Prentice Hall, 1989.

[131] M. Raijmakers. *Epigensis in Neural Network Models of Cognitive Development: Bifurcations, More Powerful Structures, and Cognitive Concepts.* PhD thesis, Universiteit van Amsterdam, 1996.

[132] J. Rhodes. *Applications of Automata Theory and Algebra via the mathematical theory of complexity to: Biology, Physics, Psychology, Philosophy, Games, Codes.* 1971.

[133] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[134] J. Rissanen. Universal coding, information, prediction, and estimation. *IEEE Trans. Info. Th.*, IT-30:629–636, 1984.

[135] J. Rissanen. *Stochastic Complexity in Statistical Inquiry.* World Scientific Publisher, Singapore, 1989.

[136] Harry S. Robertson. *Statistical Theromphysics.* Prentice Hall, 1993.

[137] J. Rothstein. Generalized entropy, boundary conditions, and biology. In R.D. Levine and M. Tribus, editors, *The Maximum Entropy Formalism.* MIT Press, Cambridge, Massachusetts, 1979.

[138] D. Ruelle. *Statistical Mechanics: Rigorous Results.* Addison Wesley, 1989.

[139] D. Saad and S. A. Solla. On-line learning in soft committee machines. *Phys. Rev. E*, 52:4225–4243, 1995.

[140] T. D. Schultz, D. C. Mattis, and E. H. Lieb. Two-dimensional Ising model as a soluble problem of many fermions. *Rev. Mod. Phys.*, 36:856–871, 1964.

[141] J. Schurmann. *Pattern Classification: A Unified View of Statistical and Neural Approaches.* Wiley, New York, 1996.

[142] C. E. Shannon. A mathematical theory of communication. *Bell System Tech. J.*, 27:379–423, 1948. as reprinted in "The Mathematical Theory of Communication", C. E. Shannon and W. Weaver, University of Illinois Press, Champaign-Urbana (1963).

[143] C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication.* University of Illinois Press, 1963.

[144] R. Shaw. *The Dripping Faucet as a Model Chaotic System.* Aerial Press, Santa Cruz, California, 1984.

[145] D. Sheinwald, A. Lempel, and J. Ziv. Two-dimensional encoding by finite-state encoders. *IEEE Transactions on Communications*, 38(3):341–347, 1990.

[146] P. H. Siegel and J. K. Wolf. Bit-stuffing bounds on the capacity of two-dimensional constrained arrays. To be presented at the Information Theory Symposium, Boston, MA, August 1998.

[147] Ja. G. Sinai. On the concept of entropy of a dynamical system. *Dokl. Akad. Nauk. SSSR*, 124:768–771, 1959.

[148] R. V. Solé, C. Blanc, and B. Luque. Statistical measures of complexity for strongly interacting systems. 1998. SFI Working Paper 97-11-083, Submitted to Physics Letters A.

[149] R. V. Solé, S. C. Manrubia, B. Luque, J. Delgado, and J. Bascompte. Phase transitions and complex systems. *Complexity*, 1(4):13–26, 1996.

[150] M. R. Spiegel. *Mathematical Handbook of Formulas and Tables.* McGraw-Hill, 1968.

[151] K. W. Sulston, B. L. Burrows, and A. N. Chishti. Recursive procedures for measuring disorder in non-periodic sequences. *Physica A*, 217:146–160, 1995.

[152] P. Szépfalusy and G. Györgyi. Entropy decay as a measure of stochasticity in chaotic systems. *Phys. Rev. A*, 33(4):2852–2855, 1986.

[153] B. A. Trakhtenbrot and Ya. M. Barzdin. *Finite Automata.* North-Holland, Amsterdam, 1973.

[154] D. R. Upper. *Theory and Algorithms for Hidden Markov Models and Generalized Hidden Markov Models.* PhD thesis, University of California, Berkeley, 1997.

[155] M. N. van Emden. *An Analysis of Complexity*, volume 35 of *Mathematical Centre Tracts.* Mathematisch Centrum Amsterdam, 1971.

[156] E. van Nimwegen, J. P. Crutchfield, and M. Mitchell. Finite populations induce metastability in evolutionary search. *Physics Letters A*, 229(3):144–150, 1997.

[157] E. van Nimwegen, J. P. Crutchfield, and M. Mitchell. Statistical dynamics of the Royal Road genetic algorithm. *Theoret. Comp. Sci.*, 1998. In press.

[158] B. Wackerbauer, A. Witt, H. Atmanspacher, J. Kurths, and H. Scheingraber. A comparative classification of complexity measures. *Chaos, Solitons & Fractals*, 4(1):133–173, 1994.

[159] D. Walgraef. *Spatio-Temporal Pattern Formation, with Examples from Physics, Chemistry, and Materials Science.* Springer-Verlag, 1997.

[160] C. S. Wallace and D. M. Boulton. An information measure for classification. *Comput. J.*, 11:185, 1968.

[161] S. Watanabe. *Knowing and Guessing; A Quantitative Study of Inference and Information.* Wiley, New York, 1969.

[162] T. L. H. Watkin, A. Rau, and M. Biehl. The statistical mechanics of learning a rule. *Rev. Mod. Phys.*, 65:499–556, 1993.

[163] W. Weeks and Richard E. Blahut. The capacity and coding gain of certain checkerboard codes. *IEEE Transactions on Information Theory*, 44(3):1193–1203, 1998.

[164] K. Wilson. Problems in physics with many scales of length. *Scientific American*, 241(2):158–179, 1979.

[165] A. Witt, A. Neiman, and J. Kurths. Characterizing the dynamics of stochastic bistable systems by measures of complexity. *Physical Review*, E55(5):5050–5059, 1997.

[166] S. Wolfram. Statistical mechanics of cellular automata. *Reviews of Modern Physics*, 55(3):601–644, 1983.

[167] S. Wolfram. Computation theory of cellular automata. *Communications in Mathematical Physics*, 96(15-57), 1984.

[168] S. Wolfram. Universality and complexity in cellular automata. *Physica*, 10D:1–35, 1984.

[169] S. Wolfram, editor. *Theory and Applications of Cellular Automata.* World Scientific, 1986.

[170] J. M. Yeomans. *Statistical Mechanics of Phase Transitions.* Clarendon Press, Oxford, 1992.

[171] K. Young. *The Grammar and Statistical Mechanics of Complex Physical Systems.* PhD thesis, University of California, Santa Cruz, 1991.

[172] K. Young and J. P. Crutchfield. Fluctuation spectroscopy. *Chaos, Solitons, and Fractals*, 4:5–39, 1993.