Causal Architecture, Complexity and Self-Organization in Time Series and Cellular Automata

Cosma Rohilla Shalizi

4 May 2001

To Kris, for being perfect

Abstract

All self-respecting nonlinear scientists know self-organization when they see it: except when we disagree. For this reason, if no other, it is important to put some mathematical spine into our floppy intuitive notion of self-organization. Only a few measures of self-organization have been proposed; none can be adopted in good intellectual conscience.

To find a decent formalization of self-organization, we need to pin down what we mean by organization. The best answer is that the organization of a process is its *causal architecture* — its internal, possibly hidden, causal states and their interconnections. *Computational mechanics* is a method for inferring causal architecture — represented by a mathematical object called the *ϵ-machine* — from observed behavior. The ϵ-machine captures all patterns in the process which have any predictive power, so computational mechanics is also a method for *pattern discovery*. In this work, I develop computational mechanics for four increasingly sophisticated types of process — memoryless transducers, time series, transducers with memory, and cellular automata. In each case I prove the optimality and uniqueness of the ϵ-machine's representation of the causal architecture, and give reliable algorithms for pattern discovery.

The ϵ-machine *is* the organization of the process, or at least of the part of it which is relevant to our measurements. It leads to a natural measure of the *statistical complexity* of processes, namely the amount of information needed to specify the state of the ϵ-machine. Self-organization is a self-generated increase in statistical complexity. This fulfills various hunches which have been advanced in the literature, seems to accord with people's intuitions, and is both mathematically precise and operational.

Contents

List of Figures

Notes on the Text and on Sources

This book was composed in LaTeX2ϵ, using the `book` document class. The `fancyheadings` package made the page headings conform to the University of Wisconsin's requirements for dissertations. I highly recommend this combination to anyone else writing a dissertation at UW, as it's much easier than trying to hack LaTeX, or even than copying the code of someone who has. The figures were produced by a large number of graphics programs, but all the box-and-arrow diagrams were done using the free program `dot` from AT&T[1], which was a life-saver.

Almost everything in this book I've either published or at least presented in other forms. It is not, however, simply those papers stapled together; I've tried to weld the separate parts into a single smooth structure, where the overall development is as streamlined, progressive and unified as possible. Here, for what it's worth, are prior sources for various parts of the text.

Chapter	Sources
2	Shalizi (1996a); Shalizi and Crutchfield (2001), Section II
3	Shalizi and Crutchfield (2000b)
4	Crutchfield and Shalizi (1999); Shalizi and Crutchfield (2001), Sections III–VI
5	Klinkner and Shalizi (2001)
6	Shalizi and Crutchfield (2001), Appendix H
7	Shalizi and Crutchfield (2000a)
9	Hordijk, Shalizi and Crutchfield (2001)
10	Shalizi, Haslinger and Crutchfield (2001)

This dissertation will be on-line at `http://www.santafe.edu/projects/CompMech/`, and probably at `arxiv.org` as well.

[1] `http://www.research.att.com/sw/tools/graphviz/`

Chapter 1

Introduction

This is a book about causal architecture, pattern discovery, complexity and self-organization. Those are vague, even grandiose themes. You may well doubt that I have anything worthwhile to say about any of them, let alone all of them; if I found a book with this title (by someone else!), I'd be skeptical myself. Still, by the end I hope to have convinced you that there *is* an optimal method for discovering causal patterns in data, that it gives us a natural and satisfying measure of complexity, and that it at least might give us an operational test for self-organization.

Let me try to unpack that a little. The problem of trying to determine causal relations from observations is very ancient, as are objections that no such thing is possible[1]. Similarly ancient is the problem of discovering patterns in the natural world, and of doing so *reliably,* without fooling ourselves. Even the notion of self-organization (though not that name) is very old, being clearly articulated by Descartes at the beginning of the scientific revolution. And similarly for complexity. All of these problems can be, and often are, phrased in vague, suggestive ways which hint at connections to each other. But phrased that way, they hint at connections to everything under the sun, and much else beyond. What I intend to do here is show that the notion of causal architecture can be made precise; that the inference of causal architecture is a special case of pattern discovery, which is also precise; and that a particular method of discovering causal patterns is optimal. The rest will follow more or less naturally from this.

But first I should say a little bit about self-organization, and why it's worth explicating.

1.1 Self-Organization

Set a child alone with a heap of Legos, and in an hour the Legos are probably a lot more ordered than they were at the start. Their organization has increased, but in a pretty unmysterious and, to us, uninteresting way: the kid organized them. On the other hand, there are some things which will organize without this kind of outside intervention, which self-organize. (Compare Figure 1.1 with Figure 1.2.)

There is a long tradition of interest in self-organizing processes (see Section 1.5 below); in recent decades, as the poet (Sterling 1985) says, "the march of science became a headlong stampede." This has obscured the fact that we do not have anything like a *theory* of self-organization, not even an unambiguous test for it. The point of this book is not to provide a full-fledged theory of self-organization but, more modestly, to formalize the concept in a way which leads to practical tests.

Currently, the state of the art for testing whether or not a process is self-organizing boils down to "I know it when I see it." This may be acceptable to artists and Supreme Court Justices, but it cannot satisfy those who fondly imagine their trades to be exact sciences. Moreover, there are likely to be many cases where my intuition conflicts with yours; this is notoriously the case with art, despite the fact that *Homo sapiens*

[1] For some of the earliest history of the problem of causation, see Kogan (1985) and his sources, ibn Rushd (Averroës) (1180–1185/1954) and al Ghazali (1100/1997).

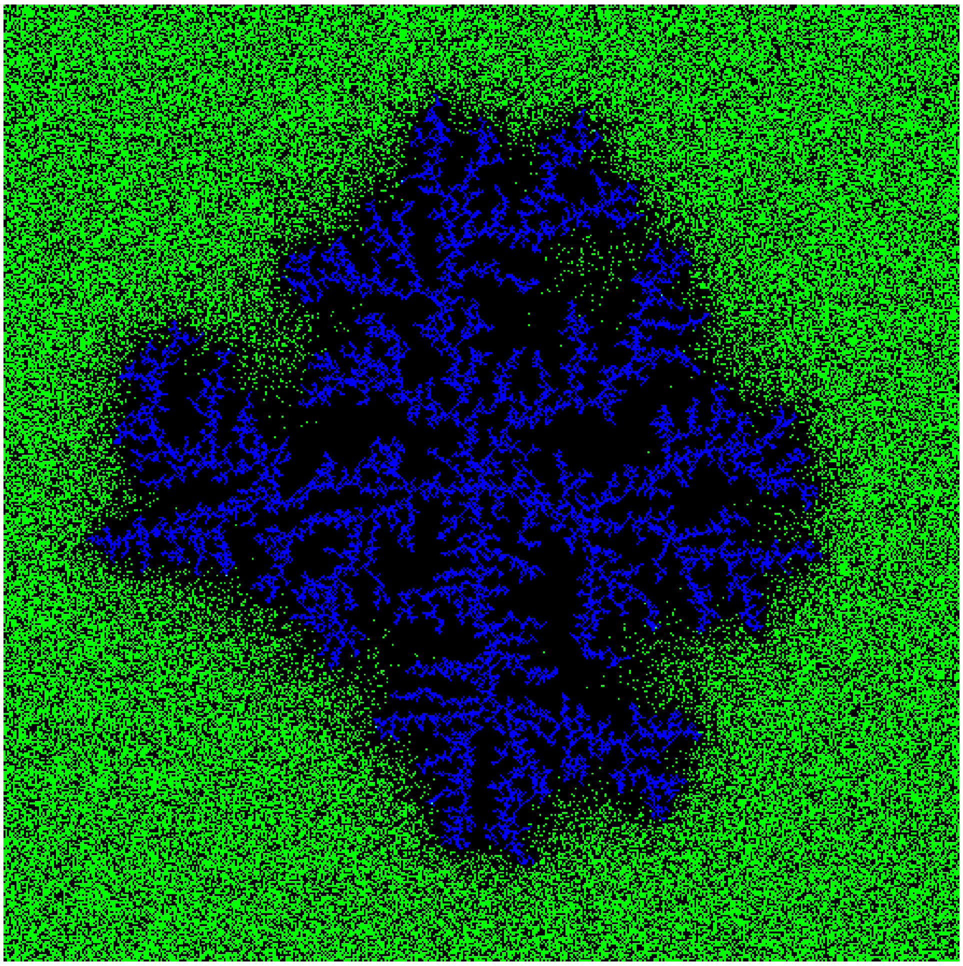

Figure 1.1: Pattern formed by diffusion-limited aggregation (Vicsek 1989; Barabási and Stanley 1995; D'Souza and Margolus 1999). At each time-step, particles (pale dots) perform independent random walks; if one of them hits the aggregate (dark mass), it stops moving and joins the aggregate. This image (made on a CAM 8 computer) shows a pattern formed on a 1024×1024 grid, starting with a single aggregate particle at the center and an initial density of free particles of 0.20, after about $1.7 \cdot 10^4$ time-steps. This cellular automaton (Chapter 8) models many natural self-organizing processes (Ball 1999, ch. 5).

Figure 1.2: Detail from *Japanese maple leaves stitched together to make a a floating chain; the next day it became a hole; supported underneath by a woven briar ring*, by Andy Goldsworthy (1990).

has been exposed to it since the Old Stone Age[2]. Given that self-organization is a recent concept, and one whose careful use hasn't been under a lot of cultural selection pressure, what's surprising is not that there are controversies over whether, e.g., turbulent flows or ecological successions are self-organizing, but that we have any agreement at all about what self-organizes.

1.2 Formalizing an Intuitive Notion

Most of our notions are intuitive, unformalized, and vague. This suits us well enough, most of the time, and arguably some degree of vagueness is inevitable. Still, from time to time we want to make a notion less vague, less intuitive and more explicit, more amenable to examination and reasoning — to *formalize* it. That's the case here: we want to make "self-organizing" a formal concept and no longer just an intuitive one. In essence, we want a definition, "x is self-organizing iff x is . . . ," followed by a list of individually necessary and jointly sufficient conditions.

Now, it's not as though "self-organizing" has some inner essence which such a definition tries to capture. (If it did, we'd be in less need of formalization!) Rather, the goal is to replace a squishy notion with a more rigid one which can do the same work, and more besides. As the usual authorities in the matter of formalizing intuitions (Quine 1961) insist, the goal is that the formal notion match the intuitive one in all the easy cases; resolve the hard ones in ways which don't make us boggle; and let us frame simple and fruitful generalizations. This is of a piece with the art of definition in mathematics generally, nicely put by Spivak (1965, p. 104):

> Stokes' theorem shares three important attributes with many fully evolved theorems:

[2] As the poet (Kipling 1940) warns us, "But the Devil whoops, as he whooped of old: 'It's clever, but is it Art?"' Of course, we want things which aren't clever, or art . . . .

1. It is trivial.
2. It is trivial because the terms appearing in it have been properly defined.
3. It has significant consequences.

Well, not quite all of a piece: we want a concept which applies to real physical systems, too, so we want to be able to decide whether it applies using just experimental data, or, in a pinch, models not too far abstracted from data. And the simple, fruitful generalizations will have to wait for someone else, or at least some other book.

1.3 The Strategy

The notion of self-organization does not yield to frontal assault. We need (as it were) to tunnel under its walls of confusion and then take it from within; and to do that, we need tunnels, and tunneling machinery. Less metaphorically, I plan to convince you that we should represent the causal architecture of processes using one particular kind of mathematical object. I will also convince you that these representations can be discovered from data — that the pattern discovery problem is soluble. I am going to do this four times over, for four different and increasingly sophisticated kinds of processes: memoryless transducers, time series, transducers with memory, and cellular automata.

Having reduced you to a suggestible state by proving optimality theorems, it will be fairly easy to convince you that the causal architecture of a process *is* its organization. And then the trick will be turned, because there is a natural quantitative measure of the complexity of the causal architecture, i.e., of the organization of the process.

A word about the math. I aim at a moderate degree of rigor throughout — but as two wise men have remarked, "One man's rigor is another man's mortis" (Bohren and Albrecht 1998). My ideal has been to keep to about the level of rigor of Shannon (1948)[3]. In some places (like Appendix B.3), I'm closer to the onset of mortis. No result on which anything else depends should have an invalid proof. There are places, naturally, where I am not even trying to be rigorous, but merely plausible, or even "physical," but it should be clear from context where those are.

1.4 A Summary

The outline of the book is as follows.

This chapter closes with a section sketching, very briefly, the histories of the idea of self-organization, of methods of pattern discovery, and of computational mechanics. Chapter 2 discusses previous proposals for how to measure organization, which leads to a more general discussion of how to measure complexity and how to define and describe patterns. I conclude that chapter by rejecting all but one of the previous proposals for quantifying self-organization, and giving a list of desiderata that any approach to patterns should fulfill. Readers who wish to go straight to the science may skip both Section 1.5 and Chapter 2, except for Section 2.3.5.

The rest of the book develops the only approach to patterns I know of which meets all those requirements, namely the method of *computational mechanics* developed by Jim Crutchfield and his minions, one of which I have the honor to be. Chapter 3 builds up computational mechanics in the simplest setting, that of causal relationships which depend only on the present value of measured variables, or *memoryless transduction* (Shalizi and Crutchfield 2000b). This is where I introduce one of the key concepts of computational mechanics, that of *causal states* (first defined by Crutchfield and Young (1989)), and show that they are the unique, optimal, minimal predictors of the transduction process.

[3] Of course, a mathematician might say that that's not rigorous at all, but this *is* supposed to be physics, and I would be *extremely happy* if annoyance at my style led people to re-do this work with all the bells, σ-algebras, whistles, categories and gongs.

Chapter 4 extends the approach to the more sophisticated setting of time series and (classical) stochastic processes. Here I introduce the other key concept of computational mechanics, that of the ϵ-*machine*, which shows the structure of connections over time between causal states. Using the ϵ-machine, we see that the causal states always form a Markov process. This is satisfying ideologically, and has interesting information-theoretic and ergodic consequences. While ϵ-machines themselves date back to Crutchfield and Young (1989), the methods of proof and results presented here follow Crutchfield and Shalizi (1999) and especially Shalizi and Crutchfield (2001).

The next three chapters build on the computational mechanics of time series. Chapter 5 describes a new procedure for reconstructing the causal states and the ϵ-machine of a time series from data, with a number of advantages over prior methods (Klinkner and Shalizi 2001). Chapter 6 compares and contrasts computational mechanics with a bunch of other, better-known ways of dealing with time series and their complexity (Shalizi and Crutchfield 2001). Chapter 7 shows how to build ϵ-machines for stochastic transducers with memory, by treating such transducers as coupled time-series (Shalizi and Crutchfield 2000a).

Chapter 8 introduces cellular automata, first in a very informal way, then in a more formal way which makes clear that they are dynamical systems, in two different ways, and much like any other bunch of maps, with all the modern conveniences (attractors, basins of attraction, etc.). Then, in Chapter 9 I discuss a very useful set of tools that developed by Crutchfield and Hanson to deal with the dynamics of spatial configurations in CA, using the notion of a "regular language" from computer science. These tools let us decompose one-dimensional CA configurations into extended *domains* and localized *particles*. The domains are regions of space and time which are, in a precise sense, doing next to nothing; the particles are propagating disturbances in (or between) the domains. Following Hordijk, Shalizi and Crutchfield (2001), I use the tools to prove a limit on how many ways the particles in a given CA can interact, and since the particles are what the CA computes with, this also limits the CA's computational power.

The domain-and-particle tools amount to a kind of purely spatial computational mechanics, and they employ a set of spatial causal states. For technical reasons, spatial computational mechanics only applied to one-dimensional cellular automata (or 1+1D, to field theorists), a restriction known as the *2D bummer* (Feldman). A fully spatio-temporal computational mechanics, like a spatial computational mechanics for higher dimensions, seemed out of reach for the longest time. (Life would be easier if the universe consisted of a single point (Calvino 1965/1968, ch. 4).) Chapter 10 explains what the difficulties were, shows how to overcome them so as to define local causal states for each point in space-time, and shows that the causal states of spatial computational mechanics are also spatio-temporal causal states (Shalizi, Haslinger and Crutchfield 2001).

The last chapter summarizes everything that's been done in the book, for the benefit of those who only read introductions and conclusions. I proceed to define emergence and self-organization, following Crutchfield (1994a, Crutchfield (1994b) with some technical refinements, and an illustrative back-of-the-envelope calculation. Then I list some unsolved problems and desirable extensions of the mathematical foundation of computational mechanics. I close by throwing out suggestions for things to examine with these tools, developed in our group at SFI, some vague, grandiose ideas about learning, and prophecy that stock-market quants will go the way of Lancashire weavers. Some of this material is frankly speculative, but I hope by that point you'll be so overwhelmed that you'll accept anything I say — that is, that you'll be swayed by the intrinsic merits of my arguments.

A couple of appendices follow, to remind you about mathematical tools (information theory, conditional measures, formal languages, etc.) you probably forgot how to use before I was out of diapers, and to hold more peripheral bits of math which would clog up the main chapters.

The key chapters, which should form a coherent sequence, are 3, 4, 7 and 10. The last two are partially independent of each other, but if you're interested in the spatial-process material in Chapter 10, you should probably read Chapter 9 as well.

1.5 Historical Sketch

[Consider] what would happen in a new world, if God were now to create somewhere in the imaginary spaces matter sufficient to compose one, and were to agitate variously and confusedly the different parts of this matter, so that there resulted a chaos as disordered as the poets ever feigned, and after that did nothing more than lend his ordinary concurrence to nature, and allow her to act in accordance with the laws which He had established . . . . I showed how the greatest part of the matter of this chaos must, in accordance with these laws, dispose and arrange itself in such a way as to present the appearance of heavens; how in the meantime some of its parts must compose an earth and some planets and comets, and others a sun and fixed stars. . . . I came next to speak of the earth in particular, and to show how . . . the mountains, seas, fountains, and rivers might naturally be formed in it, and the metals produced in the mines, and the plants grow in the fields and in general, how all the bodies which are commonly denominated mixed or composite might be generated . . . [S]o that even although He had from the beginning given it no other form than that of chaos, provided only He had established certain laws of nature, and had lent it His concurrence to enable it to act as it is wont to do, it may be believed, without discredit to the miracle of creation, that, in this way alone, things purely material might, in course of time, have become such as we observe them at present; and their nature is much more easily conceived when they are beheld coming in this manner gradually into existence, than when they are only considered as produced at once in a finished and perfect state.
Rene Descartes (1637, part 5)

This section consists of a few brief remarks on the invention and use of the idea of self-organization, so readers who just want to go straight to the science can skip it. If you're of the opposite inclination, and want more details, there is, alas, no decent history of self-organization for me to point you to. In the unlikely event that historians of science read these pages, I should like to bring this little-needed gap to their attention[4].

1.5.1 Origins of the Concept

While the notion of spontaneous, dynamically-produced organization is very old[5], it only crystallized into the term "self-organization" in the years after the Second World War, in circles connected with cybernetics and computing machinery (Yovits and Cameron 1960; Von Foerester and Zopf Jr 1962). The first appearance of the term seems to be in a 1947 paper by W. Ross Ashby[6].

Remarkably, Ashby gave a pretty clear explanation of what he meant by "organization": to paraphrase, the organization of a system was the functional dependence of its future state on its present state and its current external inputs, if any. That is, if the state space is S and the input space is I, the organization of the system is the function $f : S \times I \mapsto S$ which gives the new state. Ashby understood a system to be self-organizing if it changed its own organization, rather than being rewired by an external agency. How is that possible?

[4] The historical remarks in Prigogine's popular books (Prigogine 1980; Prigogine and Stegners 1979/1984) are at best badly informed, at worst tendentious. Krohn, Küppers and Nowotny (1990) is highly unreliable scientifically and historically.

[5] The first articulation of the concept I have found is that by Descartes, in the epigraph to this section. (See also Descartes 1664.) It subsequently played an important, if subterranean, role in European culture, in naturalistic philosophy (Vartanian 1953), in associationist psychology (Hume 1739) and in political and economic liberalism (Mayr 1986). Before the early modern period, naturalistic philosophies seem to have relied on "time and chance" explanations of organization, along the lines of the ancient atomists. But these are matters for another time, and another book.

[6] Most sources which say anything about the origins of the term, attribute it to Farley and Clark (1954), but this is plainly wrong — the latter cite Ashby. As to Ashby himself, he was a British psychiatrist who in the 1940s independently arrived at many of the ideas which Norbert Wiener bundled as "cybernetics," and was active in the cybernetics movement after the war. He is a seriously underappreciated figure in the pre-history of the sciences of complexity. Not only is there no biography of him, but he isn't even mentioned in the standard historical reference works, and there's one sentence on him in Heims's *The Cybernetics Group* (1991). See, however, Ashby's books (1956, 1960), Wiener (1954), and the obituary notice by Conant (1974).

Ashby's ingenious answer is that it is not. Organization is invariant. It may be, however, that the function f is well-approximated by another function g in a certain region of the state space and by a different function h in another region. If the dynamics then drive the system from the first region to the second, we will see an *apparent* change in the organization, from g to h, though the true, underlying dynamics remains the same. (Ashby elaborated on this argument in his 1962 paper. For suggestive remarks on the importance of thresholds in this process, see Ashby (1960).)

At the end of the day, the concepts of organization and self-organization we will emerge with will be quite similar, verbally, to Ashby's. There are three reasons why this book doesn't end right here. The first is that Ashby's ideas about what constitutes self-organization have been pretty thoroughly ignored by everyone else who's used the idea. The second is that they don't go far enough: they don't let us distinguish changes that lead to more organization from those which lead to less, or even from those which are neutral with respect to how organized the process looks. The third is that, while the wordy version of organization, in my sense, will be very close to Ashby's, the math will be pretty different, much more rigorous, and will resolve the second problem, of distinguishing *increases* in organization from simple *changes*.

1.5.2 Uses of the Concept

After its introduction, the main incubators of self-organization were physics, computer science, and the nebulous, ill-fated enterprise of "systems theory". In the physical sciences it was extensively applied, from the 1970s onwards, to pattern formation and spontaneous symmetry breaking (Nicolis and Prigogine 1977) and to cooperative phenomena (Haken 1977). To put it kindly, the real value of these early works was inspiring the production of theories which actually explain things (Manneville 1990; Cross and Hohenberg 1993; Ball 1999). The work of Eigen and Schuster (1979) and of Winfree (1980) were notable exceptions to this rule, since they were both early *and* genuinely explanatory.

Some (Klimontovich 1990/1991) have claimed that the transition from lamellar to turbulent flow is an instance of self-organization; others have just as vigorously denied this; there has been no resolution of the controversy, and no means of resolving it (Frisch 1995). More recently, there has been great interest in the idea that some systems can self-organize into critical states (Bak, Tang and Weisenfield 1987; Jensen 1998). Some people make very grand claims indeed for this idea (Bak 1996); others contend that it hasn't been demonstrated to apply to a single natural or even experimental system. In any case, the dispute does nothing to clarify what "self-organized" means.

Within computer science, the primary applications have been to learning (Selfridge 1959; Yovits and Cameron 1960), especially unsupervised learning (Hinton and Sejnowski 1999) and memory (Kohonen 1984; Kohonen 2001); to adaptation (Holland 1992; Farmer, Lapedes, Packard and Wendroff 1986); and to "emergent" or distributed computation (Forrest 1990; Resnick 1994; Crutchfield and Mitchell 1995). More recently, self-organization has begun to feature in economics (Schelling 1978; Krugman 1996; Shalizi 1996b), and in ecology (Arthur 1990), complete with the now-expected disputes about whether certain processes (such as the succession of plant communities) are self-organizing.

In the 1980s, self-organization became one of the ideas, models and techniques bundled together as the "sciences of complexity" (Pagels 1988) — for good reason, as we'll see when we get to the connection between complexity and organization (Chapter 2). This bundle has been remarkably good at getting itself adopted by at least some researchers in essentially every science, so the idea of self-organization is now used in a huge range of disciplines (see, e.g., Ortoleva 1994), though often not very well (again see, e.g., Ortoleva 1994).

1.5.3 History of Pattern Discovery and Computational Mechanics

I'll close this chapter with a few brief remarks on the histories of pattern discovery and computational mechanics. For more on these matters, especially on techniques akin to computational mechanics, see Shalizi and Crutchfield (2000c) and Chapter 6 below. It might even be a good idea to read this section after reading Chapters 2–4.

The ideal of algorithmic pattern discovery — of automatically constructing a model directly from data, without prior assumptions as to the form of the model — has been the goal, sometimes more or less obscured, of much work in computer science on unsupervised learning. It was very plainly part of the motivation of (Hebb 1949) when he founded the field of unsupervised learning. For reasons which would take too long to go into, however, machine learning changed directions, to become a study of how (in essence) to find the best model from within a given, pre-specified class of models, rather than building a model from the data. Techniques of *system identification* in control theory (Stengel 1994) are similarly limited.

In the 1970s, however, statisticians and information theorists (Akaike 1998; Rissanen 1978; Rissanen 1989) developed model-selection and model-identification techniques which sought to balance accuracy against complexity, defining both information-theoretically. Every stochastic model assigns a certain probability (technically, the *likelihood*) to a given body of data. The classical statistical technique of *maximum likelihood* (Cramér 1945) is simple to select that model from the class of those considered admissible which makes the data most likely. A fundamental result in information theory says that the optimal (minimal mean) length for the bit-string encoding a given signal is proportional to the negative logarithm of the probability of the signal. Maximum likelihood thus corresponds to minimizing the length of the bit string needed to encode the data. The *minimum description-length principle* of Jorma Rissanen says to pick the model which minimizes the sum of this bit string, *plus* the length of the bit string needed to specify the model from within the class of admissible models. This not only generalizes maximum likelihood, it generalizes algorithmic information — by letting us use stochastic models, it lets us describe random process very concisely. This was not yet pattern discovery, however, since the model class often had to be tightly constrained for tractability.

The first sustained effort at what we may reasonable call pattern discovery instead came from statistical physics and nonlinear dynamics. In the early 1980s, dynamicists (Packard, Crutchfield, Farmer and Shaw 1980; Takens 1981) developed techniques for automatically identifying, up to a diffeomorphism, a the attractor of a smooth dynamical system from a time-series of one of its coordinates. Despite occasional abuses, this method, variously known as "attractor reconstruction," "delay embedding," "geometry from a time series," etc., has become the single most valuable tool in experimental dynamics (Kantz and Schreiber 1997).

It was on this background that, in 1986, Peter Grassberger proposed his "true measure complexity," identifying the complexity of a dynamical system with the information needed to specify the state of its optimal predictor. He did not, however, give any indication of how such a predictor and its states might be found, nor even what "optimal prediction," in this sense, might mean. Simultaneously and independently, "geometry from a time series" evolved into "equations of motion from a data series" (Crutchfield and McNamara 1987; Timmer, Rust, Horbelt and Voss 2000). In this method, each small region of the state-space had a vector field of specified functional form fitted to it. The fitting was calculated to preserve the predictive information in the data series, as well as satisfying whatever smoothness constraints were imposed.

The crucial step to computational mechanics was to realize that a "pattern basis" (Crutchfield and Hanson 1993b) could be constructed directly from the data, and that it would give the optimal predictor, as well as the equations of motion. This step was taken more than a decade ago by Crutchfield and Young (1989), who introduced the essential concepts of time-series computational mechanics. Since then, their ideas have been used to analyze many aspects of dynamical systems, such as *intrinsic computation* (a concept introduced by Crutchfield; Crutchfield and Young 1990), multifractal fluctuation statistics (Young and Crutchfield 1993), the automatic construction of Markov partitions for sofic systems (Perry and Binder 1999), stochastic resonance (Witt, Neiman and Kurths 1997) and hidden Markov models (Upper 1997). This part of the theory has been successfully applied to real-world data, from the dripping faucet experiment (Gonçalves, Pinto, Sartorelli and de Oliveira 1998), and from atmospheric turbulence (Palmer, Fairall and Brewer 2000). Feldman and Crutchfield (1998a) extended the theory to equilibrium spin systems. Crutchfield and Hanson (1993b) extended it to such spatial processes as cellular automata. The spatial version of the theory has been used to understand emergent phenomena in cellular automata (Hanson and Crutchfield 1997) and, perhaps most importantly, evolved spatial computation (Crutchfield and Mitchell 1995).

Chapter 2

Measuring Pattern, Complexity and Organization

2.1 Organization

Organized matter is ubiquitous in the natural world, and there is even a branch of physics which studies it: statistical mechanics. But that field has no coherent, principled way of describing, detecting or quantifying the many different kinds of organization found in nature. Statistical mechanics has a good measure of one kind of disorder in thermodynamic entropy, and many people think this will do the job. For instance, Wolfram (1983) and Klimontovich (1990/1991) are among the handful of physicists who are explicit about what they mean by "self-organizing," and both identify it with "decreasing entropy". But thermodynamic entropy fails as a measure of organization in many ways. The most basic problem is that it doesn't distinguish between the many different *kinds* of organization matter can exhibit. Just in equilibrium, a very partial list would include:

- Dilute homogeneous gases;
- Crystals, with many different sorts of symmetry;
- Quasicrystals;
- Low T_c superconductors;
- High T_c superconductors;
- The long-range order of ferromagnets;
- The different long-range order of antiferromagnets;
- The short-range order and long-range disorder of amorphous solids (Zallen 1983) and spin glasses (Fischer and Hertz 1988);
- The partial positional and orientational orders of the many different liquid crystal phases (Collings 1990; de Gennes and Prost 1993);
- The very intricate structures formed by amphiphilic molecules in solution (Gompper and Schick 1994).

Now, statistical mechanics *does* have a procedure for classifying and quantifying these kinds of order. It goes like this (Sethna 1991). A theorist guesses an order parameter, informed by some mixture of what worked in other problems, experimental findings, dubious symmetry arguments and luck. She then further guesses an expansion for the free energy of the system in powers of this order parameter. Finally, if she is very lucky, she not only extracts some quantitative predictions from that expansion, but persuades an experimentalist to test them, at which point they are most likely found to be wrong, and the whole cycle starts over. It is a remarkable testimony to the insight, skill and tenacity of condensed matter theorists, not to mention their sheer numbers, that this works anywhere near as well as it has (Forster 1975; Yeomans 1992; Chaikin and Lubensky 1995).

Despite oft-expressed hopes to the contrary (by Prigogine (1980), Haken (1977), etc.), the ideal of expanding the free energy, or some other Lyapunov functional, in powers of an order parameter fails completely outside of equilibrium (Anderson and Stein 1987; Cross and Hohenberg 1993). This is not to say that the idea of broken symmetry isn't useful in understanding pattern formation, nor that there aren't some techniques (such as "phase equations") which apply to a wide range of pattern-forming systems, and look a bit like what we're used to in equilibrium (Manneville 1990; Cross and Hohenberg 1993). But it is to say that matters are even more *ad hoc*, and there is even less reason to think that current techniques are *universally* applicable. Nobody, for instance, expects to be able to find an order-parameter-type theory of turbulence, though it's obvious to visual inspection (Van Dyke 1982) that turbulent flows *do* have a significant degree of organization, which seems to involve vorticity (Chorin 1994).

Going on beyond conventional condensed matter physics, it is hard to see how any serious case could be made for taking thermodynamic entropy as a measure of *biological* order, though some very great scientists (most notably, Schrödinger) have done so without, it appears, a second thought. Biological systems are open, so what matters, even from the perspective of energetics, is free energy or some other thermodynamic potential, not entropy as such. Worse, there are many biological processes which everyone agrees lead to more organization which are driven by increases in entropy (Fox 1988). Fundamentally, as Medawar (1982, p. 224) nicely put it, "biological order is not, or not merely, unmixedupness." Indeed, he goes on to say that (p. 226)

> In my opinion the audacious attempt to reveal the formal equivalence of the ideas of biological organisation and of thermodynamic order, of non-randomness and information must be judged to have failed. We still seek a theory of order in its most interesting and important form, that which is represented by the complex functional and structural integration of living organisms.

The only thing wrong with this passage is that, as we've just seen, we don't even have a good theory of organization for substances in thermodynamic equilibrium!

If attempts to deal with organization and structure by physicists have been disappointing, at least there have been *some* which are worthy of the name. The literature on biological organization (Lotka 1924; Needham 1936; Needham 1943a; Needham 1943b; Lwoff 1962; Miller 1978; Peacocke 1983; Mittenthal and Baskin 1992; Harrison 1993) consists not so much of theories, as of expressions of more or less intense desires for theories, and more or less clear suggestions for what such theories ought to do — or else they're really not about biological organization at all, but, say, recycled physical chemistry (Peacocke 1983; Harrison 1993). (The work of Fontana and Buss (1994) is a welcome exception.) The best that can be hoped for from this quarter is an array of problems, counter-examples and suggestions, which is not to be sneezed at, but not enough either. In fact, those of us who work on computational mechanics suspect that it could be the basis of a theory of biological order; but that's yet another expression of desire, and not even a suggestion so much as a hint.

Outside biology, attempts to get a grip on what "organization" might or should mean are even fewer, and of even lower quality. There is a large literature in economics and sociology on *organization*, some of which is quite interesting (March and Simon 1993; Arrow 1974; Williamson 1975; Simon 1991). But here "organization" means something like "collection of people with explicitly designated roles and relations of authority", and is contrasted with informal groupings such as "institutions" (Schelling 1978; Elster 1989a; Elster 1989b; Eggertsson 1990; Young 1998; Shalizi 1998b; Shalizi 1999), though both are *organized.*

Bennett (1985, 1986, 1990), apparently in despair, suggested defining "complexity" as whatever increases whenever something self-organizes. The problem with this, as Bennett himself realized, is that it's not at all clear when something self-organizes! But perhaps we can turn this around: for something to self-organize, it must become more complex. Is it possible to come up with a measure of complexity?

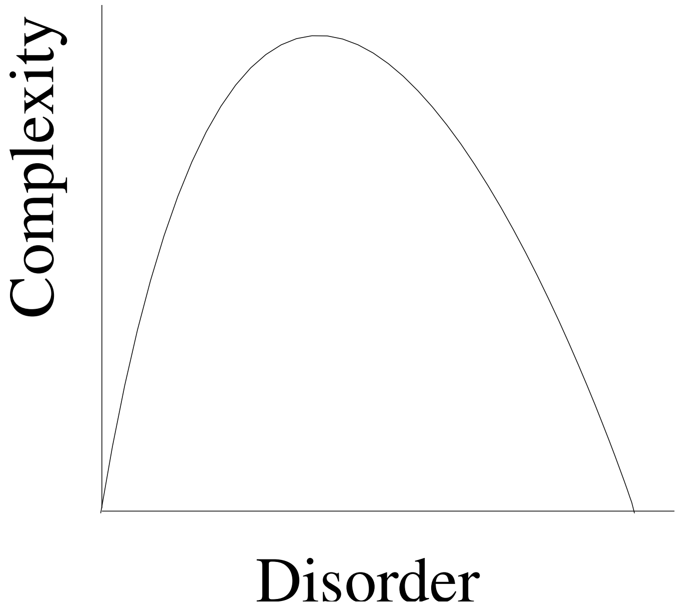

Figure 2.1: Generic complexity vs. disorder curve.

2.2 Complexity Measures, or, the History of One-Humped Curves

It is altogether too easy to come up with complexity measures. An on-line database of them (Edmonds 1997) contains 386 entries, despite not having been updated since 1997, and was surely not comprehensive even then. Every few months seems to produce another paper proposing yet another measure of complexity, generally a quantity which can't be computed for anything you'd actually care to know about, if at all. These quantities are almost never related to any other variable, so they form no part of any theory telling us when or how things get complex, and are usually just quantification for its own sweet sake.

The first and still classic measure of complexity is Kolmogorov's, which is (roughly) the shortest computer program capable of generating a given string. This quantity is in general uncomputable, in the sense that there is simply no algorithm which can calculate it. This comes from a result in computer science known as the halting problem, which in turn is a disguised form of Gödel's theorem, and so is a barrier that will not be overturned any time soon. Moreover, the Kolmogorov complexity is maximized by random strings, so it's really telling us what's random, not what's complex, and it's gradually come to be called the "algorithmic information." It plays a very important role in every discussion of measuring complexity: in a pious act of homage to our intellectual ancestors, it is solemnly taken out, exhibited, and solemnly put away as useless for any practical application.[1]

So we don't want to conflate complexity with randomness, while at the same time we don't want to say that things which are *completely* uniform and orderly are complex either. Complex and interesting stuff should be some place "in the middle". This is conventionally illustrated by a drawing like Figure 2.1. The first such curves to appear in the literature seems to have been the "complexity-entropy diagrams" of Crutchfield and Young (1989). One is reminded of kudzu, which was introduced as a useful plant, and became a weed only through thoughtless replication.

There are an immense number of ways of cooking up curves which look like that, especially since you're free to choose what you mean by "disorder," i.e., what you put on the x axis. A remarkably common prescription is to multiply "disorder" by "one minus disorder," which of course gives a one-humped curve right away (Lopez-Ruiz, Mancini and Calbet 1995; Shiner, Davison and Landsberg 1999). There are two

[1] I perform this ritual in Section 2.3.2 below, with citations.

problems with all such measures. The first is that they don't really agree with us about what things are complex (Solé and Luque 1999; Crutchfield, Feldman and Shalizi 2000a; Binder and Perry 2000). The second is that they are, to use a term introduced by Feldman and Crutchfield (1998b), *over-universal*, failing to distinguish between structurally distinct kinds of organization which just so happen to have the same amount of disorder. In other words, they really don't tell us anything about structure or organization or pattern at all; they just give us a number, which we may admire at our leisure.

It *would* be nice to have a measure of complexity that gave us a one-humped curve, but only if we can do it without cheating, without putting the hump in by hand. And the complexity measure had better not be over-universal — it must distinguish between different kinds of organization; between different patterns.

So how do patterns work?

2.3 Patterns

> These ambiguities, redundancies, and deficiencies recall those attributed by Dr. Franz Kuhn to a certain Chinese encyclopedia entitled *Celestial Emporium of Benevolent Knowledge*. On those remote pages it is written that animals are divided into (a) those that belong to the Emperor, (b) embalmed ones, (c) those that are trained, (d) suckling pigs, (e) mermaids, (f) fabulous ones, (g) stray dogs, (h) those that are included in this classification, (i) those that tremble as if they were mad, (j) innumerable ones, (k) those drawn with a very fine camel's hair brush, (l) others, (m) those that have just broken a flower vase, (n) those that resemble flies from a distance.
> — J. L. Borges (1964, p. 103), "The Analytical Language of John Wilkins"

The passage illustrates the profound gulf between patterns, and classifications derived from patterns, that are appropriate to the world and help us to understand it and those patterns which, while perhaps just as legitimate as logical entities, are not at all informative. What makes the *Celestial Emporium's* scheme inherently unsatisfactory, and not just strange, is that it tells us nothing about animals. We want to find patterns in a process that "divide it at the joints, as nature directs, not breaking any limbs in half as a bad carver might" (Plato, *Phaedrus*, 265D). (Cf. Crutchfield (1992).)

I'm not talking, here, about pattern formation. I'm not even talking about pattern recognition as a practical matter as found in, say, neuropsychology (Luria 1973), psychophysics and perception (Graham 1989), cognitive ethology (Shettleworth 1998), computer programming (Tou and Gonzalez 1974; Ripley 1996), or signal and image processing (Banks 1990; Lim 1990). Instead, I'm asking *what patterns are* and *how patterns should be represented*. I want pattern *discovery*, not pattern *recognition*.

Most of what work there is on what patterns are has been philosophical; the part of it worth bothering with is tied to mathematical logic. Within this, I distinguish two strands. One uses (highly) abstract algebra and the theory of relations; the other, the theory of algorithms and effective procedures.

The general idea, in both approaches, is that some object $\mathcal{O}$ has a pattern $\mathcal{P}$ — $\mathcal{O}$ has a pattern "represented", "described", "captured", and so on by $\mathcal{P}$ — if and only if we can use $\mathcal{P}$ to predict or compress $\mathcal{O}$. The ability to predict implies the ability to compress, but not vice versa, so I'll stick to prediction. The algebraic and algorithmic strands differ mainly on how to represent $\mathcal{P}$ itself.

I should emphasize here that "pattern" in this sense implies a kind of regularity, structure, symmetry, organization, and so on. Ordinary usage sometimes accepts, for example, speaking about the "pattern" of pixels in a particular slice of between-channels video snow; but I'll always call that the *configuration* of pixels.

2.3.1 Algebraic Patterns

Although the problem of pattern discovery appears early, in Plato's *Meno* for example, perhaps the first attempt to make the notion of "pattern" mathematically rigorous was that of Whitehead and Russell in *Principia Mathematica*. They viewed patterns as properties, not of sets, but of relations within or between sets, and accordingly they work out an elaborate *relation-arithmetic* (Whitehead and Russell 1925–27, vol. II, part IV; cf. Russell 1920, ch. 5–6). This starts by defining the *relation-number* of a relation between two sets as the class of all the relations that are equivalent to it under one-to-one, onto mappings of the two sets. In this framework relations share a common pattern or structure if they have the same relation-number. For instance, all square lattices have similar structure since their elements share the same neighborhood relation; as do all hexagonal lattices. Hexagonal and square lattices, however, exhibit different patterns since they have non-isomorphic neighborhood relations — i.e., since they have different relation-numbers. Less work has been done on this than they — especially Russell (1948) — had hoped.

A more recent attempt at developing an algebraic approach to patterns builds on semi-group theory and its Krohn-Rhodes decomposition theorem. Rhodes (1971) discusses a range of applications of this approach to patterns. Along these lines, Rhodes and Nehaniv have tried to apply semi-group complexity theory to biological evolution (Nehaniv and Rhodes 1997). They suggest that the complexity of a biological structure can be measured by the number of subgroups in the decomposition of an automaton that describes the structure.

Yet another algebraic approach has been developed by Grenander and co-workers, primarily for pattern recognition (Grenander 1996). Essentially, this is a matter of trying to invent a minimal set of *generators* and *bonds* for the pattern in question. Generators can adjoin each other, in a suitable n-dimensional space, only if their bonds are compatible. Each pair of compatible bonds specifies at once a binary algebraic operation and an observable element of the configuration built out of the generators. (The construction in Appendix B.2, linking an algebraic operation with concatenations of strings, is analogous in a rough way, as are the "observable operator models" of Jaeger (2000).) Probabilities can be attached to these bonds; these are postulated to be such as to give a Gibbs distribution over entire configurations. Grenander and his colleagues have used these methods to characterize, *inter alia*, several biological phenomena (Grenander, Chow and Keenan 1991; Grenander and Manbeck 1993). While the theory we'll end up with in chapters 4 and 10 could be phrased in terms of generators and bonds, we give a constructive procedure for making them (unlike the trial-and-error approach of Grenander), and our Gibbs distributions are derived, not postulated.

2.3.2 Turing Mechanics: Patterns and Effective Procedures

The other path to patterns follows the traditional exploration of the logical foundations of mathematics, as articulated by Frege and Hilbert and pioneered by Church, Gödel, Post, Russell, Turing, and Whitehead. This relatively more popular approach begins with Kolmogorov and Chaitin, who were interested in the *exact* reproduction of an individual object (Kolmogorov 1965; Chaitin 1966; Kolmogorov 1983; Li and Vitanyi 1993); in particular, they cared about discrete symbol systems, rather than (say) real numbers or smooth vector fields. The candidates for expressing the pattern $\mathcal{P}$ were universal Turing machine (UTM) programs — specifically, the shortest UTM program that can exactly produce the object $\mathcal{O}$. This program's length is called $\mathcal{O}$'s *Kolmogorov-Chaitin complexity*. Note that any scheme — automaton, grammar, or what-not — that is Turing equivalent and for which a notion of "length" is well defined will do as a representational scheme. Since we can convert from one such device to another — say, from a Post tag system (Minsky 1967) to a Turing machine — with only a finite description of the first system, such constants are easily assimilated when measuring complexity in this approach.

In particular, consider the first n symbols $\mathcal{O}_n$ of $\mathcal{O}$ and the shortest program $\mathcal{P}_n$ that produces them. What happens to the limit

$$\lim_{n \to \infty} \frac{|\mathcal{P}_n|}{n} , \tag{2.1}$$

where $|\mathcal{P}|$ is the length in bits of program $\mathcal{P}$? On the one hand, if there is a fixed-length program $\mathcal{P}$ that generates arbitrarily many digits of $\mathcal{O}$, then this limit vanishes. Most of our interesting numbers, rational or irrational — such as 7, π, e, $\sqrt{2}$ — are of this sort. These numbers are eminently compressible: the program $\mathcal{P}$ is the compressed description, and so it captures the pattern obeyed by the sequence describing $\mathcal{O}$. If the limit goes to 1, on the other hand, we have a completely incompressible description and conclude, following Kolmogorov, Chaitin, and others, that $\mathcal{O}$ is random (Kolmogorov 1965; Chaitin 1966; Kolmogorov 1983; Li and Vitanyi 1993; Martin-Löf 1966; Levin 1974). This conclusion is the desired one: the Kolmogorov-Chaitin framework establishes, formally at least, the randomness of an individual object without appeals to probabilistic descriptions or to ensembles of reproducible events. And it does so by referring to a deterministic, algorithmic representation — the UTM.

There are many well-known difficulties with applying Kolmogorov complexity to natural processes. First, as a quantity, it is uncomputable in general, owing to the halting problem (Li and Vitanyi 1993). Second, it is maximal for random sequences; this is either desirable, as just noted, or a failure to capture structure. Third, it only applies to a single sequence; again this can be either good or bad. Fourth, it makes no allowance for noise or error, demanding exact reproduction. Finally, $\lim_{n\to\infty} |\mathcal{P}_n|/n$ can vanish, although the computational resources needed to run the program, such as time and memory, grow without bound.

None of these impediments have kept researchers from attempting to use Kolmogorov-Chaitin complexity for practical tasks — such as measuring the complexity of natural objects (e.g. Gurzadyan (1999)), as a basis for theories of inductive inference (Solomonoff 1964; Vitányi and Li 1999), and generally as a means of capturing patterns (Flake 1998). Rissanen's comments on this can hardly be bettered, so I'll quote him (Rissanen 1989, p. 49):

> It has been sometimes enthusiastically claimed that the algorithmic [i.e., Kolmogorov] complexity provides an ideal solution to the inductive inference problem, and that 'all' we need is to find an approximation to the non-computable algorithmic complexity and use the result to do prediction and the other inferences of interest. Well, this is a tall order, for there is nothing in a universal computer that helps us to find a good model of a string. In fact, if we already know the relevant properties of the string we can always write good programs for it, but we don't learn the properties by writing programs in the hopes of finding short ones!

Some of these difficulties have been addressed by later workers. Bennett's *logical depth*, the number of computational steps the minimal-length program $\mathcal{P}$ needs to produce $\mathcal{O}$, tries to account for time resources (Bennett 1985; Bennett 1986; Bennett 1990). Koppel's *sophistication* attempts to separate out the "regularity" portion of the program from the random or instance-specific input data (Koppel 1987; Koppel and Atlan 1991). Ultimately, however, all these extensions and generalizations remain in the UTM, exact-reproduction setting and so inherit inherent uncomputability. None of them is any good for anything practical.

2.3.3 Patterns with Error

An obvious next step is to allow our pattern $\mathcal{P}$ some degree of approximation or error, in exchange for shorter descriptions. We lose perfect reproduction of the original configuration from the pattern. Given the ubiquity of noise in nature, this is a small price to pay. We might also say that sometimes we are willing to accept small deviations from a regularity, without really caring what the precise deviation is. As many have pointed out (e.g., Crutchfield 1992), this is what we do in thermodynamics, where we throw away vast amounts of useless microscopic detail in order to get workable macroscopic descriptions.

Some interesting philosophical work on patterns-with-error has been done by Dennett, with reference not just to questions about the nature of patterns and their emergence but also to psychology (Dennett 1991). The intuition is that truly random processes can be modeled very simply — to model coin-tossing, toss a coin. Any prediction scheme that is more accurate than assuming complete independence *ipso facto* captures a pattern in the data. There is thus a spectrum of potential pattern-capturers ranging from the assumption of pure noise to the exact reproduction of the data, if that is possible. Dennett notes that there is generally a trade-off between the simplicity of a predictor and its accuracy, and he plausibly describes emergent phenomena (Crutchfield 1994a; Holland 1998) as patterns that allow for a large reduction in complexity for only a small reduction in accuracy[2]. Of course, Dennett was not the first to consider predictive schemes that tolerate error and noise; we'll look at some of the earlier work in Chapter 6. However, to my knowledge, he was the first to have made such predictors a central part of an explicit account of *what patterns are*. His account lacks the mathematical detail of the other approaches we have considered so far, and it relies on the inexact prediction of a single configuration. In fact, it relies on exact predictors that are "fuzzed

[2] I develop this idea quantitatively in Chapter 11.2.

up" by noise. The introduction of noise, however, brings in probabilities, and their natural setting is in ensembles. It is in that setting that the ideas computational mechanics shares with Dennett can receive a proper quantitative treatment, and in which we will see that we don't need to invoke *exact* predictors at all.

2.3.4 Causation

We want our representations of patterns in dynamical processes to be causal — to say how one state of affairs leads to or produces another. Although a key property, causality enters the theory only in an extremely weak sense, the weakest one can use mathematically, which is Hume's (Hume 1739): one class of event causes another if the latter always follows the former; the effect invariably succeeds the cause. As good indeterminists, in the following I replace this invariant-succession notion of causality with a more probabilistic one, substituting a homogeneous distribution of successors for the solitary invariable successor. (A precise statement appears in Definition 13's definition of *causal states*.) This approach results in a purely phenomenological statement of causality, and so it is amenable to experimentation in ways that stronger notions of causality — e.g., that of Bunge (1959) — are not. Salmon (1984) independently reached essentially the same concept of causality by philosophical arguments.

2.3.5 Synopsis of Pattern

We want an approach to patterns which is at once

- *Algebraic*, giving us an explicit breakdown or decomposition of the pattern into its parts;
- *Computational*, showing how the process stores and uses information;
- *Calculable*, analytically or by systematic approximation;
- *Causal*, telling us how instances of the pattern are actually produced; and
- *Naturally stochastic*, not merely tolerant of noise but explicitly formulated in terms of ensembles.

Computational mechanics satisfies all these desiderata.

Chapter 3

The Basic Case of Computational Mechanics: Memoryless Transducers

3.1 The Setup

Consider two discrete random variables X (taking values from $\mathbf{X}$) and Y (taking values in $\mathbf{Y}$)[1]. We think of X as the causes or inputs to some process, and Y as the effects or outputs. Causation is in general stochastic, so we represent this by saying that Y is a random function of X. We assume that Y depends only on the *current* value of X, and not on any previous history of inputs. Let's call something which fits this description a *memoryless transducer*. Many different physical systems are memoryless transducers. So, a little more abstractly, are many problems in biology and social science, e.g., the output might be whether a person dies of lung disease, and the inputs various risk factors (genotype, smoking, working in a mine, whether or not the mine has a union, etc.). The task is to predict Y as well as possible from X. We'd like to know which aspects of the input are relevant to the output, because in general not all of them are, though whether a given feature is relevant can depend on what values other features take on. We want to know all the distinctions we can make about X which make a difference to the distribution of Y.

3.2 Effective States

Any prediction scheme treats some inputs the same when it calculates its predictions. That is, any prediction scheme is sensitive, not to the inputs themselves, but to equivalence classes[2] of inputs. Generally it does so implicitly; but it is much better to be explicit about this.

Definition 1 (Effective States of Memoryless Transducers) *An* effective state *is an equivalence class of inputs. A partition of* $\mathbf{X}$ *is an* effective state class. *For each effective state class, written* $\boldsymbol{\mathcal{R}}$*, there is a function* $\eta : \mathbf{X} \mapsto \boldsymbol{\mathcal{R}}$ *which map the current input into the effective state in which it resides. We write the random variable for the current effective state as* $\mathcal{R}$*, and its realizations as* ρ*;* $\mathcal{R} = \eta(X)$*,* $\rho = \eta(x)$*. When two inputs* x_1, x_2 *belong to the same effective state, we write* $x_1 \sim_\eta x_2$*.*

The collection of all effective state classes is called *Occam's pool*.

At this point, we need a way to measure how well a class of effective states lets us predict the output. The tools to do this are provided by information theory, which is explained in Appendix A.2.

[1] Here, and as nearly as possible through this book, upper-case italic letters will indicate random variables, and lower-case ones their realizations.

[2] For a review of equivalence classes, partitions, and equivalence relations, see Appendix A.1.

Definition 2 (Predictive Power of Effective States) *We measure the predictive power of an effective state class* $\boldsymbol{\mathcal{R}}$ *by the entropy of outputs, conditional on the effective state,* $H[Y|\mathcal{R}]$. $\boldsymbol{\mathcal{R}}$ *has more predictive power than* $\boldsymbol{\mathcal{R}}'$ *if and only if* $H[Y|\mathcal{R}] < H[Y|\mathcal{R}']$.

In general, effective states have less predictive power than the original input.

Lemma 1 (The Old Country Lemma) *For any class of effective states* $\boldsymbol{\mathcal{R}}$,

$$H[Y|\mathcal{R}] \geq H[Y|X] \ . \tag{3.1}$$

Proof. By Eq. A.25, for any function f, $H[Y|f(X)] \geq H[Y|X]$. But for every $\boldsymbol{\mathcal{R}}$, there is an η such that $\mathcal{R} = \eta(X)$. Hence $H[Y|\mathcal{R}] = H[Y|\eta(X)]$, and the lemma is proved.

Remark. The reason this is the "Old Country Lemma" will become clear when we consider its application to time series in Chapter 4.

However, some effective state classes are as predictive as the original inputs; we call such states *prescient.*

Definition 3 (Prescient States for Memoryless Transduction) *A set of states* $\boldsymbol{\mathcal{R}}$ *is* prescient *if and only if it has as much predictive power as the complete input space, i.e., iff* $H[Y|\mathcal{R}] = H[Y|X]$. *We mark prescient states (and sets of states, etc.) by putting a hat over the variables names:* $\widehat{\boldsymbol{\mathcal{R}}}, \widehat{\mathcal{R}}, \widehat{\rho}$, *etc.*

We now establish a link between prescience and the statistical notion of "sufficiency" (explained in Appendix A.5).

Lemma 2 (Prescient States Are Sufficient Statistics) *If* $\widehat{\boldsymbol{\mathcal{R}}}$ *is a prescient class of effective states, then* $\widehat{\mathcal{R}}$ *is a sufficient statistic for predicting* Y*, and vice versa.*

Proof. By the definition of mutual information, $I(Y;\widehat{\mathcal{R}}) = H[Y] - H[Y|\widehat{\mathcal{R}}]$. But, by the definition of prescient states, $H[Y|\widehat{\mathcal{R}}] = H[Y|X]$. So $I(Y;\widehat{\mathcal{R}}) = H[Y] - H[Y|X] = I(Y;X)$. So by Proposition 6, prescient states are sufficient statistics. Essentially the same reasoning run in reverse proves the converse part of the theorem. QED.

3.2.1 Minimality and Prediction

Let's invoke Occam's Razor: "It is vain to do with more what can be done with less" (Ockham 1964). To use the razor, we need to fix what is to be "done" and what "more" and "less" mean. The job we want done is accurate prediction, reducing the conditional entropies $H[Y|\mathcal{R}]$ as far as possible, the goal being to attain the bound set by Lemma 1, with a prescient set of states. But we want to do this as simply as possible, with as few resources as possible. We already have a measure of uncertainty, so we need a measure of resources. Since there is a probability measure over inputs, there is an induced measure on the η-states.[3] Accordingly, we define the following measure of complexity.

Definition 4 (Statistical Complexity of States) *The statistical complexity of a class* $\boldsymbol{\mathcal{R}}$ *of states is*

$$C_\mu(\boldsymbol{\mathcal{R}}) \quad \equiv \quad H[\mathcal{R}] \ . \tag{3.2}$$

The μ in C_μ reminds us that it is a measure-theoretic property and depends ultimately on the distribution over the inputs, which induces a measure over states.

The statistical complexity of a state class is the average uncertainty (in bits) in the transducer's current state. This, in turn, is the same as the average amount of memory (in bits) that the transducer *appears*

[3]This assumes η is at least nearly measurable. See Appendix B.3.2.2.

to retain about the input, given the chosen state class $\boldsymbol{\mathcal{R}}$. (We will later see how to define the statistical complexity of the transducer itself.) The goal is to do with as little of this memory as possible. Restated then, we want to minimize statistical complexity, subject to the constraint of maximally accurate prediction.

The idea behind calling the collection of all partitions of $\mathbf{X}$ Occam's pool should now be clear: One wants to find the shallowest point in the pool. This we now do.

3.3 Causal States

Definition 5 (Causal State for Memoryless Transduction) *The causal state are the range of the function*

$$\epsilon(x) = \{x' | \mathrm{P}(Y|X = x) = \mathrm{P}(Y|X = x')\} \ . \tag{3.3}$$

If $\epsilon(x) = \epsilon(x')$*, then* $\mathrm{P}(Y|X = x) = \mathrm{P}(Y|X = x')$*, and we write* $x \sim_\epsilon x'$*. We denote the class of causal states by* $\boldsymbol{\mathcal{S}}$*, the random variable for the current causal state by* $\mathcal{S}$*, and a particular causal state by* σ*.*

Each causal state σ has a unique associated distribution of outputs $\mathrm{P}(Y = y|\mathcal{S} = \sigma)$, called its *morph*. In general every *effective* state has a morph, but two effective states in the same state class may very well have the same morph. Moreover, the causal states have the important property that all of their parts have the same morph. We make this notion a little more precise in the following definitions, which build to important results later on, especially the crucial Refinement Lemma (Lemma 4).

3.3.1 Homogeneity

The following definitions are inspired by Salmon (1984).

Definition 6 (Strict Homogeneity) *A set* $\mathbf{X}$ *is* strictly homogeneous *with respect to a random variable* Y *when the conditional distribution for* Y*,* $\mathrm{P}(Y|\mathbf{X})$*, is the same for all measurable subsets of* $\mathbf{X}$*.*

Definition 7 (Weak Homogeneity) *A set* $\mathbf{X}$ *is* weakly homogeneous *with respect to* Y *if* $\mathbf{X}$ *is not strictly homogeneous with respect to* Y*, but* $\mathbf{X} \setminus \mathbf{X}_0$ *(*$\mathbf{X}$ *with* $\mathbf{X}_0$ *removed) is, where* $\mathbf{X}_0$ *is a subset of* $\mathbf{X}$ *of measure* 0*.*

Lemma 3 (Strict Homogeneity of Causal States) *A process's causal states are the largest subsets of inputs that are all strictly homogeneous with respect to the output.*

Proof. We must show that, first, the causal states are strictly homogeneous with respect to output and, second, that no larger strictly homogeneous subsets of inputs could be made. The first point, the strict homogeneity of the causal states, is evident from Definition 5: By construction, all elements of a causal state have the same conditional distribution for the output, so any part of a causal state will have the conditional distribution as the whole state. The second point likewise follows from Definition 5, since the causal state contains *all* the inputs with a given conditional distribution of output. Any other set strictly homogeneous with respect to output must be smaller than a causal state, and any set that includes a causal state as a proper subset cannot be *strictly* homogeneous. QED.

3.3.2 Optimalities and Uniqueness

Let's see what the causal states are good for. Let's start by seeing how well we can predict the output from knowing the causal state.

Theorem 1 (Prescience and Sufficiency of Causal States) *The causal states* $\boldsymbol{\mathcal{S}}$ *are prescient, and sufficient statistics.*

Proof. It is clear that $\mathrm{P}(Y = y|\mathcal{S} = \epsilon(x)) = \mathrm{P}(Y = y|X = x)$, for all x, y. Thus, by Definition 66, the causal states are sufficient statistics for predicting the output, and so, by Lemma 2, they are prescient.

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

Proof. We invoke a straightforward extension of Theorem 2.7.3 of Cover and Thomas (1991): If $X_1, X_2, \ldots X_n$ are random variables over the same set $\mathcal{A}$, each with distinct probability distributions, Θ a random variable over the integers from 1 to n such that $\mathrm{P}(\Theta = i) = \lambda_i$, and Z a random variable over $\mathcal{A}$ such that $Z = X_\Theta$, then

$$\begin{aligned} H[Z] &= H\left[\sum_{i=1}^{n} \lambda_i X_i\right] \\ &\geq \sum_{i=1}^{n} \lambda_i H[X_i] \ . \end{aligned} \tag{3.4}$$

In words, the entropy of a mixture of distributions is at least the mean of the entropies of those distributions. This follows since H is strictly concave, which in turn follows from $x \log x$ being strictly convex for $x \geq 0$. We obtain equality in Eq. 3.4 if and only if all the λ_i are either 0 or 1, i.e., if and only if Z is at least weakly homogeneous (Definition 7).

The conditional distribution of outputs for each rival state ρ can be written as a weighted mixture of the distributions of one or more causal states. Thus, by Eq. 3.4, unless every ρ is at least weakly homogeneous with respect to outputs, the entropy of Y conditioned on $\mathcal{R}$ will be higher than the minimum, the entropy conditioned on $\mathcal{S}$. So, in the case of the maximally predictive $\widehat{\mathcal{R}}$, every $\widehat{\rho} \in \widehat{\boldsymbol{\mathcal{R}}}$ must be at least weakly homogeneous with respect to Y. But the causal states are the largest classes that are strictly homogeneous with respect to Y (Lemma 7). Thus, the strictly homogeneous part of each $\widehat{\rho} \in \widehat{\boldsymbol{\mathcal{R}}}$ must be a subclass, possibly improper, of some causal state $\sigma \in \boldsymbol{\mathcal{S}}$. QED.

Remark 1. One can provide a more elaborately algebraic and less verbal proof of this Lemma. We do this for the case of time series in Appendix B.4, but the reader may easily adapt the argument there to this simpler case.

Remark 2. The content of the lemma can be made quite intuitive, if we ignore for a moment the measure-0 set $\widehat{\rho}_0$ of inputs mentioned in its statement. It then asserts that any alternative partition $\widehat{\boldsymbol{\mathcal{R}}}$ that is as prescient as the causal states must be a refinement of the causal-state partition. That is, each $\widehat{\mathcal{R}}_i$ must be a (possibly improper) subset of some $\mathcal{S}_j$. Otherwise, at least one $\widehat{\mathcal{R}}_i$ would have to contain parts of at least two causal states. And so, using this $\widehat{\mathcal{R}}_i$ to predict the output would lead to more uncertainty about Y than using the causal states.

Adding the measure-0 set $\widehat{\rho}_0$ of inputs to this picture does not change its heuristic content much. Precisely because these inputs have zero probability, treating them in the wrong way makes no discernible difference to predictions, morphs, and so on. There is a problem of terminology, however, since there seems to be no standard name for the relationship between the partitions $\widehat{\boldsymbol{\mathcal{R}}}$ and $\boldsymbol{\mathcal{S}}$. We propose to say that the former is a refinement of the latter *almost everywhere* or, simply, a *refinement a.e.*

Remark 3. One cannot work the proof the other way around to show that the causal states have to be a refinement of the equally prescient $\widehat{\mathcal{R}}$-states. This is because the theorem borrowed from Cover and Thomas (1991), Eq. 3.4 only applies when we can reduce uncertainty by specifying from *which* distribution one chooses. Since the causal states are constructed so as to be strictly homogeneous with respect to futures, this is not the case. Lemma 3 and Theorem 1 together protect us.

Remark 4. Because almost all of each prescient rival state is wholly contained within a single causal state, we can construct a function $g : \widehat{\boldsymbol{\mathcal{R}}} \mapsto \boldsymbol{\mathcal{S}}$, such that, if $\eta(x) = \widehat{\rho}$, then $\epsilon(x) = g(\widehat{\rho})$ almost always. We can even say that $\mathcal{S} = g(\widehat{\mathcal{R}})$ almost always, with the understanding that this means that, for each $\widehat{\rho}$, $\mathrm{P}(\mathcal{S} = g(\widehat{\rho})|\widehat{\mathcal{R}} = \widehat{\rho}) = 1$.

Theorem 2 (Minimality of Causal States) *For any prescient state class $\widehat{\boldsymbol{\mathcal{R}}}$,*

$$C_\mu(\widehat{\boldsymbol{\mathcal{R}}}) \geq C_\mu(\boldsymbol{\mathcal{S}}) \ . \tag{3.5}$$

Proof. By Lemma 4, Remark 4, there is a function g such that $\mathcal{S} = g(\widehat{\mathcal{R}})$ almost always. But $H[f(X)] \leq H[X]$ (Eq. A.22) and so

$$H[\mathcal{S}] = H[g(\widehat{\mathcal{R}})] \leq H[\widehat{\mathcal{R}}] \ . \tag{3.6}$$

but $C_\mu(\widehat{\boldsymbol{\mathcal{R}}}) = H[\widehat{\mathcal{R}}]$ (Definition 4). QED.

Remark. If the distribution over inputs $\mathrm{P}(X)$ changes, but, for each x, but the conditional distribution of outputs $\mathrm{P}(Y|X = x)$ does not, the causal states also do not change. In general, the numerical value of the statistical complexity of the causal states *will* change, but their minimality among the prescient states will not.

Corollary 1 (Causal States Are Minimal Sufficient) *The causal states are minimal sufficient statistics for predicting the output.*

Proof. We saw in the proof of Theorem 2 how to construct a function from any prescient state class to the causal states. From Lemma 2, the prescient state classes consist of all and only the predictively sufficient statistics. Therefore, the causal states are functions of all the sufficient statistics, and so by Definition 67, they are the minimal sufficient statistics.

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

Proof. From Lemma 4, we know that $\mathcal{S} = g(\widehat{\mathcal{R}})$ almost always. We now show that there is a function f such that $\widehat{\mathcal{R}} = f(\mathcal{S})$ almost always, implying that $g = f^{-1}$ and that f is the desired relation between the two sets of states. To do this, by Eq. A.23 it is sufficient to show that $H[\widehat{\mathcal{R}}|\mathcal{S}] = 0$. Now, it follows from an information-theoretic identity (Eq. A.19) that

$$H[\mathcal{S}] - H[\mathcal{S}|\widehat{\mathcal{R}}] = H[\widehat{\mathcal{R}}] - H[\widehat{\mathcal{R}}|\mathcal{S}] \ . \tag{3.7}$$

Since, by Lemma 4 $H[\mathcal{S}|\widehat{\mathcal{R}}] = 0$, both sides of Eq. 3.7 are equal to $H[\mathcal{S}]$. But, by hypothesis, $H[\widehat{\mathcal{R}}] = H[\mathcal{S}]$. Thus, $H[\widehat{\mathcal{R}}|\mathcal{S}] = 0$ and so there exists an f such that $\widehat{\mathcal{R}} = f(\mathcal{S})$ almost always. We have then that $f(g(\widehat{\mathcal{R}})) = \widehat{\mathcal{R}}$ and $g(f(\mathcal{S})) = \mathcal{S}$, so $g = f^{-1}$. This implies that f preserves equivalence of states almost always: for almost all $x, x' \in \mathbf{X}$, $\eta(x) = \eta(x')$ if and only if $\epsilon(x) = \epsilon(x')$. QED.

Remark. As in the case of the Refinement Lemma 4, on which the theorem is based, the measure-0 caveats seem unavoidable. A rival that is as predictive and as simple (in the sense of Definition 4) as the causal states, can assign a measure-0 set of inputs to different states than ϵ does, but no more. This makes sense: such a measure-0 set makes no difference, since its members are never observed, by definition. By the same token, however, nothing prevents a minimal, prescient rival from disagreeing with the causal states on those inputs.

Definition 8 (Statistical Complexity of Memoryless Transduction) *The statistical complexity of a transduction process, written simply C_μ, is equal to the statistical complexity of its causal states, $C_\mu(\boldsymbol{\mathcal{S}}) = H[\mathcal{S}]$.*

Remark. This definition is motivated by the minimal statistical complexity of the causal states, and by their uniqueness.

Theorem 4 (Control Theorem for Memoryless Transduction) *For any set of effective states $\boldsymbol{\mathcal{R}}$, the reduction in the uncertainty of the outputs, conditional on knowing the effective state, $H[Y] - H[Y|\mathcal{R}]$, is at most C_μ.*

Proof. This one, honestly, is simple.

$$\begin{aligned} H[Y] - H[Y|\mathcal{R}] &\leq H[Y] - H[Y|\mathcal{S}] && (3.8) \\ &= I(Y;\mathcal{S}) && (3.9) \\ &= H[\mathcal{S}] - H[\mathcal{S}|Y] && (3.10) \\ &\leq H[\mathcal{S}] = C_\mu && (3.11) \end{aligned}$$

QED.

Remark. This result is inspired by, and is a version of, Ashby's "Law of Requisite Variety" (Ashby 1956, ch. 11), which states that applying a controller can reduce the uncertainty in the controlled variable by at most the entropy of the control variable. (Touchette and Lloyd (1999) recently restated this result, without credit to Ashby.) Our control theorem is a statement about the degree of control we can exert over the output by fixing the input, and so the causal state. Note that the inequality will be saturated if $H[\mathcal{S}|Y] = 0$, which will be the case if each output is due to a unique causal state. Since this can't be ruled out *a priori*, we cannot, in general, sharpen the upper bound any further.

3.4 Other Approaches to Memoryless Transduction

This is, of course, a very old, very general and very important problem. In recent years a wide array of methods have arisen for tackling it. We consider here three which are particularly akin to computational mechanics.

3.4.1 Graphical Models

Some of the most widely applied methods for this problem are those that travel under the label of "graphical models" (Loehlin 1992; Lauritzen 1996; Pearl 2000; Spirtes, Glymour and Scheines 2001). These involve representing the input and the output as a number of distinct variables (one for each quantity we can measure, essentially), and positing a number of hidden or "latent" variables in between. Each variable, manifest or latent, is represented by a node in a graph. A directed edge runs from variable A to variable B if and only if A is a direct cause of B. Assuming that what's called the "causal Markov condition" is met[4] and some other, more technical requirements are satisfied, reliable techniques exist for inferring which variables cause which, and through what intermediate, latent variables.

While these methods are ideologically akin to computational mechanics (Spirtes, Glymour and Scheines (2001) in particular), they are not quite the same. In particular, they do not seek to directly partition the space of inputs $\mathbf{X}$ into the divisions which are relevant to the output; at best this is implicit in the structure of connections between the manifest inputs and the latent variables. Moreover, mathematical tractability generally restricts practitioners to fairly simple forms of dependence between variables, often even to linearity. Our method does not labor under these restrictions. It begins directly with a partition of the input space, to which everything is referred. In effect, the computational mechanics approach is to always construct a graph with only three variables, the input, the causal state, and the output, connected in that order. The work comes in constructing the middle part!

[4] Consider any variable A in the graph $\mathbf{G}$. Write the set of variables which are direct causes of A as $\mathbf{C}(A)$. Write the set of variables which are effects of A, whether direct or indirect, as $\mathbf{E}(A)$, i.e., $B \in \mathbf{E}(A)$ if and only if there is a path from A to B. Finally, let $\mathbf{N}(A) = \mathbf{G} \setminus (\mathbf{C}(A) \cup \mathbf{E}(A))$. Then the causal Markov condition is that A is conditionally independent of all variables in $\mathbf{N}(A)$ given $\mathbf{C}(A)$, that $A \perp\!\!\!\perp \mathbf{N}(A) | \mathbf{C}(A)$.

3.4.2 The Information-Bottleneck Method

Tishby, Pereira and Bialek (1999) poses the following problem. Given a joint distribution over the input X and the output Y, find an intermediate or "bottleneck" variable $\tilde{X}$ which is a (possibly stochastic) function of X such that $\tilde{X}$ is more compressed than X, but retains predictive information about Y. More exactly, they ask for a conditional distribution $\mathrm{P}(\tilde{X} = \tilde{x}|X = x)$ that minimizes the functional

$$\mathcal{F} = I(\tilde{X}; X) - \beta I(\tilde{X}; Y) \ , \tag{3.12}$$

where β is a positive real number. Minimizing the first term represents the desire to find a compression of the original input data X; maximizing the second term represents the desire to retain the ability to predict Y.[5] The coefficient β governs the trade-off between these two goals: as $\beta \to 0$, we lose interest in prediction in favor of compression; whereas as $\beta \to \infty$, predictive ability becomes paramount.

Extending classical rate-distortion theory, the authors are not only able to state self-consistent equations that determine which distributions satisfy this variational problem, but give a convergent iterative procedure that finds one of these distributions. They do not address the rate of convergence.

Now, $I(Y; \tilde{X}) = H[Y] - H[Y|\tilde{X}]$. Since $H[Y]$ is fixed, maximizing $I(Y; \tilde{X})$ is the same as minimizing $H[Y|\tilde{X}]$. That is, to maximize the predictive information, the bottleneck variable should be prescient. But the most compressed prescient states — the ones with the smallest entropy — are the causal states. Thus, they are precisely what should be delivered by the information-bottleneck method in the limit where $\beta \to \infty$. It is not immediately obvious that the iterative procedure of Tishby, Pereira and Bialek (1999) is still valid in this limit. Nonetheless, that ϵ is the partition satisfying their original constraints is evident.

We note in passing that Tishby, Pereira and Bialek (1999) assert that, when sufficient statistics exist, then compression-with-prediction is possible. Conversely, we have shown that the causal states are always sufficient statistics.

3.4.3 The Statistical Relevance Basis

Here is one last solution to the problem of discovering concise and predictive hidden variables. In his books of 1971 and 1984, Wesley Salmon put forward a construction, under the name of the "statistical relevance basis", that is identical in its essentials with that of causal states for memoryless transducers.[6] Owing to the rather different aims for which Salmon's construction was intended — explicating the notion of "causation" in the philosophy of science — no one seems to have proved its information-theoretic optimality properties nor even to have noted its connection to sufficient statistics. Briefly: if a nontrivial sufficient partition of the input variables exists, then the relevance basis is the minimal sufficient partition.

3.5 Summary

Let's recap what we've done in this chapter, since we're going to be going through a similar exercise over and over again.

We start with one variable (or set of variables) which causes, in some statistical fashion, another variable. We want to predict the output, given the input, as accurately and as simply as possible. We summarize the input in an effective state, and measure predictive power by the entropy of the output conditional on the effective state, and the complexity of the predictor by the entropy of the effective state, i.e., the amount of information the state retains from the input. The predictive power of effective states is limited by that of the original input; states which attain this limit are prescient. Our goal is to minimize complexity, subject to the constraint of prescience.

[5] Since $\tilde{X} = g(X, \Omega)$ for some auxiliary random variable Ω, a theorem of Shannon's assures us that $I(\tilde{X}; Y) \leq I(X; Y)$ and the transformation from X to $\tilde{X}$ cannot *increase* our ability to predict Y (Shannon 1948, App. 7).

[6] I discovered Salmon's work by accident in May 1998, browsing in a used book store, so it's not cited in computational mechanics papers up to and including Crutchfield and Shalizi (1999).

We introduce a particular partition of the inputs, the causal states, which treats inputs as equivalent if they lead to the same conditional distribution of outputs. This is prescient, since the distribution of outputs conditional on the causal state is, by construction, the same as that conditional on the input. We then use homogeneity to prove a refinement lemma, telling us that any prescient rival to the causal states must be a refinement of them almost everywhere. The refinement lemma, in turn, leads directly to the result that the causal states are the minimal prescient states, and to the uniqueness of the causal states.

The bulk of the work in the rest of this book will be setting up these same tricks for processes which are more subtle than memoryless transduction, and examining the extra implications for the causal states of those subtleties. We start with time series.

Chapter 4

Computational Mechanics of Time Series

> The next chapter is devoted to the statistical mechanics of time series. This is another field in which conditions are very remote from those of the statistical mechanics of heat engines and which is thus very well suited to serve as a model of what happens in the living organism.
> —Norbert Wiener (1961, p. 59)

4.1 Paddling Around in Occam's Pool

4.1.1 Processes

Let's restrict ourselves to discrete-valued, discrete-time stationary stochastic processes. (See Section 11.4 for ways in which these restrictions might be lifted.) Intuitively, such processes are sequences of random variables S_i, the values of which are drawn from a countable set $\mathcal{A}$. We let i range over all the integers, and so get a bi-infinite sequence

$$\overleftrightarrow{S} = \ldots S_{-1} S_0 S_1 \ldots \ . \tag{4.1}$$

In fact, we can define a process in terms of the distribution of such sequences (cf. Billingsley 1965; Gray 1990).

Definition 9 (A Process) *Let $\mathcal{A}$ be a countable set. Let $\Omega = \mathcal{A}^{\mathbb{Z}}$ be the set of bi-infinite sequences composed from $\mathcal{A}$, $T_i : \Omega \mapsto \mathcal{A}$ be the measurable function that returns the i^{th} element s_i of a bi-infinite sequence $\omega \in \Omega$, and $\mathcal{F}$ the σ-algebra of cylinder sets of Ω. Adding a probability measure* P *gives us a probability space $(\Omega, \mathcal{F}, \mathrm{P})$, with an associated random variable $\overleftrightarrow{S}$. A* process *is a sequence of random variables $S_i = T_i(\overleftrightarrow{S}), i \in \mathbb{Z}$.*

It follows from Definition 9 that there are well defined probability distributions for sequences of every finite length. Let $\overrightarrow{S}^L_t$ be the sequence of $S_t, S_{t+1}, \ldots, S_{t+L-1}$ of L random variables beginning at S_t. $\overrightarrow{S}^0_t \equiv \emptyset$, the null sequence. Likewise, $\overleftarrow{S}^L_t$ denotes the sequence of L random variables going up to S_t, but not including it; $\overleftarrow{S}^L_t = \overrightarrow{S}^L_{t-L}$. Both $\overrightarrow{S}^L_t$ and $\overleftarrow{S}^L_t$ take values from $s^L \in \mathcal{A}^L$. Similarly, $\overrightarrow{S}_t$ and $\overleftarrow{S}_t$ are the semi-infinite sequences starting from and stopping at t and taking values $\overrightarrow{s}$ and $\overleftarrow{s}$, respectively.

Intuitively, we can imagine starting with distributions for finite-length sequences and extending them gradually in both directions, until the infinite sequence is reached as a limit. While this can be a useful picture to have in mind, defining a process in this way raises some subtle measure-theoretic issues, such as how distributions over finite-length sequences limit on the infinite-length distribution. To evade these questions, we *start* with the latter, and obtain the former by "marginalization". (The first chapter of Gray (1990) has a particularly clear exposition of this approach.)

Definition 10 (Stationarity) *A process S_i is* stationary *if and only if*

$$\mathrm{P}(\overrightarrow{S}_t^L = s^L) = \mathrm{P}(\overrightarrow{S}_0^L = s^L) \ , \tag{4.2}$$

for all $t \in \mathbb{Z}$, $L \in \mathbb{Z}^+$, and all $s^L \in \mathcal{A}^L$.

In other words, a stationary process is one that is time-translation invariant. Consequently, $\mathrm{P}(\overrightarrow{S}_t = \overrightarrow{s}) = \mathrm{P}(\overrightarrow{S}_0 = \overrightarrow{s})$ and $\mathrm{P}(\overleftarrow{S}_t = \overleftarrow{s}) = \mathrm{P}(\overleftarrow{S}_0 = \overleftarrow{s})$, and so I'll drop the subscripts from now on.

I'll call $\overleftarrow{S}$ and $\overleftarrow{S}^L$ *pasts* or *histories* and $\overrightarrow{S}$ and $\overrightarrow{S}^L$, *futures.* I'll need to refer to the class of all measurable sets of histories; this will be $\mu(\overleftarrow{S})$[1] Similarly, the class of all measurable sets of futures is $\mu(\overrightarrow{S})$. It is readily checked (Upper 1997) that $\mu(\overleftarrow{S}) = \bigcup_{L=1}^{\infty} \mu(\overleftarrow{S}^L)$, and likewise for $\mu(\overrightarrow{S})$.

This is a good place to note that strict stationarity, as defined above, is actually a stronger property than this chapter needs. All we really require is that $\mathrm{P}(\overrightarrow{S}_t \in F | \overleftarrow{S}_t = \overleftarrow{s}) = \mathrm{P}(\overrightarrow{S}_0 \in F | \overleftarrow{S}_0 = \overleftarrow{s})$, for all t and for all $F \in \mu(\overrightarrow{S})$. This property, of time-invariant transition probabilities, should I guess be named some form of "homogeneity," by analogy with the corresponding property for Markov processes, but that name is pre-empted. So let's call this *conditional stationarity* instead.

4.1.2 The Pool

Our goal is to predict all or part of $\overrightarrow{S}$ using some function of some part of $\overleftarrow{S}$. As before, let's start with effective states, and classes of effective states.

Definition 11 (Effective States) *A partition of $\overleftarrow{\mathbf{S}}$ is an* effective state class. *Each $\rho \in \boldsymbol{\mathcal{R}}$ will be called a* state *or an* effective state. *When the current history $\overleftarrow{s}$ is included in the set ρ, the process is in state ρ. Define a function η from histories to effective states:*

$$\eta : \overleftarrow{\mathbf{S}} \mapsto \boldsymbol{\mathcal{R}} \ . \tag{4.3}$$

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

It makes little difference whether one thinks of η as being a function from a history to a subset of histories or a function from a history to the *label* of that subset. Each interpretation is convenient at different times, and I'll use both.

We could use *any* function defined on $\overleftarrow{\mathbf{S}}$ to partition that set, by assigning to the same ρ all the histories $\overleftarrow{s}$ on which the function takes the same value. Similarly, any equivalence relation on $\overleftarrow{\mathbf{S}}$ partitions it. (See Appendix A.1 for more on equivalence relations.) Due to the way I defined a process's distribution, each effective state has a well-defined distribution of futures[2], though other states could have the same conditional distribution. Specifying the effective state thus amounts to making a prediction about the process's future. All the histories belonging to a given effective state are treated as *equivalent for purposes of predicting the future.* (In this way, the framework formally incorporates traditional methods of time-series analysis; see Section 6.1.)

The definition of statistical complexity, Definition 4, applies unchanged to time series effective states.

Call the collection of all partitions $\boldsymbol{\mathcal{R}}$ of the set of histories $\overleftarrow{\mathbf{S}}$ *Occam's pool.*

[1]Conventionally, this ought to be $\sigma(\overleftarrow{S})$, but, as the reader will see, that notation would be confusing later on.

[2]This is not true if η is not at least nearly measurable (see Appendix B.3.2.2). To paraphrase Schutz (1980), you should assume that all my effective-state functions are sufficiently tame, measure-theoretically, that whatever induced distributions I invoke will exist.

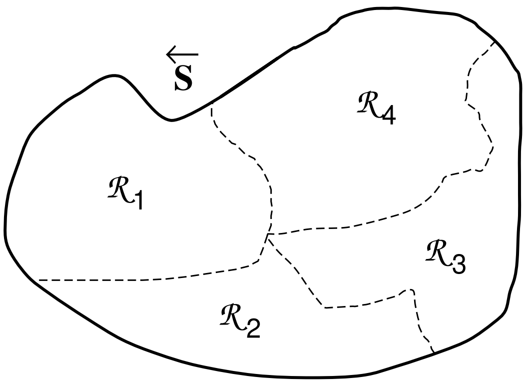

Figure 4.1: A schematic picture of a partition of the set $\overleftarrow{\mathbf{S}}$ of all histories into some class of effective states: $\boldsymbol{\mathcal{R}} = \{\mathcal{R}_i : i = 1, 2, 3, 4\}$. Note that the $\mathcal{R}_i$ need not form compact sets; they're drawn that way for clarity. Imagine Cantor sets or other, more pathological, structures.

4.1.3 Patterns in Ensembles

It will be convenient to have a way of talking about the uncertainty of the future. Intuitively, this would just be $H[\overrightarrow{S}]$, but in general that quantity is infinite and awkward to manipulate. (The special case in which $H[\overrightarrow{S}]$ is finite is dealt with in Appendix B.5.) Normally, I'll evade this by considering $H[\overrightarrow{S}^L]$, the uncertainty of the next L symbols, treated as a function of L. On occasion, I'll refer to the entropy per symbol or *entropy rate* (Shannon 1948; Cover and Thomas 1991):

$$h[\overrightarrow{S}] \equiv \lim_{L\to\infty} \frac{1}{L} H[\overrightarrow{S}^L] \ , \tag{4.4}$$

and the *conditional entropy rate*,

$$h[\overrightarrow{S}|X] \equiv \lim_{L\to\infty} \frac{1}{L} H[\overrightarrow{S}^L|X] \ , \tag{4.5}$$

where X is some random variable and the limits exist. For stationary stochastic processes, the limits always exist (Cover and Thomas 1991, Theorem 4.2.1, p. 64).

These entropy rates are also always bounded above by $H[S]$; which is a special case of Eq. A.14. Moreover, if $h[\overrightarrow{S}] = H[S]$, the process consists of independent variables — independent, identically distributed (IID) variables, in fact, for stationary processes.

Definition 12 (Capturing a Pattern) $\boldsymbol{\mathcal{R}}$ captures a pattern *if and only if there exists an L such that*

$$H[\overrightarrow{S}^L|\mathcal{R}] < LH[S] \ . \tag{4.6}$$

This says that $\boldsymbol{\mathcal{R}}$ captures a pattern when it tells us something about how the distinguishable parts of a process affect each other: $\boldsymbol{\mathcal{R}}$ exhibits their dependence. (I'll also speak of η, the function associated with pasts, as capturing a pattern, since this is implied by $\boldsymbol{\mathcal{R}}$ capturing a pattern.) Supposing that these parts *do not* affect each other, then we have IID random variables, which is as close to the intuitive notion of "patternless" as one is likely to state mathematically. Note that, because of the independence bound on joint entropies (Eq. A.14), if the inequality is satisfied for some L, it is also satisfied for every $L' > L$. Thus, the difference $H[S] - H[\overrightarrow{S}^L|\mathcal{R}]/L$, for the smallest L for which it is nonzero, is the *strength of the pattern* captured by $\boldsymbol{\mathcal{R}}$. Let's now mark an upper bound (Lemma 5) on the strength of patterns; later we'll see how to attain this upper bound (Theorem 5).

4.1.4 The Lessons of History

We are now in a position to prove a result about patterns in ensembles that will be useful in connection with later theorems about causal states.

Lemma 5 (Old Country Lemma) *For all* $\boldsymbol{\mathcal{R}}$ *and for all* $L \in \mathbb{Z}^+$,

$$H[\overrightarrow{S}^L|\mathcal{R}] \geq H[\overrightarrow{S}^L \mid \overleftarrow{S}] \ . \tag{4.7}$$

Proof. By construction (Eq. 4.3), for all L,

$$H[\overrightarrow{S}^L|\mathcal{R}] = H[\overrightarrow{S}^L|\eta(\overleftarrow{S})] \ . \tag{4.8}$$

But

$$H[\overrightarrow{S}^L|\eta(\overleftarrow{S})] \geq H[\overrightarrow{S}^L \mid \overleftarrow{S}] \ , \tag{4.9}$$

since the entropy conditioned on a variable is never more than the entropy conditioned on a function of the variable (Eq. A.25). QED.

Remark 1. That is, conditioning on the whole of the past reduces the uncertainty in the future to as small a value as possible. Carrying around the whole semi-infinite past is rather bulky and uncomfortable and is a somewhat dismaying prospect. Put a bit differently: we want to forget as much of the past as possible and so reduce its burden. It is the contrast between this desire and the result of Eq. 4.7 that leads me to call this the *Old Country Lemma*.

Remark 2. Lemma 5 establishes the promised upper bound on the strength of patterns: viz., the strength of the pattern is at most $H[S] - H[\overrightarrow{S}^L \mid \overleftarrow{S}]/L_{past}$, where L_{past} is the least value of L such that $H[\overrightarrow{S}^L \mid \overleftarrow{S}] < LH[S]$.

4.2 The Causal States

Here I'm going to define the causal states for stochastic processes, very much as I did in the last chapter for memoryless transducers. As was the case there, the definitions and constructions in this section use conditional probabilities over and over again. That's fine so long as I condition on events of nonzero probability. However, I need to condition on events, such as particular histories, whose probability generally *is* zero. There are standard ways of dealing with this, but their technicalities tend to obscure the main lines of the argument. To keep those lines as clear as possible, in this section I state my definitions as though classical conditional probability was adequate, reserving the measure-theoretic treatment, and its limitations and caveats, for Appendix B.3. The proofs are compatible with the proper use of conditional measures, but they should be intelligible without them.

Definition 13 (A Process's Causal States) *The* causal states *of a process are the members of the range of the function* ϵ *that maps from histories to sets of histories:*

$$\epsilon(\overleftarrow{s}) \ \equiv \{\overleftarrow{s}' \ \ |\mathrm{P}(\overrightarrow{S} \in F| \overleftarrow{S} = \overleftarrow{s}) = \mathrm{P}(\overrightarrow{S} \in F| \overleftarrow{S} = \overleftarrow{s}') \ , \forall \ F \in \mu(\overrightarrow{S}), \overleftarrow{s}' \in \overleftarrow{S}\} \ , \tag{4.10}$$

where $\mu(\overrightarrow{S})$ *is the collection of all measurable future events. Write the* i^{th} *causal state as* σ_i *and the set of all causal states as* $\boldsymbol{\mathcal{S}}$*; the corresponding random variable is denoted* $\mathcal{S}$*, and its realization* σ.

The cardinality and topology of $\boldsymbol{\mathcal{S}}$ are unspecified. $\boldsymbol{\mathcal{S}}$ can be finite, countably infinite, a continuum, a Cantor set, or something stranger still. Examples of these are given in Crutchfield (1994a) and Upper (1997); see especially the examples for hidden Markov models given there.

Alternately and equivalently, I could define an equivalence relation $\sim_\epsilon$ such that two histories are equivalent if and only if they have the same conditional distribution of futures, and then define causal states as the equivalence classes generated by $\sim_\epsilon$. (In fact, this was the original approach (Crutchfield and Young 1989).) Either way, the divisions of this partition of $\overleftarrow{\mathbf{S}}$ are made between regions that leave us in different conditions of ignorance about the future.

This last statement suggests another, still equivalent, description of ϵ:

$$\epsilon(\overleftarrow{s}) \quad = \quad \{\overleftarrow{s}' | \mathrm{P}(\overrightarrow{S}^L = \overrightarrow{s}^L | \overleftarrow{S} = \overleftarrow{s}) = \mathrm{P}(\overrightarrow{S}^L = \overrightarrow{s}^L | \overleftarrow{S} = \overleftarrow{s}') \ , \forall \overrightarrow{s}^L \in \overrightarrow{S}^L, \overleftarrow{s}' \in \overleftarrow{S} \ , L \in \mathbb{Z}^+\} \ . \tag{4.11}$$

Using this we can make the original definition, Eq. 4.10, more intuitive by picturing a sequence of partitions of the space $\overleftarrow{\mathbf{S}}$ of all histories in which each new partition, induced using futures of length $L+1$, is a refinement of the previous one induced using L. At the coarsest level, the first partition ($L = 1$) groups together those histories that have the same distribution for the very next observable. These classes are then subdivided using the distribution of the next two observables, then the next three, four, and so on. The limit of this sequence of partitions — the point at which every member of each class has the same distribution of futures, of whatever length, as every other member of that class — is the partition of $\overleftarrow{\mathbf{S}}$ induced by $\sim_\epsilon$.

Although they will not be of direct concern in the following, due to the time-asymptotic limits taken, there are transient causal states in addition to those (recurrent) causal states defined above in Eq. 4.10. Roughly speaking, the transient causal states describe how a lengthening sequence of observations allows us to identify the recurrent causal states with increasing precision. See Upper (1997) and Feldman and Crutchfield (1998a) for details on transient causal states.

Causal states are a particular kind of effective state, and they have all the properties common to effective states (Section 4.1.2). In particular, each causal state $\mathcal{S}_i$ has several structures attached:

1. The index i — the state's "name".

2. The set of histories that have brought the process to $\mathcal{S}_i$, $\{\overleftarrow{s} \in \mathcal{S}_i\}$.

3. A conditional distribution over futures, denoted $\mathrm{P}(\overrightarrow{S} | \mathcal{S}_i)$ and equal to $\mathrm{P}(\overrightarrow{S} | \overleftarrow{s})$, $\overleftarrow{s} \in \mathcal{S}_i$. Since I refer to this type of distribution frequently and since it is the "shape of the future", I'll call it the state's *morph*, following Crutchfield and Young (1989).

Ideally, each of these should be denoted by a different symbol, and there should be distinct functions linking each of these structures to their causal state. To keep the growth of notation under control, however, I'll be tactically vague about these distinctions. Readers may variously picture ϵ as mapping histories to (i) simple indices, (ii) subsets of histories, (iii) distributions over futures or (iv) ordered triples of indices, subsets, and morphs; or one may even leave ϵ uninterpreted, as preferred, without interfering with the development that follows.

4.2.1 Morphs

Each causal state has a unique morph, i.e., no two causal states have the same conditional distribution of futures. This follows directly from Definition 13, and it is not true of effective states in general. Another immediate consequence of that definition is that, for any measurable future event F,

$$\mathrm{P}(\overrightarrow{S} \in F | \mathcal{S} = \epsilon(\overleftarrow{s})) = \mathrm{P}(\overrightarrow{S} \in F | \overleftarrow{S} = \overleftarrow{s}). \tag{4.12}$$

(Again, this is not generally true of effective states.) This observation lets us prove a useful lemma about the conditional independence of the past $\overleftarrow{S}$ and the future $\overrightarrow{S}$.

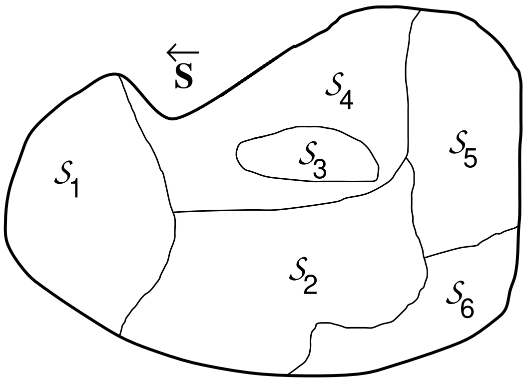

Figure 4.2: A schematic representation of the partitioning of the set $\overleftarrow{\mathbf{S}}$ of all histories into causal states $\mathcal{S}_i \in \boldsymbol{\mathcal{S}}$. Within each causal state all the individual histories $\overleftarrow{s}$ have the same morph — the same conditional distribution $\mathrm{P}(\overrightarrow{S}|\overleftarrow{s})$ for future observables.

Lemma 6 *The past and the future are independent, conditioning on the causal states:* $\overleftarrow{S} \perp\!\!\!\perp \overrightarrow{S} | \mathcal{S}$.

Proof. By Proposition 9 of Appendix B.3, $\overleftarrow{S}$ and $\overrightarrow{S}$ are independent given $\mathcal{S}$ if and only if, for any measurable set F of futures, $\mathrm{P}(\overrightarrow{S} \in F | \overleftarrow{S} = \overleftarrow{s}, \mathcal{S} = \sigma) = \mathrm{P}(\overrightarrow{S} \in F | \mathcal{S} = \sigma)$. Since $\mathcal{S} = \epsilon(\overleftarrow{S})$, it is automatically true (Eq. B.6) that $\mathrm{P}(\overrightarrow{S} \in F | \overleftarrow{S} = \overleftarrow{s}, \mathcal{S} = \epsilon(\overleftarrow{s})) = \mathrm{P}(\overrightarrow{S} \in F | \overleftarrow{S} = \overleftarrow{s})$. But then, $\mathrm{P}(\overrightarrow{S} \in F | \overleftarrow{S} = \overleftarrow{s}) = \mathrm{P}(\overrightarrow{S} \in F | \mathcal{S} = \epsilon(\overleftarrow{s}))$, so $\mathrm{P}(\overrightarrow{S} \in F | \overleftarrow{S} = \overleftarrow{s}, \mathcal{S} = \sigma) = \mathrm{P}(\overrightarrow{S} \in F | \mathcal{S} = \sigma)$. QED.

Lemma 7 (Strict Homogeneity of Causal States) *A process's causal states are the largest subsets of histories that are all strictly homogeneous with respect to futures of all lengths.*

The proof is identical to that for memoryless transducers (Lemma 3).

4.2.2 Causal State-to-State Transitions

The causal state at any given time and the next value of the observed process together determine a new causal state; this is proved shortly in Lemma 10. Thus, there is a natural relation of succession among the causal states; recall the discussion of causality in Section 2.3.4. Moreover, given the current causal state, all the possible next values of the observed sequence ($\overrightarrow{S}^1$) have well defined conditional probabilities. In fact, by construction the entire semi-infinite future ($\overrightarrow{S}$) does. Thus, there is a well defined probability $T_{ij}^{(s)}$ of the process generating the value $s \in \mathcal{A}$ and going to causal state $\mathcal{S}_j$, if it is in state $\mathcal{S}_i$.

Definition 14 (Causal Transitions) *The labeled transition probability* $T_{ij}^{(s)}$ *is the probability of making the transition from state* $\mathcal{S}_i$ *to state* $\mathcal{S}_j$ *while emitting the symbol* $s \in \mathcal{A}$*:*

$$T_{ij}^{(s)} \equiv \mathrm{P}(\mathcal{S}' = \mathcal{S}_j, \overrightarrow{S}^1 = s | \mathcal{S} = \mathcal{S}_i) , \quad (4.13)$$

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

Lemma 8 (Transition Probabilities) $T_{ij}^{(s)}$ *is given by*

$$T_{ij}^{(s)} = \mathrm{P}(\overleftarrow{S} s \in \mathcal{S}_j | \overleftarrow{S} \in \mathcal{S}_i) , \quad (4.14)$$

where $\overleftarrow{S}\, s$ is read as the semi-infinite sequence obtained by concatenating $s \in \mathcal{A}$ onto the end of $\overleftarrow{S}$.

Proof. It's enough to show that the events concerned are really the same. That is, I want to show that

$$\left\{\mathcal{S}' = \mathcal{S}_j, \overrightarrow{S}^1 = s, \mathcal{S} = \mathcal{S}_i\right\} \quad = \quad \left\{\overleftarrow{S}\, s \in \mathcal{S}_j,\ \overleftarrow{S} \in \mathcal{S}\right\} \ .$$

Now, that $\mathcal{S} = \mathcal{S}_i$ and $\overleftarrow{S} \in \mathcal{S}_i$ are the same event is clear by construction. So, too, for $\overleftarrow{S}' \in \mathcal{S}_j$ and $\mathcal{S}' = \mathcal{S}_j$. So I can certainly assert that

$$\left\{\mathcal{S}' = \mathcal{S}_j, \overrightarrow{S}^1 = s, \mathcal{S} = \mathcal{S}_i\right\} \quad = \quad \left\{\overleftarrow{S}' \in \mathcal{S}_j, \overrightarrow{S}^1 = s,\ \overleftarrow{S} \in \mathcal{S}_i\right\} \ .$$

The conjunction of the first and third events implies that, for all $\overleftarrow{s}$, if $\overleftarrow{S} = \overleftarrow{s}$, then $\overleftarrow{S}' = \overleftarrow{s}\, a$, for some symbol $a \in \mathcal{A}$. But the middle event ensures that $a = s$. Hence,

$$\left\{\mathcal{S}' = \mathcal{S}_j, \overrightarrow{S}^1 = s, \mathcal{S} = \mathcal{S}_i\right\} \quad = \quad \left\{\overleftarrow{S}\, s \in \mathcal{S}_j, \overrightarrow{S}^1 = s,\ \overleftarrow{S} \in \mathcal{S}_i\right\} \ .$$

But now the middle event is redundant and can be dropped. Thus,

$$\left\{\mathcal{S}' = \mathcal{S}_j, \overrightarrow{S}^1 = s, \mathcal{S} = \mathcal{S}_i\right\} \quad = \quad \left\{\overleftarrow{S}\, s \in \mathcal{S}_j,\ \overleftarrow{S} \in \mathcal{S}_i\right\} \ ,$$

as promised. Since the events have the same probability, when conditioned on $\mathcal{S}$, the events $\left\{\overleftarrow{S}\, s \in \mathcal{S}_j\right\}$ and $\left\{\mathcal{S}' = \mathcal{S}_j, \overrightarrow{S}^1 = s\right\}$ will yield the same conditional probability.[3] QED.

Notice that $T_{ij}^{(\emptyset)} = \delta_{ij}$; that is, the transition labeled by the null symbol $\emptyset$ is the identity.

4.2.3 ϵ-Machines

The combination of the function ϵ from histories to causal states with the labeled transition probabilities $T_{ij}^{(s)}$ is called the *ϵ-machine* of the process (Crutchfield 1994a; Crutchfield and Young 1989).

Definition 15 (An ϵ-Machine Defined) *The* ϵ-machine *of a process is the ordered pair $\{\epsilon, \mathbf{T}\}$, where ϵ is the causal state function and $\mathbf{T}$ is set of the transition matrices for the states defined by ϵ.*

Equivalently, you can denote an ϵ-machine by $\{\boldsymbol{\mathcal{S}}, \mathbf{T}\}$.

I promised that computational mechanics would be "algebraic" back in Section 2.3.5, so here is an explicit connection with semi-group theory, and can you get more algebraic?

Lemma 9 (ϵ-Machines Are Monoids) *The algebra generated by the ϵ-machine $\{\epsilon, \mathbf{T}\}$ is a monoid — a semi-group with an identity element.*

Proof. See Appendix B.2.

Remark. Due to this, ϵ-machines can be interpreted as capturing a process's *generalized symmetries.* Any subgroups of an ϵ-machine's semi-group are, in fact, symmetries in the usual sense.

Lemma 10 (ϵ-Machines Are Deterministic) *For each $\mathcal{S}_i \in \boldsymbol{\mathcal{S}}$ and each $s \in \mathcal{A}$, there is at most one $\mathcal{S}_j \in \boldsymbol{\mathcal{S}}$ such that, for every history $\overleftarrow{s} \in \mathcal{S}_i$, the history $\overleftarrow{s}s \in \mathcal{S}_j$. If such a $\mathcal{S}_j$ exists, then for all other $\mathcal{S}_k \in \boldsymbol{\mathcal{S}}$, $T_{ik}^{(s)} = 0$. If there is no such $\mathcal{S}_j$, then $T_{ik}^{(s)} = 0$ for all $\mathcal{S}_k \in \boldsymbol{\mathcal{S}}$ whatsoever. That is, the ϵ-machine is deterministic in the sense of Definition 59.*

[3]Technically, they will yield versions of the same conditional probability, i.e., they will agree with probability 1. See Appendix B.3.

Proof. The first part of the lemma asserts that for all $s \in \mathcal{A}$ and $\overleftarrow{s}, \overleftarrow{s}' \in \overleftarrow{\mathbf{S}}$, if $\epsilon(\overleftarrow{s}) = \epsilon(\overleftarrow{s}')$, then $\epsilon(\overleftarrow{s}s) = \epsilon(\overleftarrow{s}'s)$. ($\overleftarrow{s}s$ is just another history and belongs to one or another causal state.) I'll show that this follows directly from causal equivalence.

Consider any pair of histories $\overleftarrow{s}, \overleftarrow{s}'$ such that $\epsilon(\overleftarrow{s}) = \epsilon(\overleftarrow{s}')$, any single symbol s, and a (measurable) set F of future events. Let sF denote the set of futures obtained by prefixing the symbol s to each future in F. (sF is also measurable.) By causal equivalence, $\mathrm{P}(\overrightarrow{S} \in sF | \overleftarrow{S} = \overleftarrow{s}) = \mathrm{P}(\overrightarrow{S} \in sF | \overleftarrow{S} = \overleftarrow{s}')$. Now, $\overrightarrow{S} \in sF$ can be decomposed into the intersection of two events: $\overrightarrow{S}^1 = s$ and $\overrightarrow{S}_1 \in F$, where $\overrightarrow{S}_1$ is the random variable for the future sequence, ignoring the next symbol. Therefore, we begin with the following equalities.

$$
\begin{aligned}
\mathrm{P}(\overrightarrow{S} \in sF | \overleftarrow{S} = \overleftarrow{s}) &= \mathrm{P}(\overrightarrow{S} \in sF | \overleftarrow{S} = \overleftarrow{s}') \\
\mathrm{P}(\overrightarrow{S}^1 = s, \overrightarrow{S}_1 \in F | \overleftarrow{S} = \overleftarrow{s}) &= \mathrm{P}(\overrightarrow{S}^1 = s, \overrightarrow{S}_1 \in F | \overleftarrow{S} = \overleftarrow{s}')
\end{aligned}
$$

For any three random variables X, Y, Z, the conditional probability $\mathrm{P}(Z \in A, Y = y | X = x)$ can be factored as $\mathrm{P}(Z \in A | Y = y, X = x)\mathrm{P}(Y = y | X = x)$ (Eq. B.5) [4].

$$
\begin{aligned}
&\mathrm{P}(\overrightarrow{S}_1 \in F | \overrightarrow{S}^1 = s, \overleftarrow{S} = \overleftarrow{s})\mathrm{P}(\overrightarrow{S}^1 = s | \overleftarrow{S} = \overleftarrow{s}) \\
&= \mathrm{P}(\overrightarrow{S}_1 \in F | \overrightarrow{S}^1 = s, \overleftarrow{S} = \overleftarrow{s}')\mathrm{P}(\overrightarrow{S}^1 = s | \overleftarrow{S} = \overleftarrow{s}')
\end{aligned}
$$

From causal equivalence, the second factors on each side of the equation are equal, so divide through for them. (I address the case where $\mathrm{P}(\overrightarrow{S}^1 = s | \overleftarrow{S} = \overleftarrow{s}) = \mathrm{P}(\overrightarrow{S}^1 = s | \overleftarrow{S} = \overleftarrow{s}') = 0$ below.)

$$
\begin{aligned}
\mathrm{P}(\overrightarrow{S}_1 \in F | \overrightarrow{S}^1 = s, \overleftarrow{S} = \overleftarrow{s}) &= \mathrm{P}(\overrightarrow{S}_1 \in F | \overrightarrow{S}^1 = s, \overleftarrow{S} = \overleftarrow{s}') \\
\mathrm{P}(\overrightarrow{S} \in F | \overleftarrow{S} = \overleftarrow{s}s) &= \mathrm{P}(\overrightarrow{S} \in F | \overleftarrow{S} = \overleftarrow{s}'s)
\end{aligned}
$$

The last step is justified by (conditional) stationarity. Since the set F of future events is arbitrary, it follows that $\overleftarrow{s}s \sim_\epsilon \overleftarrow{s}'s$. Consequently, for each $\mathcal{S}_i$ and each s, there is at most one $\mathcal{S}_j$ such that $T_{ij}^{(s)} > 0$.

As remarked, causal equivalence tells us that $\mathrm{P}(\overrightarrow{S}^1 = s | \overleftarrow{S} = \overleftarrow{s}) = \mathrm{P}(\overrightarrow{S}^1 = s | \overleftarrow{S} = \overleftarrow{s}')$. But they could both be equal to zero, in which case we can't divide through for them. But then, again as promised, it follows that every entry in the transition matrix $T_{ij}^{(s)} = 0$, when $\mathcal{S}_i = \epsilon(\overleftarrow{s})$. Thus, the labeled transition probabilities have the promised form. QED.

Remark 1. This use of "determinism" is entirely standard in automata theory (see Appendix A.4), but obviously is slightly confusing. Many simple stochastic processes, such as Markov chains, are deterministic in this sense. Indeed, some computer scientists are so shameless as to say things like "stochastic deterministic finite automata". Sadly, nothing can be done about this. Whenever there is a possibility of confusion between determinism in the automata-theoretic sense, and determinism in the ordinary, physical sense, I'll call the latter "non-stochasticity" or "non-randomness".

Remark 2. Starting from a fixed state, a given symbol always leads to at most one single state. But there can be several transitions from one state to another, each labeled with a different symbol.

Remark 3. Clearly, if $T_{ij}^{(s)} > 0$, then $T_{ij}^{(s)} = \mathrm{P}(\overrightarrow{S}^1 = s | \mathcal{S} = \mathcal{S}_i)$. In automata theory the disallowed transitions ($T_{ij}^{(s)} = 0$) are sometimes explicitly represented and lead to a "reject" state indicating that the particular history does not occur.

Lemma 11 (ϵ-Machines Are Markovian) *Given the causal state at time $t-1$, the causal state at time t is independent of the causal state at earlier times.*

[4] This assumes the regularity of the conditional probabilities, which is valid for our discrete processes. Again, see Appendix B.3.

Proof. I'll start by showing that, writing $\mathcal{S}$, $\mathcal{S}'$, $\mathcal{S}''$ for the sequence of causal states at three successive times, $\mathcal{S}$ and $\mathcal{S}''$ are conditionally independent, given $\mathcal{S}'$.

Let M be a (measurable) set of causal states.

$$\mathrm{P}(\mathcal{S}'' \in \mathrm{M}|\mathcal{S}' = \sigma', \mathcal{S} = \sigma) \quad = \quad \mathrm{P}(\overrightarrow{S}^1 \in A|\mathcal{S}' = \sigma', \mathcal{S} = \sigma) \ ,$$

where $A \subseteq \mathcal{A}$ is the set of all symbols that lead from σ' to some $\sigma'' \in \mathrm{M}$. This is a well-defined and measurable set, in virtue of Lemma 10 immediately preceding, which also guarantees (see Remark 3 to the Lemma) the equality of conditional probabilities I used. Invoking Lemma 7, conditioning on $\mathcal{S}$ has no further effect once we've conditioned on $\mathcal{S}'$,

$$\begin{aligned}
\mathrm{P}(\overrightarrow{S}^1 \in A|\mathcal{S}' = \sigma', \mathcal{S} = \sigma) &= \mathrm{P}(\overrightarrow{S}^1 \in A|\mathcal{S}' = \sigma') \\
&= \mathrm{P}(\mathcal{S}'' \in \mathrm{M}|\mathcal{S}' = \sigma')
\end{aligned}$$

But (Proposition 9 and Eq. B.4) this is true if and only if conditional independence holds. Now the lemma follows by straightforward mathematical induction. QED.

Remark 1. This lemma strengthens the claim that the causal states are, in fact, the causally efficacious states: given knowledge of the present state, what has gone before makes no difference. (Again, recall the philosophical preliminaries of Section 2.3.4.)

Remark 2. This result indicates that the causal states, considered as a process, define a Markov process. Thus, causal states are a kind of generalization of hidden Markovian states. Of course, the class of ϵ-machines is substantially richer (Crutchfield 1994a; Upper 1997) than what's normally associated with Markov processes (Kemeny and Snell 1976; Kemeny, Snell and Knapp 1976) or even hidden Markov processes (Elliot, Aggoun and Moore 1995). In fact, we've just shown that *every* conditionally stationary discrete stochastic process has a Markovian representation!

Definition 16 (ϵ-Machine Reconstruction) ϵ-Machine reconstruction *is any procedure that given a process* $\mathrm{P}(\overleftrightarrow{S})$ *(respectively an approximation of* $\mathrm{P}(\overleftrightarrow{S})$*), produces the process's* ϵ*-machine* $\{\boldsymbol{\mathcal{S}}, \mathbf{T}\}$ *(respectively an approximation of* $\{\boldsymbol{\mathcal{S}}, \mathbf{T}\}$*).*

Given a mathematical description of a process, one can often calculate analytically its ϵ-machine. (For example, see the computational mechanics analysis of spin systems in Feldman and Crutchfield 1998a.) There is also a wide range of algorithms which reconstruct ϵ-machines from empirical estimates of $\mathrm{P}(\overleftrightarrow{S})$. I give such an algorithm in the next chapter.

4.3 Optimalities and Uniqueness, or, Why Causal States Are the Funk

I now show that: causal states are maximally accurate predictors of minimal statistical complexity; they are unique in sharing both properties; and their state-to-state transitions are minimally stochastic. In other words, they satisfy both of the constraints borrowed from Occam, and they are the only representations that do so. The overarching moral here is that causal states and ϵ-machines are *the* goals in any learning or modeling scheme. The argument is made by the time-honored means of proving optimality theorems.

All the theorems, and some of the lemmas, will be established by comparing causal states, generated by ϵ, with other rival sets of states, generated by other functions η. In short, none of the rival states — none of the other patterns — can out-perform the causal states.

It is convenient to recall some notation before plunging in. Let $\mathcal{S}$ be the random variable for the current causal state, $\overrightarrow{S}^1 \in \mathcal{A}$ the next "observable" we get from the original stochastic process, $\mathcal{S}'$ the next causal state, $\mathcal{R}$ the current state according to η, and $\mathcal{R}'$ the next η-state. σ will stand for a particular value (causal state) of $\mathcal{S}$ and ρ a particular value of $\mathcal{R}$. When I quantify over alternatives to the causal states, I'll quantify over $\boldsymbol{\mathcal{R}}$.

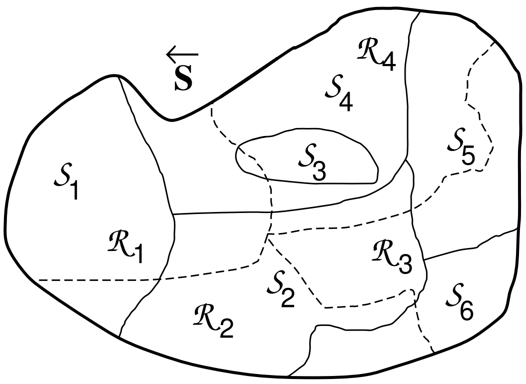

Figure 4.3: An alternative class $\boldsymbol{\mathcal{R}}$ of states (delineated by dashed lines) that partition $\overleftarrow{\mathbf{S}}$ overlaid on the causal states $\boldsymbol{\mathcal{S}}$ (outlined by solid lines). Here, for example, $\mathcal{S}_2$ contains parts of $\mathcal{R}_1$, $\mathcal{R}_2$, $\mathcal{R}_3$ and $\mathcal{R}_4$. The collection of all such alternative partitions form *Occam's pool*. Note again that the $\mathcal{R}_i$ need not be compact nor simply connected, as drawn.

Theorem 5 (Causal States Are Prescient) *(Crutchfield and Shalizi 1999)*
For all $\boldsymbol{\mathcal{R}}$ and all $L \in \mathbb{Z}^+$,

$$H[\overrightarrow{S}^L|\mathcal{R}] \geq H[\overrightarrow{S}^L|\mathcal{S}] \ . \tag{4.15}$$

Proof. We have already seen that $H[\overrightarrow{S}^L|\mathcal{R}] \geq H[\overrightarrow{S}^L|\overleftarrow{S}]$ (Lemma 5). But by construction (Definition 13),

$$\mathrm{P}(\overrightarrow{S}^L = \overrightarrow{s}^L | \overleftarrow{S} = \overleftarrow{s}) = \mathrm{P}(\overrightarrow{S}^L = \overrightarrow{s}^L | \mathcal{S} = \epsilon(\overleftarrow{s})) \ . \tag{4.16}$$

Since entropies depend only on the probability distribution, $H[\overrightarrow{S}^L|\mathcal{S}] = H[\overrightarrow{S}^L|\overleftarrow{S}]$ for every L. Thus, $H[\overrightarrow{S}^L|\mathcal{R}] \geq H[\overrightarrow{S}^L|\mathcal{S}]$, for all L. QED.

Remark. That is to say, causal states are as good at predicting the future — are as *prescient* — as complete histories. In this, they satisfy the first requirement borrowed from Occam. Since the causal states are well defined and since they can be systematically approximated, this shows that the upper bound on the strength of patterns (Definition 12 and Lemma 5, Remark) can be reached. Intuitively, the causal states achieve this because, unlike effective states in general, they do not throw away any information about the future which might be contained in $\overleftarrow{S}$. Even more colloquially, to paraphrase Bateson's (1979) definition of information, the causal states record every difference (about the past) that makes a difference (to the future). We can actually make this intuition quite precise, in an easy corollary to the theorem.

Corollary 2 (Causal States Are Sufficient Statistics) *The causal states $\boldsymbol{\mathcal{S}}$ of a process are sufficient statistics for predicting it.*

Proof. It follows from Theorem 5 and Eq. A.10 that, for all $L \in \mathbb{Z}^+$,

$$I[\overrightarrow{S}^L; \mathcal{S}] = I[\overrightarrow{S}^L; \overleftarrow{S}] \ , \tag{4.17}$$

where I was defined in Eq. A.10. Consequently, by Proposition 6, the causal states are sufficient statistics for futures of any length. QED.

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

Definition 17 (Prescient Rivals) Prescient rivals $\widehat{\boldsymbol{\mathcal{R}}}$ *are states that are as predictive as the causal states; viz., for all* $L \in \mathbb{Z}^+$,

$$H[\overrightarrow{S}^L | \widehat{\mathcal{R}}] = H[\overrightarrow{S}^L | \mathcal{S}] \ . \tag{4.18}$$

Remark. Prescient rivals are also sufficient statistics.

Theorem 6 (Sufficiency and Determinism Imply Prescience) *If $\mathcal{R}$ is a sufficient statistic for the next symbol, i.e., if* $\mathrm{P}(\overrightarrow{S}^1 = a | \mathcal{R} = \eta(\overleftarrow{s})) = \mathrm{P}(\overrightarrow{S}^1 = a | \mathcal{S} = \epsilon(\overleftarrow{s}))$ *for all* $a \in \mathcal{A}$, *and if* $\boldsymbol{\mathcal{R}}$ *is deterministic (in the sense of Definition 59), then* $\boldsymbol{\mathcal{R}}$ *is prescient. That is, deterministic states which get the distribution of the next symbol right are prescient.*

Proof: It will be enough to show that, for any L, $\mathrm{P}(\overrightarrow{S}^L | \mathcal{R}) = \mathrm{P}(\overrightarrow{S}^L | \mathcal{S})$, since then the equality of conditional entropies is obvious. I do this by induction; suppose that the equality of conditional probabilities holds for all lengths of futures up to some L, and consider futures of length $L + 1$.

$$\begin{aligned} \mathrm{P}(\overrightarrow{S}^{L+1} &= s^L a | \mathcal{R} = \eta(\overleftarrow{s})) = \\ &= \mathrm{P}(\overrightarrow{S}_{L+1} = a | \mathcal{R} = \eta(\overleftarrow{s}), \overrightarrow{S}^L = s^L) \mathrm{P}(\overrightarrow{S}^L = s^L | \mathcal{R} = \eta(\overleftarrow{s})) \\ &= \mathrm{P}(\overrightarrow{S}_{L+1} = a | \mathcal{R} = \eta(\overleftarrow{s}), \overrightarrow{S}^L = s^L) \mathrm{P}(\overrightarrow{S}^L = s^L | \mathcal{S} = \epsilon(\overleftarrow{s})) \end{aligned} \tag{4.19}$$

where the second line uses the inductive hypothesis. Since we assume the $\mathcal{R}$ states are deterministic, the combination of the current effective state ($\eta(\overleftarrow{s})$) and the next L symbols (s^L) fixes a unique future effective state, namely $\eta(\overleftarrow{s} s^L)$. Thus, by Proposition 8, Appendix B.3, we see that $\mathrm{P}(\overrightarrow{S}_{L+1} = a | \mathcal{R} = \eta(\overleftarrow{s}), \overrightarrow{S}^L = s^L) = \mathrm{P}(\overrightarrow{S}^1 = a | \mathcal{R} = \eta(\overleftarrow{s} s^L))$. Substituting back in,

$$\begin{aligned} \mathrm{P}(\overrightarrow{S}^{L+1} = s^L a | \mathcal{R} = \eta(\overleftarrow{s})) &= \mathrm{P}(\overrightarrow{S}^1 = a | \mathcal{R} = \eta(\overleftarrow{s} s^L)) \mathrm{P}(\overrightarrow{S}^L = s^L | \mathcal{S} = \epsilon(\overleftarrow{s}) & (4.20) \\ &= \mathrm{P}(\overrightarrow{S}^1 = a | \mathcal{S} = \epsilon(\overleftarrow{s} s^L)) \mathrm{P}(\overrightarrow{S}^L = s^L | \mathcal{S} = \epsilon(\overleftarrow{s}) & (4.21) \\ &= \mathrm{P}(\overrightarrow{S}^{L+1} = s^L a | \mathcal{S} = \epsilon(\overleftarrow{s})) \ , & (4.22) \end{aligned}$$

so the induction is established. Since (by hypothesis) it holds for $L = 1$, it holds for all positive L. QED.

Remark. The causal states satisfy the hypotheses of this proposition. Since, as we shall see (Theorem 7), the causal states are the minimal prescient states, they are also the minimal deterministic states which get the distribution of the next symbol right. This is handy when doing ϵ-machine reconstruction (Chapter 5).

Lemma 12 (Refinement Lemma) *For all prescient rivals* $\widehat{\boldsymbol{\mathcal{R}}}$ *and for each* $\widehat{\rho} \in \widehat{\boldsymbol{\mathcal{R}}}$, *there is a* $\sigma \in \boldsymbol{\mathcal{S}}$ *and a measure-*0 *subset* $\widehat{\rho}_0 \subset \widehat{\rho}$, *possibly empty, such that* $\widehat{\rho} \setminus \widehat{\rho}_0 \subseteq \sigma$, *where* $\setminus$ *is set subtraction.*

The proof is identical to that for the memoryless case (Lemma 4). An alternative, more algebraic, proof appears in Appendix B.4. The Lemma is illustrated by the contrast between Figures 4.4 and 4.3.

Theorem 7 (Causal States Are Minimal) *(Crutchfield and Shalizi 1999) For all prescient rivals* $\widehat{\boldsymbol{\mathcal{R}}}$,

$$C_\mu(\widehat{\boldsymbol{\mathcal{R}}}) \geq C_\mu(\boldsymbol{\mathcal{S}}) \ . \tag{4.23}$$

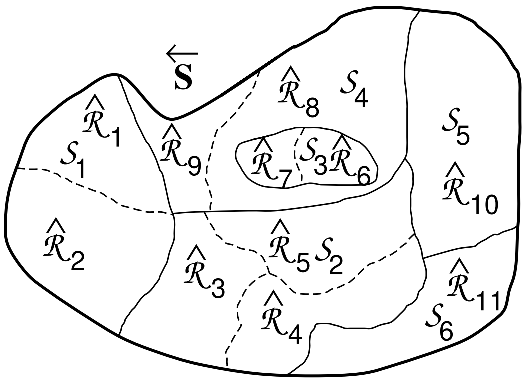

Figure 4.4: A prescient rival partition $\widehat{\boldsymbol{\mathcal{R}}}$ must be a refinement of the causal-state partition *almost everywhere*. That is, almost all of each $\widehat{\mathcal{R}}_i$ must contained within some $\mathcal{S}_j$; the exceptions, if any, are a set of histories of measure 0. Here for instance $\mathcal{S}_2$ contains the positive-measure parts of $\widehat{\mathcal{R}}_3$, $\widehat{\mathcal{R}}_4$, and $\widehat{\mathcal{R}}_5$. One of these rival states, say $\widehat{\mathcal{R}}_3$, could have member-histories in any or all of the other causal states, provided the total measure of such exceptional histories is zero. Cf. Figure 4.3.

The proof is identical to that in the memoryless case (Theorem 2).

Remark 1. No rival pattern, which is as good at predicting the observations as the causal states, is any simpler than the causal states. (This is the theorem of Crutchfield and Young (1989).) Occam therefore tells us that there is no reason not to use the causal states. The next theorem shows that causal states are uniquely optimal and so that Occam's Razor all but forces us to use them.

Remark 2. Here it becomes important that we are trying to predict the whole of $\overrightarrow{S}$ and not just some piece, $\overrightarrow{S}^L$. Suppose two histories $\overleftarrow{s}$ and $\overleftarrow{s}'$ have the same conditional distribution for futures of lengths up to L, but differing ones after that. They would then belong to different causal states. An η-state that merged those two causal states, however, would have just as much ability to predict $\overrightarrow{S}^L$ as the causal states. More, these $\mathcal{R}$-states would be simpler, in the sense that the uncertainty in the current state would be lower. Causal states are optimal, but for the hardest job — that of predicting futures of all lengths.

Corollary 3 (Causal States Are Minimal Sufficient Statistics) *The causal states are minimal sufficient statistics for predicting futures of all lengths.*

The proof is identical to that for the memoryless case (Corollary 1).

I can now, as promised, define the *statistical complexity of a process* (Crutchfield 1994a; Crutchfield and Young 1989).

Definition 18 (Statistical Complexity of a Process) *The statistical complexity "$C_\mu(\mathcal{O})$" of a process $\mathcal{O}$ is that of its causal states: $C_\mu(\mathcal{O}) \equiv C_\mu(\boldsymbol{\mathcal{S}})$.*

Due to the minimality of causal states, the statistical complexity measures the average amount of historical memory stored in the process. Since we can trivially elaborate internal states, while still generating the same observed process — arbitrarily complex sets of states can be prescient. If we didn't have the minimality theorem, we couldn't talk about the complexity of the process, just that of various predictors of it (Crutchfield 1992).

Theorem 8 (Causal States Are Unique) *For all prescient rivals $\widehat{\boldsymbol{\mathcal{R}}}$, if $C_\mu(\widehat{\boldsymbol{\mathcal{R}}}) = C_\mu(\boldsymbol{\mathcal{S}})$, then there exists an invertible function between $\widehat{\boldsymbol{\mathcal{R}}}$ and $\boldsymbol{\mathcal{S}}$ that almost always preserves equivalence of state: $\widehat{\boldsymbol{\mathcal{R}}}$ and η are the same as $\boldsymbol{\mathcal{S}}$ and ϵ, respectively, except on a set of histories of measure* 0.

The proof is the same as for the memoryless case (Theorem 3); the same remarks apply.

Theorem 9 (ϵ-Machines Are Minimally Stochastic) *(Crutchfield and Shalizi 1999) For all prescient rivals* $\widehat{\boldsymbol{\mathcal{R}}}$*,*

$$H[\widehat{\mathcal{R}}'|\widehat{\mathcal{R}}] \geq H[\mathcal{S}'|\mathcal{S}] \ , \tag{4.24}$$

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

Proof. From Lemma 10, $\mathcal{S}'$ is fixed by $\mathcal{S}$ and $\overrightarrow{S}^1$ together, thus $H[\mathcal{S}'|\mathcal{S}, \overrightarrow{S}^1] = 0$ by Eq. A.23. Therefore, from the chain rule for entropies Eq. A.17,

$$H[\overrightarrow{S}^1|\mathcal{S}] \quad = \quad H[\mathcal{S}', \overrightarrow{S}^1|\mathcal{S}] \ . \tag{4.25}$$

There's no result like the Determinism Lemma 10 for the rival states $\widehat{\mathcal{R}}$, but entropies are always non-negative: $H[\widehat{\mathcal{R}}'|\widehat{\mathcal{R}}, \overrightarrow{S}^1] \geq 0$. Since for all L, $H[\overrightarrow{S}^L|\widehat{\mathcal{R}}] = H[\overrightarrow{S}^L|\mathcal{S}]$ by the definition (Definition 17) of prescient rivals, $H[\overrightarrow{S}^1|\widehat{\mathcal{R}}] = H[\overrightarrow{S}^1|\mathcal{S}]$. Now apply the chain rule again,

$$\begin{aligned}
H[\widehat{\mathcal{R}}', \overrightarrow{S}^1|\widehat{\mathcal{R}}] &= H[\overrightarrow{S}^1|\widehat{\mathcal{R}}] + H[\widehat{\mathcal{R}}'|\overrightarrow{S}^1, \widehat{\mathcal{R}}] && (4.26)\\
&\geq H[\overrightarrow{S}^1|\widehat{\mathcal{R}}] && (4.27)\\
&= H[\overrightarrow{S}^1|\mathcal{S}] && (4.28)\\
&= H[\mathcal{S}', \overrightarrow{S}^1|\mathcal{S}] && (4.29)\\
&= H[\mathcal{S}'|\mathcal{S}] + H[\overrightarrow{S}^1|\mathcal{S}', \mathcal{S}] \ . && (4.30)
\end{aligned}$$

To go from Eq. 4.28 to Eq. 4.29 use Eq. 4.25, and in the last step use the chain rule once more.

Using the chain rule one last time, with feeling, we have

$$H[\widehat{\mathcal{R}}', \overrightarrow{S}^1|\widehat{\mathcal{R}}] = H[\widehat{\mathcal{R}}'|\widehat{\mathcal{R}}] + H[\overrightarrow{S}^1|\widehat{\mathcal{R}}', \widehat{\mathcal{R}}] \ . \tag{4.31}$$

Putting these expansions, Eqs. 4.30 and 4.31, together we get

$$\begin{aligned}
H[\widehat{\mathcal{R}}'|\widehat{\mathcal{R}}] + H[\overrightarrow{S}^1|\widehat{\mathcal{R}}', \widehat{\mathcal{R}}] &\geq H[\mathcal{S}'|\mathcal{S}] + H[\overrightarrow{S}^1|\mathcal{S}', \mathcal{S}] && (4.32)\\
H[\widehat{\mathcal{R}}'|\widehat{\mathcal{R}}] - H[\mathcal{S}'|\mathcal{S}] &\geq H[\overrightarrow{S}^1|\mathcal{S}', \mathcal{S}] - H[\overrightarrow{S}^1|\widehat{\mathcal{R}}', \widehat{\mathcal{R}}] \ .
\end{aligned}$$

From Lemma 12, we know that $\mathcal{S} = g(\widehat{\mathcal{R}})$, so there is another function g' from ordered pairs of η-states to ordered pairs of causal states: $(\mathcal{S}', \mathcal{S}) = g'(\widehat{\mathcal{R}}', \widehat{\mathcal{R}})$. Therefore, Eq. A.25 implies

$$H[\overrightarrow{S}^1|\mathcal{S}', \mathcal{S}] \geq H[\overrightarrow{S}^1|\widehat{\mathcal{R}}', \widehat{\mathcal{R}}] \ . \tag{4.33}$$

And so, we have that

$$\begin{aligned}
H[\overrightarrow{S}^1|\mathcal{S}', \mathcal{S}] - H[\overrightarrow{S}^1|\widehat{\mathcal{R}}', \widehat{\mathcal{R}}] &\geq 0\\
H[\widehat{\mathcal{R}}'|\widehat{\mathcal{R}}] - H[\mathcal{S}'|\mathcal{S}] &\geq 0\\
H[\widehat{\mathcal{R}}'|\widehat{\mathcal{R}}] &\geq H[\mathcal{S}'|\mathcal{S}] \ . && (4.34)
\end{aligned}$$

QED.

Remark. What this theorem says is that there is no more uncertainty in transitions between causal states, than there is in the transitions between any other kind of prescient effective states. In other words, the causal states approach as closely to perfect determinism — in the usual physical, non-computation-theoretic sense — as any rival that is as good at predicting the future.

4.4 Bounds

In this section I develop bounds between measures of structural complexity and entropy derived from ϵ-machines and those from ergodic and information theories, which are perhaps more familiar.

Definition 19 (Excess Entropy) *The excess entropy* $\mathbf{E}$ *of a process is the mutual information between its semi-infinite past and its semi-infinite future:*

$$\mathbf{E} \equiv I[\overrightarrow{S}; \overleftarrow{S}] \ . \tag{4.35}$$

The excess entropy is a frequently-used measure of the complexity of stochastic processes and appears under a variety of names; e.g., "predictive information", "stored information", "effective measure complexity", and so on (Crutchfield and Packard 1983; Shaw 1984; Grassberger 1986; Lindgren and Nordahl 1988; Li 1991; Arnold 1996; Bialek and Tishby 1999). $\mathbf{E}$ measures the amount of *apparent* information stored in the observed behavior about the past. But $\mathbf{E}$ is not, in general, the amount of memory that the process stores *internally* about its past; that's C_μ.

Theorem 10 (The Bounds of Excess) *The statistical complexity* C_μ *bounds the excess entropy* $\mathbf{E}$*:*

$$\mathbf{E} \leq C_\mu \ , \tag{4.36}$$

with equality if and only if $H[\mathcal{S}| \overrightarrow{S}] = 0$*.*

Proof. $\mathbf{E} = I[\overrightarrow{S}; \overleftarrow{S}] = H[\overrightarrow{S}] - H[\overrightarrow{S} \mid \overleftarrow{S}]$ and, by the construction of causal states, $H[\overrightarrow{S} \mid \overleftarrow{S}] = H[\overrightarrow{S} \mid \mathcal{S}]$, so

$$\mathbf{E} = H[\overrightarrow{S}] - H[\overrightarrow{S} \mid \mathcal{S}] = I[\overrightarrow{S}; \mathcal{S}] \ . \tag{4.37}$$

Thus, since the mutual information between two variables is never larger than the self-information of either one of them (Eq. A.20), $\mathbf{E} \leq H[\mathcal{S}] = C_\mu$, with equality if and only if $H[\mathcal{S}| \overrightarrow{S}] = 0$. QED.

Remark 1. Note that I invoked $H[\overrightarrow{S}]$, not $H[\overrightarrow{S}^L]$, but only while subtracting off quantities like $H[\overrightarrow{S} \mid \overleftarrow{S}]$. We needn't worry, therefore, about the existence of a finite $L \to \infty$ limit for $H[\overrightarrow{S}^L]$, just that of a finite $L \to \infty$ limit for $I[\overrightarrow{S}^L; \overleftarrow{S}]$ and $I[\overrightarrow{S}^L; \mathcal{S}]$. There are many elementary cases (e.g., the fair coin process) where the latter limits exist, while the former do not. (See Gray (1990) for details on how to construct such a mutual information with full rigor.)

Remark 2. At first glance, it is tempting to see $\mathbf{E}$ as the amount of information stored in a process. As Theorem 10 shows, this temptation should be resisted. $\mathbf{E}$ is only a lower bound on the true amount of information the process stores about its history, namely C_μ. You can, however, say that $\mathbf{E}$ measures the *apparent* information in the process, since it is defined directly in terms of observed sequences and not in terms of hidden, intrinsic states, as C_μ is.

Remark 3. Perhaps another way to describe what $\mathbf{E}$ measures is to note that, by its implicit assumption of block-Markovian structure, it takes sequence-blocks as states. But even for the class of block-Markovian sources, for which such an assumption is appropriate, excess entropy and statistical complexity measure different kinds of information storage. Feldman and Crutchfield (1998a) and Crutchfield and Feldman (1997) showed that in the case of one-dimensional range-R spin systems, or any other block-Markovian source where block configurations are isomorphic to causal states,

$$C_\mu = \mathbf{E} + R h_\mu \ , \tag{4.38}$$

for finite R. Only for zero-entropy-rate block-Markovian sources will the excess entropy, a quantity estimated directly from sequence blocks, equal the statistical complexity, the amount of memory stored in the process. Examples of such sources include periodic processes, for which $C_\mu = \mathbf{E} = \log_2 p$, where p is the period.

Corollary 4 *For all prescient rivals* $\widehat{\boldsymbol{\mathcal{R}}}$*,*

$$\mathbf{E} \leq H[\widehat{\mathcal{R}}] \ . \tag{4.39}$$

Proof. This follows directly from Theorem 7, since $H[\widehat{\mathcal{R}}] \geq C_\mu$. QED.

Lemma 13 (Conditioning Does Not Affect Entropy Rate) *For all prescient rivals* $\widehat{\boldsymbol{\mathcal{R}}}$*,*

$$h[\overrightarrow{S}] = h[\overrightarrow{S}|\widehat{\mathcal{R}}] \ , \tag{4.40}$$

where the entropy rate $h[\overrightarrow{S}]$ *and the conditional entropy rate* $h[\overrightarrow{S}|\widehat{\mathcal{R}}]$ *were defined in Eq. 4.4 and Eq. 4.5, respectively.*

Proof. From Theorem 10 and its Corollary 4,

$$\lim_{L\to\infty} \left(H[\overrightarrow{S}^L] - H[\overrightarrow{S}^L|\widehat{\mathcal{R}}] \right) \leq \lim_{L\to\infty} H[\widehat{\mathcal{R}}] \ , \tag{4.41}$$

or,

$$\lim_{L\to\infty} \frac{H[\overrightarrow{S}^L] - H[\overrightarrow{S}^L|\widehat{\mathcal{R}}]}{L} \leq \lim_{L\to\infty} \frac{H[\widehat{\mathcal{R}}]}{L} \ . \tag{4.42}$$

Since, by Eq. A.15, $H[\overrightarrow{S}^L] - H[\overrightarrow{S}^L|\widehat{\mathcal{R}}] \geq 0$,

$$h[\overrightarrow{S}] - h[\overrightarrow{S}|\widehat{\mathcal{R}}] = 0 \ . \tag{4.43}$$

QED.

Remark. Forcing the process into a certain state $\widehat{\mathcal{R}} = \widehat{\rho}$ is akin to applying a controller, once. But in the infinite-entropy case, $H[\overrightarrow{S}^L] \to_{L\to\infty} \infty$, which is the general one, the future could contain (or consist of) an infinite sequence of disturbances. In the face of this "grand disturbance", the effects of the finite control are simply washed out.

Another way of viewing this is to reflect on the fact that $h[\overrightarrow{S}]$ accounts for the effects of all the dependencies between all the parts of the entire semi-infinite future. This, owing to the time-translation invariance of (conditional) stationarity, is equivalent to taking account of all the dependencies in the entire process, including those between past and future. But these are what is captured by $h[\overrightarrow{S}|\widehat{\mathcal{R}}]$. It is not that conditioning on $\mathcal{R}$ fails to reduce our uncertainty about the future; it does so, for all finite times, and conditioning on $\mathcal{S}$ achieves the maximum possible reduction in uncertainty. Rather, the lemma asserts that such conditioning cannot affect the asymptotic rate at which such uncertainty grows with time.

Theorem 11 (Control Theorem) *Given a class* $\widehat{\boldsymbol{\mathcal{R}}}$ *of prescient rivals,*

$$H[S] - h[\overrightarrow{S}|\widehat{\mathcal{R}}] \leq C_\mu \ , \tag{4.44}$$

where $H[S]$ *is the entropy of a single symbol from the observable stochastic process.*

Proof. As is well known (Cover and Thomas 1991, Theorem 4.2.1, p. 64), for any stationary stochastic process,

$$\lim_{L\to\infty} \frac{H[\overrightarrow{S}^L]}{L} = \lim_{L\to\infty} H[S_L|\overrightarrow{S}^{L-1}] \ . \tag{4.45}$$

Moreover, the limits always exist. Up to this point, I defined $h[\overrightarrow{S}]$ in the manner of the left-hand side; recall Eq. 4.4. It's now more convenient to use the right-hand side.

From the definition of conditional entropy,

$$\begin{aligned} H[\overleftarrow{S}^{L}] &= H[\overleftarrow{S}^{1}|\overleftarrow{S}^{L-1}] + H[\overleftarrow{S}^{L-1}] \\ &= H[\overleftarrow{S}^{L-1}|\overleftarrow{S}^{1}] + H[\overleftarrow{S}^{1}] \ . \end{aligned} \tag{4.46}$$

So we can express the entropy of the last observable the process generated before the present as

$$H[\overleftarrow{S}^{1}] = H[\overleftarrow{S}^{L}] - H[\overleftarrow{S}^{L-1}|\overleftarrow{S}^{1}] \tag{4.47}$$

$$= H[\overleftarrow{S}^{1}|\overleftarrow{S}^{L-1}] + H[\overleftarrow{S}^{L-1}] - H[\overleftarrow{S}^{L-1}|\overleftarrow{S}^{1}] \tag{4.48}$$

$$= H[\overleftarrow{S}^{1}|\overleftarrow{S}^{L-1}] + I[\overleftarrow{S}^{L-1};\overleftarrow{S}^{1}] \ . \tag{4.49}$$

To go from Eq. 4.47 to Eq. 4.48, substitute the first RHS of Eq. 4.46 for $H[\overleftarrow{S}^{L}]$.

Taking the $L \rightarrow \infty$ limit has no effect on the LHS,

$$H[\overleftarrow{S}^{1}] = \lim_{L\rightarrow\infty} \left(H[\overleftarrow{S}^{1}|\overleftarrow{S}^{L-1}] + I[\overleftarrow{S}^{L-1};\overleftarrow{S}^{1}] \right) \ . \tag{4.50}$$

Since the process is stationary, we can move the first term in the limit forward to $H[S_L|\overrightarrow{S}^{L-1}]$. This limit is $h[\overrightarrow{S}]$, by Eq. 4.45. Furthermore, because of stationarity, $H[\overleftarrow{S}^{1}] = H[\overrightarrow{S}^{1}] = H[S]$. Shifting the entropy rate $h[\overrightarrow{S}]$ to the LHS of Eq. 4.50 and appealing to time-translation once again,

$$H[S] - h[\overrightarrow{S}] = \lim_{L\rightarrow\infty} I[\overleftarrow{S}^{L-1};\overleftarrow{S}^{1}] \tag{4.51}$$

$$= I[\overleftarrow{S};\overrightarrow{S}^{1}] \tag{4.52}$$

$$= H[\overrightarrow{S}^{1}] - H[\overrightarrow{S}^{1}|\overleftarrow{S}] \tag{4.53}$$

$$= H[\overrightarrow{S}^{1}] - H[\overrightarrow{S}^{1}|\mathcal{S}] \tag{4.54}$$

$$= I[\overrightarrow{S}^{1};\mathcal{S}] \tag{4.55}$$

$$\leq H[\mathcal{S}] = C_\mu \ , \tag{4.56}$$

where the last inequality comes from Eq. A.20. QED.

Remark 1. Thinking of the controlling variable as the causal state, this is a limitation on the controller's ability to reduce the entropy *rate*.

Remark 2. This is the only result so far where the difference between the finite-L and the infinite-L cases is important. For the analogous result in the finite case, see Appendix B.5, Theorem 25.

Remark 3. By applying Theorem 7 and Lemma 13, we could go from the theorem as it stands to $H[S] - h[\overrightarrow{S}|\widehat{\mathcal{R}}] \leq H[\widehat{\mathcal{R}}]$. This has a pleasing appearance of symmetry to it, but is actually a weaker limit on the strength of the pattern or, equivalently, on the amount of control that fixing the causal state (or one of its rivals) can exert.

4.5 The Physical Meaning of Causal States

All this has been very abstract, and not particularly "physical." This is the price for a general method, one which is not tied to particular assumptions about the physical character of the processes to which it can be applied. That said, nothing prevents us from applying the formalism to the kind of things we came to know and love while reading Landau and Lifshitz (1980). In particular, the computational mechanics of time series can be applied to the time evolution of ordinary statistical-mechanical systems, and the result helps clarify the meaning of causal states — and the meaning of macrostates in statistical mechanics.[5]

Consider a collection of physical particles, obeying the usual laws of classical mechanics with some Hamiltonian or other, and described by an ensemble distribution in the microscopic phase space Γ. The ensemble is not necessarily any of the usual equilibrium ensembles, and we don't suppose that the system is anywhere near equilibrium or a steady state. Now think of your favorite macroscopic variable S. The value of S will be a function of where the system happens to be in Γ when you make your measurement, i.e., $S = s(x)$, $x \in \Gamma$. The macrovariable S induces a partition on the phase space Γ; call the partition A. Conversely, a (measurable) partition of Γ corresponds to some macroscopic variable. If you measure several macrovariables $S, R \ldots$ simultaneously (which is always possible, classically), the induced partition of Γ is simply the product of the partitions of the individual variables, $A \times B \times \ldots$. We may regard this joint variable as simply yet another macroscopic variable, which could be measured directly with the appropriate instrument. So, without loss of generality, let's just think about a single macrovariable. With only minor loss of generality, moreover, let's assume that it's discrete, and measured at discrete times.[6] Restricting ourselves to discrete times allows us to write the time-evolution of the phase space in the form of a discrete map, $\mathrm{T} : \Gamma \mapsto \Gamma$.

Histories of measurements of this macrovariable induce yet another partition of Γ, in the following manner. Each observation value s corresponds to a set A_s of points in phase space. The sequence of measurements ss' thus corresponds to the set $A_{ss'} \equiv \mathrm{T}A_s \cap A_{s'}$, where T is the phase-space mapping operator. Since the sets A_s form a partition, it's easy to see that the sets $A_{ss'}$ form a refinement of that partition. An exercise in mathematical induction extends this to any sequence of measurements of countable length. The partition induced by histories of length $L+1$ is always a refinement of histories of length L. So far this is an entirely standard construction of symbolic dynamics for a statistical-mechanical system, as found in, e.g., Dorfman (1998). Normally, to get useful results from such a construction, the initial partition must be a Markov or generating partition, or otherwise pretty special. Here we have just started with whatever observable partition we liked.

Now comes the trick. By making the time evolution of the statistical mechanical system look like an ordinary discrete stochastic process, we have brought it within the range of application of the theory developed in this chapter. We can construct causal states for it, and those states have three key properties: they are optimal predictors of the original sequence of measurements; they are Markovian; and they are the minimal set of states of which both those things are true. But $\boldsymbol{\mathcal{S}}$ is a partition of $\overleftarrow{\mathbf{S}}$, which in turn is a partition of Γ. Therefore $\boldsymbol{\mathcal{S}}$ induces a partition on Γ (which is coarser, generally considerably coarser, than that induced by $\overleftarrow{\mathbf{S}}$). The causal state, therefore, corresponds to a measurable macroscopic variable, call it C, which is the coarsest one that can both predict the macrovariable(s) with which we started, and whose own dynamics are Markovian. But these are the properties of a "good" set of macroscopic variables, of ones which define a useful macrostate: they are dynamically autonomous (Markovian), the present value of them predicts future behavior optimally, and nothing simpler does the job.[7] Thermodynamic macrostates, then, are causal states, and conversely causal states are a kind of generalized macrostate, with the value of the causal state acting as a generalized order parameter.

Put slightly differently, what we have done is construct a partition of the phase space Γ which is Marko-

[5]This section derives from Shalizi and Moore (2001). That in turn is based on earlier work connecting statistical and computational mechanics (Crutchfield 1992; Crutchfield 1994a; Crutchfield and Feldman 1997; Feldman and Crutchfield 1998a; Feldman 1998; Crutchfield and Shalizi 1999). Cf. Lloyd and Pagels (1988).

[6]The limited accuracy and precision of all instruments arguably imposes something like discretization on all our measurements anyway, but that's a bit of a tricky point, which I'd like to evade.

[7]An apparent exception is found in systems, like glasses (Zallen 1983) and spin glasses (Fischer and Hertz 1988), where there are memory effects over very long time scales. These are due, however, to the very large number of metastable states in these systems, transitions between which are slow. The memory effects can be eliminated by introducing the occupations of these metastable states as order parameters — by adding a macroscopic number of degrees of freedom, as Fischer and Hertz put it. For more on this point, see Shalizi and Moore (2001).

vian, starting from an arbitrary observational partition. Each causal state thus corresponds not only to a history of observations, but also to a region in phase space. (Perry and Binder (1999) have mapped these regions, albeit for an unusually simple phase space.) Even better, since the causal states form a Markov chain, the distribution of sequences of causal states is a Gibbs distribution.[8] Yet we haven't had to assume that our system is in equilibrium, or in a steady state, or has any particular kind of ensemble (such as a maximum entropy ensemble). This is, perhaps, part of the justification for why the assumption of Gibbs distributions is often fruitful in non-equilibrium statistical mechanics.[9]

Of course, this argument is very, very far from a complete story for macrostates and macrovariables. It says nothing, for instance, about why *extensive* quantities are good macrovariables. Nor does it say anything about why macrovariables are, so to speak, recyclable, why pressure (say) is a good macrovariable for many systems with little in common microscopically. The explanation of such regularities presumably is to be found, not in the very general statistical properties captured by computational mechanics, but in the more detailed dynamical properties studied by ergodic theory (Ruelle 1989; Dorfman 1998; Gaspard 1998; Ruelle 1999), and to some extent in the theory of large deviations (Ellis 1985; Ellis 1999).

[8]The proof that Markovianity implies a Gibbs measure over sequences, and vice versa, while fairly straightforward, is outside the scope of this book. See Guttorp (1995) for an elementary proof.

[9]Thanks to Erik van Nimwegen for this observation.

Chapter 5

A Machine Reconstruction Algorithm

> Those who are good at archery learnt from the bow and not from Yi the Archer. Those who know how to manage boats learnt from the boats and not from Wo. Those who can think learnt from themselves, and not from the Sages.
> —Anonymous (T'ang Dynasty).

A natural and appropriate reaction to the theory developed in Chapters 3 and 4 is that it may be all well and good as a pure mathematical construction, but that it will only matter if it can be implemented, if it can be put into practice. This is as it should be. Consider the difference in fate between two similar ideas proposed at roughly the same time, namely attractor reconstruction, a.k.a. "geometry from a time series" (Packard, Crutchfield, Farmer and Shaw 1980), and the Turing-machine test for the presence of deterministic structure in a time series (Takens 1983). The former has become a fundamental tool of nonlinear dynamics, not just because it is mathematically important, but because it can be reduced to practice. The latter is almost completely ignored, because it is simply impossible to implement. Implementation separates Neat Ideas from Real Tools.

This has been recognized since the first days of computational mechanics, when an algorithm was developed for ϵ-machine reconstruction (Crutchfield and Young 1989; Crutchfield and Young 1990), which merged distinct histories together into states when their morphs seemed "close". (I will briefly describe this algorithm, and related but distinct approaches, in Section 5.1.) This has since become the standard one, to the point where some conflate it with computational mechanics as such. People have used the algorithm on discrete maps (Crutchfield and Young 1990), on sequences from cellular automata (Hanson 1993) and on one-dimensional spin systems (Feldman and Crutchfield 1998a; Feldman 1998). It has even been applied to experimental data, from the dripping faucet system (Gonçalves, Pinto, Sartorelli and de Oliveira 1998), from stochastic resonance experiments (Witt, Neiman and Kurths 1997), and from turbulent geophysical fluid flows (Palmer, Fairall and Brewer 2000; Nicholas Watkins, personal communication, 2000).

While the Crutchfield-Young algorithm has considerable intuitive appeal, and has a record of success in practice, it is not altogether satisfactory. We are essentially dealing with a problem in statistical inference, and its statistical justification is weak. Because it works by merging, it effectively makes the most complicated model of the process it can. This gross rejection of Occam's Razor is not only ideologically repugnant, but hard to analyze statistically. Finally, the algorithm does not make use of any of the known properties of causal states and ϵ-machines to guide the search, e.g., though the causal states are deterministic, the states it returns often aren't.

This chapter presents a new algorithm which improves on the old Crutchfield-Young algorithm in all these respects. It operates on the opposite principle, namely creating or splitting off new states only when absolutely forced to. I specify the new algorithm, prove its asymptotic reliability or convergence on the true states, and describe its successful function. I then speculate about how the rate of convergence varies with characteristics of the process, such as its statistical complexity C_μ, and make hand-wavy arguments for a particular form of dependence.

Erik van Nimwegen originally suggested the core of this idea, inspired by Bussemaker, Li and Siggia (2000), which, however, looks at "words" in biosequence data and natural-language corpora rather than causal states. (Thanks to Erik as well for providing a preprint of that paper.) The development of this algorithm is joint work with Kristina Klinkner, and a more extensive report on it can be had in Klinkner and Shalizi (2001).

5.1 Reconstructing States by Merging

Previous procedures for reconstructing the states operate by using what one might call compression or merging. The default is that each distinct history encountered in the data is a distinct causal state. Histories are then merged into states on the basis of equality of conditional probabilities of futures, or at least of closeness of those probabilities.

The standard Crutchfield-Young merging algorithm is a tree method. Assume the process takes values from an alphabet $\mathcal{A}$ of size k. Then the algorithm is to build a k-ary tree of some pre-set depth L, where paths through the tree correspond to sequences of observations of length L, obtained by sliding a window through the data stream (or streams, if there are several). If $L = 4$, say, and the sequence $abba$ is encountered, the path in the tree will start at the root node, take the edge labeled a to a new node, then take the outgoing edge labeled b to a third node, then the edge labeled b from that, and finally the edge labeled a to a fifth node, which is a leaf. An edges of the tree is labeled, not just with a symbol, but also with the number of times that edge has been traversed in scanning through the data stream. Call the number on the a_i edge going out of node n, $\nu(a_i|n)$, and the total number of sequences we have entered into the tree N.

The traversal-counts are converted into empirical conditional probabilities by simple division:

$$\hat{\mathrm{P}}_N(a_i|n) = \frac{\nu(a_i|n)}{\sum_{a_j} \nu(a_j|n)}$$

(We write $\hat{\mathrm{P}}_N$ to remind ourselves that the probability estimate is a function of the number of data points N.) Thus attached to each non-leaf node is an empirical conditional distribution for the next symbol. If n has descendants to depth K, then it has (by implication) a conditional distribution for futures of length K.

The merging procedure is now as follows. Consider all nodes with sub-trees of depth $L/2$. Take any two of them. If all the empirical probabilities attached to the edges in their sub-trees are within some constant δ of one another, then the two nodes are equivalent, and they (and their descendants) should be merged with one another. The new node for the root will have incoming links from both the parents of the old nodes. This procedure is to be repeated until no further merging is possible.[1]

All other methods for causal state reconstruction currently in use are also based on merging. Take, for instance, the "topological" or "modal" merging procedure of Perry and Binder (1999). They consider the relationship between histories and futures, both (in the implementation) of length L. Two histories are assigned to the same state if the sets of futures which can succeed them are identical.[2] The distribution over those futures is then estimated for each state, not for each history.

5.1.1 What's Wrong with Merging Methods?

The basic problem with all merging methods is that their default is to treat each history as belonging to its own causal state, creating larger causal states only when they must. The implicit null model of the process is thus the most complicated one that can be devised, given the length of histories available to the algorithm. This seems perverse, especially given computational mechanics's strong commitment to Occam's Razor and the like. Worse, it makes it very hard, if not impossible, to apply standard tools of statistical inference to the estimation procedure.

[1] Since the criterion for merging is not a true equivalence relation (it isn't transitive), the order in which states are examined for merging matters, and various tricks exist for dealing with this. See, e.g., Hanson (1993).

[2] This *is* an equivalence relation, but it isn't causal equivalence.

For instance: what is a reasonable value of δ? Clearly, as the amount of data increases, and the Law of Large Numbers makes empirical probabilities converge to true probabilities, δ should grow smaller. But it is grossly impractical to calculate what δ should be, since the null model itself is so complicated. (Current best practice is to pick δ as though the process were an IID multinomial, which is just the opposite of the algorithm's default estimate!) Furthermore, using the same δ for every pair of nodes is a bad idea, since one node might have been sampled much less often than the other, and so the conditional probabilities in its sub-tree are less accurate than those in the other.

The results summarized in Chapter 4 tell us a lot about what the causal states are like; for instance, they are deterministic, they are Markovian, etc. No existing reconstruction algorithm makes use of this information to guide its search. The Crutchfield-Young algorithm frequently returns a non-deterministic set of states, for instance, which can't possibly be the true causal states.[3] This sort of behavior should be discouraged.

None of this is to say that merging algorithms do not work in practice, since they have. It's even clear that, given enough data, and a small enough δ, if the true causal states can be identified on the basis of finite histories, the Crutchfield-Young algorithm will identify them. Still, their limitations and deficiencies are deeply unsatisfying.

5.2 Reconstructing States by Splitting

5.2.1 Description of the Method

We assume we are given a sequence of length N over the finite alphabet $\mathcal{A}$.[4] We wish to calculate from this a class of states, $\hat{\mathcal{S}}$. Each member $\hat{\sigma}$ of $\hat{\mathcal{S}}$ is a set of histories, or suffixes to histories. The function $\hat{\epsilon}$ maps a finite history $\overleftarrow{s}$ to that $\hat{\sigma}$ containing the longest sequence terminating $\overleftarrow{s}$, i.e., to the state containing the longest suffix of $\overleftarrow{s}$.

Each $\hat{\sigma} \in \hat{\mathcal{S}}$, is associated with a distribution for the next observable $\overrightarrow{S}^1$, i.e., $\mathrm{P}(\overrightarrow{S}^1 = a|\hat{\mathcal{S}} = \hat{\sigma})$ is defined for each $a \in \mathcal{A}$ and each $\hat{\sigma}$. We will call this conditional distribution the *morph* of the state.

The null hypothesis is that the process is Markovian on the basis of the states in $\hat{\mathcal{S}}$,

$$\begin{aligned} \mathrm{P}(\overrightarrow{S}^1 | \overleftarrow{S}^L = as^{L-1}) &= \mathrm{P}(\overrightarrow{S}^1 | \overleftarrow{S}^{L-1} = s^{L-1}) && (5.1)\\ &= \mathrm{P}(\overrightarrow{S}^1 | \hat{\mathcal{S}} = \hat{\epsilon}(s^{L-1})) && (5.2) \end{aligned}$$

We apply a standard statistical test to this hypothesis, e.g. the Kolmogorov-Smirnov test[5], at some specified significance level. (If we use the KS test, we can actually avoid estimating the conditional distribution, and just use the empirical frequency counts.) This controls directly the probability of type I error (rejecting the null when it is true), and generally the KS test has higher power (lower probability of type II error, of accepting the null when it's false) than other, similar tests, such as χ^2 (Rayner and Best 1989). We modify $\hat{\mathcal{S}}$ only when the null is rejected.

I. *Initialization.* Set $L = 0$, and $\hat{\mathcal{S}} = \{\hat{\sigma}_0\}$, where $\hat{\sigma}_0 = \{\emptyset\}$, i.e., $\hat{\sigma}_0$ contains only the null sequence. We assume that the null sequence can be regarded as a suffix of any history, so that initially all histories are mapped to $\hat{\sigma}_0$. The morph of $\hat{\sigma}_0$ is defined by

$$\mathrm{P}(\overrightarrow{S}^1 = a|\hat{\mathcal{S}} = \hat{\sigma}_0) \quad = \quad \mathrm{P}(\overrightarrow{S}^1 = a) \ ,$$

[3]It is sometimes claimed (Jay Palmer, personal communication) that the non-determinism is due to non-stationarity in the data stream. While a non-stationary source can cause the Crutchfield-Young algorithm to return non-deterministic states, the algorithm quit capable of doing this when the source is IID.

[4]The modification to handle multiple sequences, multiple samples from the same process, is discussed at the end of this section.

[5]See Press, Teukolsky, Vetterling and Flannery (1992, sec. 14.3) and Hollander and Wolfe (1999, pp. 178–187) for details of this test.

so the initial model is that the process is a sequence of independent, identically-distributed random variables.

II. *Homogeneity.* We first generate states whose members are homogeneous (Definition 6) for the next symbol — whose parts all have the same morph. Or rather, we generate states whose members have no *significant* differences in their morphs.

1. For each $\hat{\sigma} \in \hat{\mathcal{S}}$, calculate $\hat{\mathrm{P}}_N(\overrightarrow{S}^1 | \hat{\mathcal{S}} = \hat{\sigma})$ — the "distribution" of that state.
   (a) For each sequence $\overleftarrow{s}^L \in \hat{\sigma}$, estimate $\mathrm{P}(\overrightarrow{S}^1 = a | \overleftarrow{S}^L = \overleftarrow{s}^L)$. The naive maximum-likelihood estimate,
   $$\hat{\mathrm{P}}_N(\overrightarrow{S}^1 = a | \overleftarrow{S}^L = \overleftarrow{s}^L) \quad = \quad \frac{\nu(\overleftarrow{S}^L = \overleftarrow{s}^L, \overrightarrow{S}^1 = a)}{\nu(\overleftarrow{S}^L = \overleftarrow{s}^L)} \; ,$$
   is simple and well-adapted to the later part of the procedure, but other estimators could be used. This distribution is the morph of $\overleftarrow{s}^L$.
   (b) The morph of $\hat{\sigma}$ is the weighted average of the morphs of the sequences $\overleftarrow{s}^L \in \hat{\sigma}$, with weights proportional to $\nu(\overleftarrow{S}^L = \overleftarrow{s}^L)$.
   (c) For the special case when $L = 0$ and the only history is the null sequence, see above.
2. For each $\hat{\sigma} \in \hat{\mathcal{S}}$, test the null (Markov) hypothesis. For each length L sequence $\overleftarrow{s}^L \in \hat{\sigma}$ and each $a \in \mathcal{A}$, generate the suffix of length $L + 1$ $a\overleftarrow{s}^L$ — a *child suffix* of $\overleftarrow{s}^L$.
   (a) Estimate the morph of $a\overleftarrow{s}^L$ by the same method as used above.
   (b) Test whether the morphs of $a\overleftarrow{s}^L$ and $\hat{\sigma}$ differ significantly.
   (c) If they do, then it is worthwhile to distinguish $a\overleftarrow{s}^L$ from $\overleftarrow{s}^L$, and from all the other histories in $\hat{\sigma}$.
      i. Test whether there are any states $\hat{\mathcal{S}}$ whose morphs do *not* differ significantly from that of $a\overleftarrow{s}^L$. If so, add $a\overleftarrow{s}^L$ to the state whose morph it matches most closely, as measured by the score of the significance test[6].
      ii. If the morph of $a\overleftarrow{s}^L$ is significantly different from the morphs of all existing states, create a new state and add $a\overleftarrow{s}^L$ to it, with its morph.
      iii. Generate all the other child suffixes of $\overleftarrow{s}^L$, and assign them to the states whose morphs they match most closely.
      iv. Delete $\overleftarrow{s}^L$ (and any of its ancestors)[7] from $\hat{\sigma}$.
      v. Recalculate the morphs of states from which sequences have been added or deleted.
   (d) If the morph of $a\overleftarrow{s}^L$ does not differ significantly from that of $\hat{\sigma}$, add $a\overleftarrow{s}^L$ to $\hat{\sigma}$.
3. Increment L by one.
4. Repeat steps 1–3 until we reach some preset maximum length $L_{\max}$.

At the end of this procedure, no history is in a state whose morph is significantly different from its own. Moreover, every state's morph is significantly different from every other state's morph. The causal states have this property, but they are also deterministic, and we need another procedure to "determinize" $\hat{\mathcal{S}}$.

III. *Determinization.*

[6] Actually, which of these states $a\overleftarrow{s}^L$ is assigned to is irrelevant in the limit where $N \to \infty$; but this choice is convenient and plausible.

[7] If any of the ancestors of $\overleftarrow{s}^L$ are around as suffixes, then they must also be in $\hat{\sigma}$.

1. For each state $\hat{\sigma} \in \hat{\mathcal{S}}$

   (a) For each $a \in \mathcal{A}$

      i. Calculate $\hat{\epsilon}(\overleftarrow{s} a)$ for all $\overleftarrow{s} \in \hat{\sigma}$ — these are the *successor states on* a of the histories.

      ii. If there is only one successor state on a, go on to the next a.

      iii. If there are $n \geq 2$ successor states on a, create $n-1$ new states, moving histories into them from $\hat{\sigma}$ so that all histories in the new states and $\hat{\sigma}$ now have the same successor on a. Go back to 1.

   (b) If every history $\overleftarrow{s}$ in $\hat{\sigma}$ has the same successor on a, for every a, go on to the next state.

2. For each state, output the list of sequences in the state, the conditional probability for each symbol $a \in \mathcal{A}$, and the successor on a.

It is clear that this procedure will terminate (in the worst case, when every history is assigned to its own state), and that when it terminates, $\hat{\mathcal{S}}$ will be deterministic. Moreover, because we create the deterministic states by splitting the homogeneous states, the deterministic states remain homogeneous.

Now, by Theorem 6, the causal states are the minimal states which have a homogeneous distribution for the next symbol and are deterministic. If we had access to the exact conditional distributions, therefore, and did not have to estimate the morphs, this procedure would return the causal states. Instead it returns a set of states which in some sense cannot be significantly distinguished from them.

5.2.2 Reliability of Reconstruction

The road to wisdom? — Well, it's plain
and simple to express:

Err
and err
and err again
but less
and less
and less.

— Piet Hein (1966, p. 34)

We wish to show that the algorithm we have given will, like the Crutchfield-Young algorithm, return the correct causal states, if $L_{\max}$ is sufficiently large, and $N \to \infty$. To be more precise, assume that $L_{\max}$ is large enough that $\overleftarrow{s}^{L_{\max}}$ is sufficient to place the system in the correct causal state. We wish to show that the probability that $\hat{\mathcal{S}} \neq \boldsymbol{\mathcal{S}}$ goes to zero as $N \to \infty$. For definiteness, we'll assume here that the algorithm employs the KS test, though nothing vital hinges on that.

Nothing can go wrong in procedure I.

Two sorts of error are possible in procedure II. A history $\overleftarrow{s}$ can be put in a class with $\overleftarrow{s}'$, even though $\overleftarrow{s} \not\sim_{\epsilon} \overleftarrow{s}'$; or two histories which are causally equivalent could be assigned to different states, $\overleftarrow{s} \sim_{\epsilon} \overleftarrow{s}'$ but $\hat{\epsilon}(\overleftarrow{s}) \neq \hat{\epsilon}(\overleftarrow{s}')$. Can we show that these events become vanishingly rare as $N \to \infty$?

Each time we see $\overleftarrow{s}$, the next symbol $\overrightarrow{S}^1$ is independent of what the next symbol is every other time we see $\overleftarrow{s}$; this is what it means for $L_{\max}$ to be large enough to make the process Markovian. Hence our naive maximum-likelihood estimate of the morph, $\hat{\mathrm{P}}_N(\overrightarrow{S}^1 \mid \overleftarrow{S} = \overleftarrow{s})$, is the empirical mean of IID random variables, and by the strong law of large numbers, converges on $\mathrm{P}(\overrightarrow{S}^1 \mid \overleftarrow{S} = \overleftarrow{s})$ with probability 1 as $N \to \infty$.[8] If

[8] *For probabilists.* Technically, the strong law just tells us this happens for each realization of $\overrightarrow{S}^1$ separately. Since there are only a finite number of them, however, it still is true for them all jointly, and so for the distribution.

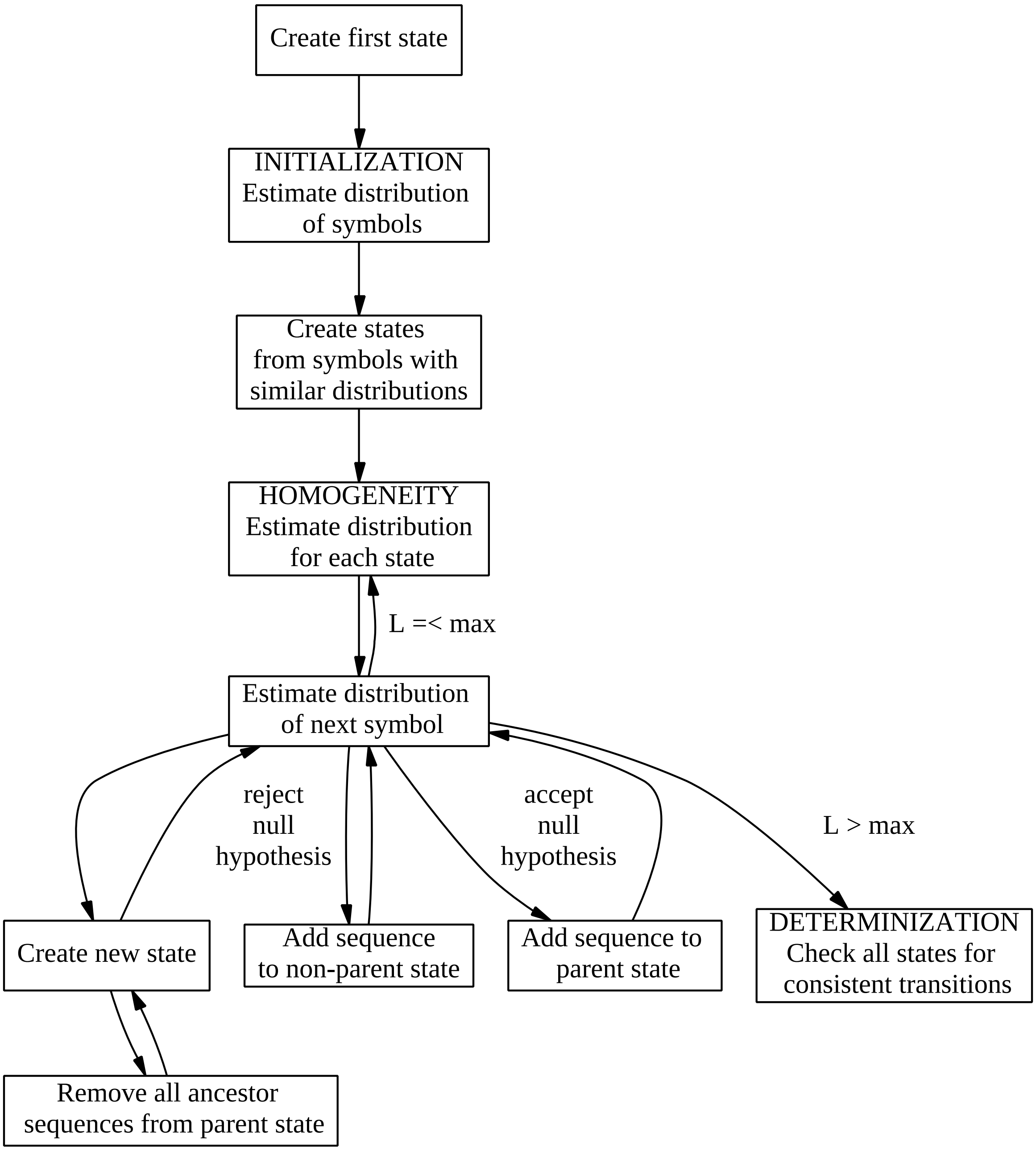

Figure 5.1: Flow chart for the operation of the state-splitting reconstruction algorithm.

$\overleftarrow{s} \sim_\epsilon \overleftarrow{s}'$, then $\mathrm{P}(\overrightarrow{S}^1 \mid \overleftarrow{S} = \overleftarrow{s}) = \mathrm{P}(\overrightarrow{S}^1 \mid \overleftarrow{S} = \overleftarrow{s}')$. Therefore, $\forall a \in \mathcal{A}$

$$\left| \hat{\mathrm{P}}_N(\overrightarrow{S}^1 = a \mid \overleftarrow{S} = \overleftarrow{s}) - \hat{\mathrm{P}}_N(\overrightarrow{S}^1 = a \mid \overleftarrow{S} = \overleftarrow{s}') \right| \quad \rightarrow \quad 0$$

as $N \rightarrow 0$, at least in probability. Therefore the KS test statistic for the difference between the morphs of $\overleftarrow{s}$ and $\overleftarrow{s}'$ will converge to zero, and the probability that two histories which belong to the same causal state will be assigned to different states goes to zero.

If $\overleftarrow{s} \not\sim_\epsilon \overleftarrow{s}'$, then there are two possibilities. (1) $\exists a \in \mathcal{A}$ such that $\mathrm{P}(\overrightarrow{S}^1 = a \mid \overleftarrow{S} = \overleftarrow{s}) \neq \mathrm{P}(\overrightarrow{S}^1 = a \mid \overleftarrow{S} = \overleftarrow{s}')$. Call the difference between those probabilities p. Then

$$\left| \hat{\mathrm{P}}_N(\overrightarrow{S}^1 = a \mid \overleftarrow{S} = \overleftarrow{s}) - \hat{\mathrm{P}}_N(\overrightarrow{S}^1 = a \mid \overleftarrow{S} = \overleftarrow{s}') \right| \quad \rightarrow \quad p$$

in probability, and so the KS test will separate $\overleftarrow{s}$ and $\overleftarrow{s}'$. (2) The morphs are the same, but $\exists s^K \in \mathcal{A}^K$ such that $\mathrm{P}(\overrightarrow{S}^1 \mid \overleftarrow{S} = \overleftarrow{s} s^K) \neq \mathrm{P}(\overrightarrow{S}^1 \mid \overleftarrow{S} = \overleftarrow{s}' s^K)$. Then (by the previous argument) $\overleftarrow{s} s^K$ and $\overleftarrow{s}' s^K$ will belong to different states, at least in the limit, and so, recursively, $\overleftarrow{s}$ and $\overleftarrow{s}'$ will be separated by procedure III.

Nothing can go wrong in procedure III.

Therefore $\mathrm{P}(\hat{\mathcal{S}} \neq \boldsymbol{\mathcal{S}}) \rightarrow 0$ as $N \rightarrow \infty$.

In the terminology of mathematical statistics (Cramér 1945), we have just shown that the algorithm is a *consistent estimator* of the causal states. In that of machine learning theory (Kearns and Vazirani 1994; Vapnik 2000), it is *probably approximately correct*. In that of the philosophy of science (Glymour 1992; Spirtes, Glymour and Scheines 2001; Kelly 1996) it is *reliable*.

5.2.3 Advantages of the Method

The main advantages of this algorithm are, naturally enough, the opposites of what I said were the disadvantages of the Crutchfield-Young algorithm.

The implicit null model is that the process is IID, which is the simplest model we could use. We add states only when the current model is definitely rejected, and so introduce complications (and complexity) only as the data demand them. By using a proper hypothesis test, instead of a simple cut-off as in the Crutchfield-Young algorithm, we take in to account the effects of sample size and the non-trivial form of the distribution. Adjusting the significance level directly controls the rate at which the algorithm creates spurious states. It also indicates our fear of over-fitting, or our willingness to accept additional complexity in return for better fits to the data. Strict fidelity to Occam not only lets us bask in the warmth of methodological virtue, it gives us a better handle on what our program is doing.

The algorithm makes full use of the known properties of the causal states — their homogeneity, their determinism, their Markovianity. This greatly reduces the space of state classes in which the algorithm must search, and so should significantly improve the rate of convergence (see below). By using homogeneity and determinism, we never have to look at futures of length greater than one, which is good both for the time it takes the algorithm to run and for the accuracy of the results. By keeping everything deterministic and Markovian, it should be possible to analytically calculate error rates (size, power, and even severity (Mayo 1996; Mayo and Spanos 2000)), at least in the asymptotic regime, by adapting results in Billingsley (1961).

5.2.3.1 Problems with the Method

We have no assurance that the set of states produced by this algorithm will be minimal. Currently there is no penalty for making spurious distinctions which do not impair prediction. Because we can only use finite quantities of data, it is always possible that, simply through bad luck and sampling errors, two histories which belong in the same causal state will have significantly different sample-distributions of futures, and be split. This might be avoided by lowering the significance level in the KS test, and so splitting only when the difference in the conditional distribution of futures is larger, but past a certain point, this will tend to lump together states which should be split — the old trade-off between false positives and false negatives in statistics.

We need to fix a value for $L_{\max}$. Normally, we imagine that this should be as large as time and memory constraints will allow — if there isn't enough data to go back that far, the significance test will handle it automatically. It is possible, however, to independently test for the Markov order of the data stream (Billingsley 1961; van der Heyden, Diks, Hoekstra and DeGoede 1998), and so place bounds on $L_{\max}$, if we want to.

The algorithm returns a single state class. But for finite N, there are generally lots of others which would do at least as well on all the tests. The one the algorithm returns depends on such details of its innards as the order in which it generates child suffixes. Rather than providing a point estimate of the causal states, it would be nice if it gave us all the adequate state classes, all the ones whose performance is over a certain threshold; this would be a kind of confidence region for the causal states. Since doing that is, for combinatorial reasons, really impractical, it might be better to randomize such things as the orders of generations and checking, and re-run the algorithm repeatedly to sample the confidence region.[9]

Lastly, any pattern which is strictly sofic — where there are subwords of allowed words which are forbidden — the algorithm will fail to pick up the pattern. A particularly annoying example, suggested by Cris Moore, is the language which consists of all strings where the *total* numbers of zeroes and ones are even. The difficulty here is that while the entire data-stream could not consist of (say) the string `000111`, that could occur as a substring (of `00011101` or `00011110` or even `000111000111`), and there is no way of telling whether or not the string as a whole is admissible until we reach its end. Existing merging algorithms also fail on this example, however[10] It's not clear how to work around this.

5.3 Some Notes on an Implementation

We implemented the algorithm in C++, running on Sun workstations. For reasons of speed and memory conservation, the conditional probability distributions were stored as a parse tree, rather as in the Crutchfield-Young algorithm. We used the Kolmogorov-Smirnov test, modifying slightly the code in Press *et al.* (1992), and, following statistical convention, set the significance level to 0.05. The absolute-worst-case run time is $O(N + |\mathcal{A}|^{L_{\max}+1})$ (Klinkner and Shalizi 2001).

We have tested our implementation on a range of processes where we can work out the correct causal states by hand. These include multinomial IID processes, periodic sequences, stationary Markov models, hidden Markov models, and master equations/biased random walks. None of the tests cases has had more than 7 states. In every case, with $N = 1000$ and $L_{\max} = 5$, the code returns the correct states at least 95% of the time. All cases were computed much faster than the worst-case analysis would lead us to fear. While these preliminary results are too scanty to support detailed quantitative analysis, qualitatively, things look good.

Currently, the algorithm scans in only a single time series. It will be easy to modify the code so that it can be given multiple series, storing them all in the same parse tree. This assumes that they all come from the same source, but that's the only way that it makes sense to use multiple series in reconstructing a single ϵ-machine anyway.

5.4 Statistical Analysis and Rates of Convergence

There are some statistical properties of the algorithm which need careful analysis.

One is value of the significance level. If we keep it at .05, then we can expect that, out of twenty times when we should not split a state, we will do so once. This will effect the error statistics (see below), but we

[9] The `TETRAD` algorithm for causal discovery in graphs does something like this (Spirtes, Glymour and Scheines 2001).

[10] The Crutchfield-Young algorithm works very well, however, on languages with parity constraints on *blocks* of symbols, say, ones only occur in blocks of even length.

would also like to know about how often we will not split states when we should. This probability, essentially the *power* of the test, is not directly given by the significance level, but it should be possible to calculate using the tools of statistical inference for Markov chains (Billingsley 1961). This in turn will tell us what is a reasonable value for the significance level.

The second major issue is the scaling of the error statistics (Mayo 1996), or the rates of convergence. We have seen that these go to zero as $N \rightarrow \infty$. Infinity is a long time, however, and we'd like to know how long we need to wait for the error to be *small*. More precisely, suppose we introduce a measure of the error involved in using the states estimated from N data points, $\hat{\mathcal{S}}_N$, rather than the true causal states — call this error $err(\hat{\mathcal{S}}_N)$. Then we would like to find a function $n(\delta, \varepsilon)$ such that, if $N > n(\delta, \varepsilon)$,

$$\mathrm{P}(err(\hat{\mathcal{S}}_N) \geq \varepsilon) \quad \leq \quad 1 - \delta \; . \tag{5.3}$$

Alternately, we fix δ and invert n to get $\varepsilon(N, \delta)$ — given N data points, with confidence level $1-\delta$, the error is ε or less. The dependence of ε on N for fixed δ is the rate of convergence of the algorithm.

The exact rate of convergence is likely to be complicated and highly dependent on the characteristics of the process generating the data, i.e., on precisely the things we want the algorithm to tell us about. We would therefore like to find functions which bound $n(\delta, \varepsilon)$ or $\varepsilon(N, \delta)$, where the bounds are fairly tight, but hold across a wide range of processes, and the bounding functions can be calculated in terms of very general characteristics; something like the statistical complexity would be ideal. We want, if not exactly a uniform rate of convergence in the technical sense, then something of that ilk.

Under the circumstances we've assumed, it's easy to adapt results from large deviation and empirical process theory (Ellis 1985; Pollard 1984; Feng and Kurtz 2000) to see that the empirical conditional distributions $\hat{\mathrm{P}}_N(\overrightarrow{S}^1 \mid \overleftarrow{S} = \overleftarrow{s})$ should converge on $\mathrm{P}(\overrightarrow{S}^1 \mid \overleftarrow{S} = \overleftarrow{s})$ exponentially in N. This does not imply that the global error converges exponentially, however. In fact, based on studies of the rate of convergence of other statistical estimators, especially for stochastic processes (Bickel and Ritov 1995; van de Geer 2000; Bosq 1998) we conjecture that the rate of convergence will be polynomial in N and in ${C_\mu}^{-1}$. Generally such rates of convergence results depend very strongly on the size of the space of possible models the estimation algorithm must search through, so we also conjecture that the splitting algorithm, with its constraints of determinism and the like, will converge faster than the Crutchfield-Young algorithm. (For preliminary results on the error statistics of the Crutchfield-Young algorithm see Crutchfield and Douglas 1999.)

Establishing analytical bounds on the rate of convergence is likely to be extremely tricky, though there are promising hints in machine learning theory (Evans, Rajagopalan and Vazirani 1993), in addition to empirical process theory and large deviations theory. A numerical-experimental approach to the problem would be to fix on a global error measure, such as the relative entropy between the actual distribution over sequences and that predicted by $\hat{\mathcal{S}}_N$, and measure how it varies with N and with characteristics of the process, such as C_μ. We could similarly look at $\hat{C}_\mu$ as a function of N, where we expect the mean to converge on the true value from below, and more rapidly the smaller C_μ is.

Chapter 6

Connections to Other Approaches

6.1 Time Series Modeling

The goal of time series modeling is to predict the future of a measurement series on the basis of its past. Broadly speaking, this can be divided into two parts: identify equivalent pasts and then produce a prediction for each class of equivalent pasts. That is, we first pick a function $\eta : \overleftarrow{\mathbf{S}} \mapsto \boldsymbol{\mathcal{R}}$ and then pick another function $p : \boldsymbol{\mathcal{R}} \mapsto \overrightarrow{\mathbf{S}}$. Of course, we can choose for the range of p futures of some finite length (length 1 is popular) or even choose distributions over these. While practical applications often demand a single definite prediction — "You will meet a tall dark stranger", there are obvious advantages to predicting a distribution — "You have a .95 chance of meeting a tall dark stranger and a .05 chance of meeting a tall familiar albino." Clearly, the best choice for p is the actual conditional distribution of futures for each $\rho \in \boldsymbol{\mathcal{R}}$. Given this, the question becomes what the best $\boldsymbol{\mathcal{R}}$ is; i.e., What is the best η? At least in the case of trying to understand the whole of the underlying process, the best η is, unambiguously, ϵ. Computational mechanics subsumes the whole of traditional time series modeling.

Computational mechanics — in its focus on letting the process speak for itself through (possibly impoverished) measurements — follows the spirit that motivated one approach to experimentally testing dynamical systems theory. Specifically, it follows in spirit the methods of reconstructing "geometry from a time series" introduced by Packard, Crutchfield, Farmer and Shaw (1980) and Takens (1981). A closer parallel is found, however, in later work on estimating minimal equations of motion from data series (Crutchfield and McNamara 1987).

6.2 Decision-Theoretic Problems

The classic focus of decision theory is "rules of inductive behavior" (Neyman 1950; Blackwell and Girshick 1954; Luce and Raiffa 1957). The problem is to chose functions from observed data to courses of action that possess desirable properties. This task has obvious affinities to considering the properties of ϵ and its rivals η. We can go further and say that what we have done *is* consider a decision problem, in which the available actions consist of predictions about the future of the process. The calculation of the optimum rule of behavior in general faces formidable technicalities, such as providing an estimate of the utility of every different course of action under every different hypothesis about the relevant aspects of the world. Remarkably enough, however, we can show that, for anything which it's reasonable to call a decision problem, the optimal rule of behavior can be implemented using ϵ (Appendix D).

6.3 Stochastic Processes

Clearly, the computational mechanics approach to patterns and pattern discovery involves stochastic processes in an intimate and inextricable way. Probabilists have, of course, long been interested in using information-theoretic tools to analyze stochastic processes, particularly their ergodic behavior (Billingsley 1965; Gel'fand and Yaglom 1956; Caines 1988; Gray 1990). There has also been considerable work in the hidden Markov model and optimal prediction literatures on inferring models of processes from data or from given distributions (Blackwell and Koopmans 1957; Ito, Amari and Kobayashi 1992; Algoet 1992; Upper 1997; Jaeger 2000). To the best of my knowledge, however, these two approaches have not been previously combined.

Perhaps the closest approach to the spirit of computational mechanics in the stochastic process literature is, surprisingly, the now-classical theory of optimal prediction and filtering for stationary processes, developed by Wiener and Kolmogorov (Kolmogorov 1941; Wiener 1949; Wiener 1958; Schetzen 1989; Wiener 1961). The two theories share the use of information-theoretic notions and the unification of prediction and structure. So far as I've been able to learn, however, no one has ever used this theory to explicitly identify causal states and causal structure, leaving these implicit in the mathematical form of the prediction and filtering operators. Moreover, the Wiener-Kolmogorov framework forces us to sharply separate the linear and nonlinear aspects of prediction and filtering, because it has a great deal of trouble calculating nonlinear operators (Wiener 1958; Schetzen 1989). Computational mechanics is completely indifferent to this issue, since it packs *all* of the process's structure into the ϵ-machine, which is equally calculable in linear or strongly nonlinear situations[1].

6.4 Formal Language Theory and Grammatical Inference

A formal language is a set of symbol strings ("words" or "allowed words") drawn from a finite alphabet. Every formal language may be described either by a set of rules (a "grammar") for creating all and only the allowed words, by an abstract automaton which also generates the allowed words, or by an automaton which accepts the allowed words and rejects all "forbidden" words.[2] ϵ-machines, stripped of probabilities, correspond to such automata — generative in the simple case or classificatory, if we add a reject state and move to it when none of the allowed symbols are encountered.

Since Chomsky (1956, 1957), it has been known that formal languages can be classified into a hierarchy, the higher levels of which have strictly greater expressive power. The hierarchy is defined by restricting the form of the grammatical rules or, equivalently, by limiting the amount and kind of memory available to the automata. The lowest level of the hierarchy is that of regular languages, which may be familiar to Unix-using readers as regular expressions. These correspond to finite-state machines, for which relatives of the minimality and uniqueness theorems are well known (Lewis and Papadimitriou 1998), and the construction of causal states is analogous to "Nerode equivalence classing" (Hopcroft and Ullman 1979). Our theorems, however, are *not* restricted to this low-memory, non-stochastic setting; for instance, they apply to hidden Markov models with both finite and infinite numbers of hidden states (Upper 1997).

The problem of learning a language from observational data has been extensively studied by linguists, and by computer scientists interested in natural-language processing. Unfortunately, well developed learning techniques exist only for the two lowest classes in the Chomsky hierarchy, the regular and the context-free languages. (For a good account of these procedures see Charniak (1993) and Manning and Schütze (1999).) Adapting and extending this work to the reconstruction of ϵ-machines should form a useful area of future research (cf. the "hierarchical ϵ-machine reconstruction" of Crutchfield (1994a)).

[1]For more on the nonlinear Wiener theory, see Section 7.6.
[2]For more on formal languages and automata, see Appendix A.4.

6.5 Computational and Statistical Learning Theory

The goal of computational learning theory (Kearns and Vazirani 1994; Vapnik 2000) is to identify algorithms that quickly, reliably, and simply lead to good representations of a target "concept". The latter is typically defined to be a binary dichotomy of a certain feature or input space. Particular attention is paid to results about "probably approximately correct" (PAC) procedures (Valiant 1984): those having a high probability of finding members of a fixed "representation class" (e.g., neural nets, Boolean functions in disjunctive normal form, or deterministic finite automata). The key word here is "fixed"; as in contemporary time-series analysis, practitioners of this discipline acknowledge the importance of getting the representation class right. (Getting it wrong can make easy problems intractable.) In practice, however, they simply take the representation class as a given, even assuming that we can always count on it having at least one representation which *exactly* captures the target concept. Although this is in line with implicit assumptions in most of mathematical statistics, it seems dubious when analyzing learning in the real world (Crutchfield 1994a; Boden 1994; Thornton 2000).

In any case, the preceding development made no such assumption. One of the goals of computational mechanics is, exactly, *discovering* the best representation. This is not to say that the results of computational learning theory are not remarkably useful and elegant, nor that one should not take every possible advantage of them in implementing ϵ-machine reconstruction. But these theories belong more to statistical inference, particularly to algorithmic parameter estimation, than to foundational questions about the nature of pattern and the dynamics of learning.

6.6 Description-Length Principles and Universal Coding Theory

Rissanen's *minimum description length* (MDL) principle, most fully described in Rissanen (1989), is a procedure for selecting the most concise generative model out of a family of models that are all statistically consistent with given data. The MDL approach starts from Shannon's results on the connection between probability distributions and codes.

Suppose we choose a representation that leads to a class $\mathcal{M}$ of models and are given data set X. The MDL principle enjoins us to pick the model $\mathrm{M} \in \mathcal{M}$ that minimizes the sum of the length of the description of X given M, plus the length of description of M given $\mathcal{M}$. The description length of X is taken to be $-\log \mathrm{P}(X|\mathrm{M})$; cf. Eq. A.7. The description length of M may be regarded as either given by some coding scheme or, equivalently, by some distribution over the members of $\mathcal{M}$. (Despite the similarities to model estimation in a Bayesian framework (Lindley 1972), Rissanen does not interpret this distribution as a Bayesian prior or regard description length as a measure of evidential support.)

The construction of causal states is somewhat similar to the states estimated in Rissanen's *context* algorithm (Rissanen 1983; Rissanen 1989; Bühlmann and Wyner 1999), and to the "vocabularies" built by universal coding schemes, such as the popular Lempel-Ziv algorithm (Lempel and Ziv 1976; Ziv and Lempel 1977). Despite the similarities, there are significant differences. For a random source — for which there is a single causal state — the context algorithm estimates a number of states that diverges (at least logarithmically) with the length of the data stream, rather than inferring a single state, as ϵ-machine reconstruction would. Moreover, the theory makes no reference to encodings of rival models or to prior distributions over them; $C_\mu(\boldsymbol{\mathcal{R}})$ is *not* a description length.

6.7 Measure Complexity

Grassberger (1986) proposed that the appropriate measure of the complexity of a process was the "minimal average Shannon information needed" for optimal prediction. This *true measure complexity* was to be taken as the Shannon entropy of the states used by some optimal predictor. The same paper suggested that it could be approximated (from below) by the excess entropy; there called the *effective measure complexity*, as noted in Section 4.4 above. This is a position closely allied to that of computational mechanics, to Rissanen's MDL principle, and to the minimal embeddings introduced by attractor-reconstruction methods.

In contrast to computational mechanics, however, the key notion of "optimal prediction" was left undefined, as were the nature and construction of the states of the optimal predictor. In fact, the predictors used required knowing the process's underlying equations of motion. Moreover, the statistical complexity $C_\mu(\boldsymbol{\mathcal{S}})$ differs from the measure complexities in that it is based on the well defined causal states, whose optimal predictive powers are in turn precisely defined. Thus, computational mechanics is an operational and constructive formalization of the insights expressed in Grassberger (1986).

6.8 Hierarchical Scaling Complexity

Introduced in Badii and Politi (1997, ch. 9), this approach seeks, like computational mechanics, to extend certain traditional ideas of statistical physics. In brief, the method is to construct a hierarchy of n^{th}-order Markov models and examine the convergence of their predictions with the real distribution of observables as $n \to \infty$. The discrepancy between prediction and reality is, moreover, defined information theoretically, in terms of the relative entropy or Kullback-Leibler distance (Kullback 1968; Cover and Thomas 1991). (I've not used this quantity.) The approach implements Weiss's discovery that for finite-state sources there is a structural distinction between block-Markovian sources (*subshifts of finite type*) and *sofic systems*. Weiss showed that, despite their finite memory, sofic systems are the limit of an infinite series of increasingly larger block-Markovian sources (Weiss 1973).

The hierarchical-scaling-complexity approach has several advantages, particularly its ability to handle issues of scaling in a natural way (see Badii and Politi (1997, sec. 9.5)). Nonetheless, it does not attain all the goals set in Section 2.3.5. Its Markovian predictors are so many black boxes, saying little or nothing about the hidden states of the process, their causal connections, or the intrinsic computation carried on by the process. All of these properties are manifest from the ϵ-machine. A productive line of future work would be to investigate the relationship between hierarchical scaling complexity and computational mechanics, and to see whether they can be synthesized. Along these lines, hierarchical scaling complexity is sort of reminiscent of hierarchical ϵ-machine reconstruction (Crutchfield 1994a).

6.9 Continuous Dynamical Computing

Using dynamical systems as computers has become increasingly attractive over the last ten years or so among physicists, computer scientists, and others exploring the physical basis of computation (Huberman 1985; Moore 1996; Moore 1998; Orponen 1997; Blum, Shub and Smale 1989). These proposals have ranged from highly abstract ideas about how to embed Turing machines in discrete-time nonlinear continuous maps (Crutchfield and Young 1990; Moore 1990) to, more recently, schemes for specialized numerical computation that could in principle be implemented in current hardware (Sinha and Ditto 1998). All of them, however, have been synthetic, in the sense that they concern *designing* dynamical systems that implement a given desired computation or family of computations. In contrast, one of the central questions of computational mechanics is exactly the converse: *given* a dynamical system, how can one detect what it is intrinsically computing?

Having a mathematical basis and a set of tools for answering this question are important to the synthetic, engineering approach to dynamical computing. Using these tools we may be able to discover novel forms of computation embedded in natural processes that operate at higher speeds, with less energy or with fewer physical degrees of freedom than currently possible.

Chapter 7

Transducers with Memory

> We watch an ant make his laborious way across and wind- and wave-molded beach. He moves ahead, angles to the right to ease his climb up a steep dunelet, detours around a pebble, stops for a moment to exchange information with a compatriot. So as not to anthropomorphize about his purposes, I sketch the path on a piece of paper. It is a sequence of irregular, angular segments — not quite a random walk, for it has an underlying sense of direction, of aiming towards a goal . . . .
>
> Viewed as a geometric figure, the ant's path is irregular, complex, hard to describe. But its complexity is really a complexity in the surface of the beach, not a complexity in the ant. On that same beach another small creature with a home at the same place as the ant might well follow a very similar path . . . .
>
> The ant, viewed as a behaving system, is quite simple. The apparent complexity of its behavior over time is largely a reflection of the complexity of the environment in which it finds itself.
>
> — Herbert Simon (1996, pp. 51–52)

7.1 Introduction

The previous chapters have developed the computational mechanics for memoryless transducers and for time series. We now "combines our information" to deal with transducers with memory. The picture is that one series, called the input, is fed into a transducer, box (or other physical process), resulting in an output series. This differs from the case of memoryless transduction because the transducer has internal states, and so a kind of memory for both the past of the input process and its own internal dynamics (which may well be stochastic). The goal is to be able to identify the internal states of the transducer and their structure of connection — to find the ϵ-transducer.

Put another way: we have two time series, and the future values of the output are a stochastic functional of the history of the input. We want to put this relationship in "transducer form," replacing the stochastic functional of the series with a stochastic function of an internal or hidden state of a transducer, which in turn is a functional of the history. That is, we want to represent these relationships by means of a generalization of what automata theory calls "finite state transducers" or "sequential machines" (Moore 1956; Booth 1967; Hartmanis and Stearns 1966; Carroll and Long 1989). We won't assume that we'll need only a finite number of states.

7.1.1 Notation and Assumptions

Adapting the notation of Chapter 4 in the obvious way, write the stochastic process of the input as $\overleftrightarrow{X}$, its past as $\overleftarrow{X}$, and its future as $\overrightarrow{X}$. It takes values from the finite alphabet $\mathcal{A}$. The symbols $\overleftrightarrow{Y}$, $\overleftarrow{Y}$, $\overrightarrow{Y}$, $\mathcal{B}$ serve the same role for the output process.

Both input and output values are in general multidimensional variables, but we don't care about that.

Write the set of all possible input histories as $\overleftarrow{\mathbf{X}}$. Similarly, write $\overleftarrow{\mathbf{Y}}$ for the set of all possible output histories.

7.2 Simplifying the Transducer Problem

It is commonly assumed that you completely specify a transducer by giving the conditional probabilities of all finite-length input-output pairs. That is, you need only specify $\mathrm{P}(Y_{n+1}, Y_n \ldots Y_1 | X_n, \ldots X_0)$ to completely specify the behavior of the transducer. I've never seen a demonstration that this is enough, but it's hard to see what else there could be, and in any case I shall appeal to proof-by-consensus.

Assume it is true, and factor that conditional probability as follows:

$$\mathrm{P}(\overrightarrow{Y}^{L} = \overrightarrow{y}^{L-1} b | \overleftarrow{X}^{L} = \overleftarrow{x}^{L-1} a) \tag{7.1}$$

$$= \mathrm{P}(\overrightarrow{Y}_L = b | \overrightarrow{Y}^{L-1} = \overrightarrow{y}^{L-1}, \overleftarrow{X}^{L} = \overleftarrow{x}^{L-1} a) \mathrm{P}(\overrightarrow{Y}^{L-1} = \overrightarrow{y}^{L-1} | \overleftarrow{X}^{L} = \overleftarrow{x}^{L-1} a) \tag{7.2}$$

$$= \mathrm{P}(\overrightarrow{Y}_L = b | \overrightarrow{Y}^{L-1} = \overrightarrow{y}^{L-1}, \overleftarrow{X}^{L} = \overleftarrow{x}^{L-1} a) \mathrm{P}(\overrightarrow{Y}^{L-1} = \overrightarrow{y}^{L-1} | \overleftarrow{X}^{L-1} = \overleftarrow{x}^{L-1}) \tag{7.3}$$

In the last line, I assumed that the future of the input is independent of the past of the output, given the past of the input. This is true just when there is no feedback from output to input. I'll deal with the feedback case below (Section 7.5).

Clearly, we can repeat this factoring with the last factor, $\mathrm{P}(\overrightarrow{Y}^{L-1} = \overrightarrow{y}^{L-1} | \overleftarrow{X}^{L-1} = \overleftarrow{x}^{L-1})$, since it has the same form as our original term. Thus, to get all the conditional probabilities needed for the transducer, it is enough to know all the probabilities of the form $\mathrm{P}(\overrightarrow{Y}^{1} | \overleftarrow{X}^{L}, \overleftarrow{Y}^{L})$. We then build up the probabilities of output sequences by multiplying these next-output conditional distributions together.[1]

Reverting to our usual habit of considering a semi-infinite history, this means that we want conditional probabilities of the form $\mathrm{P}(\overrightarrow{Y}^{1} | \overleftarrow{X}, \overleftarrow{Y})$; all the other conditional probabilities we require can be obtained from this distribution by "marginalizing" the histories down to the needed finite length. Finding transducer states reduces to finding states which "get right" the next output, given the complete input and output histories.

7.3 Effective Transducer States

The definitions of effective states, of predictive ability and of statistical complexity all transfer in the obvious way, except that I define alternate states as equivalence classes over the *joint* history of inputs and outputs. I'll give all these definitions over again for convenience here.

Definition 20 (Joint History) *The joint history of a transducer system is the random variable $(\overleftarrow{X}, \overleftarrow{Y})$, which takes values from the space $\overleftarrow{\mathbf{X}} \times \overleftarrow{\mathbf{Y}}$. $(\overleftarrow{x}, \overleftarrow{y}) \oplus (a, b)$ denotes the joint history obtained by appending a to the input history and b to the output history, $(\overleftarrow{x}\, a, \overleftarrow{y}\, b)$.*

Definition 21 (Effective States of Transducers) *Transducer effective states are equivalence classes of joint histories. To each class of effective states $\boldsymbol{\mathcal{R}}$ there corresponds a function $\eta : \overleftarrow{\mathbf{X}} \times \overleftarrow{\mathbf{Y}} \mapsto \boldsymbol{\mathcal{R}}$. The random variable for the current effective state is $\mathcal{R}$, its realizations ρ.*

Definition 22 ***(Predictive Power for Transducer Effective States)*** *The predictive power of $\boldsymbol{\mathcal{R}}$ is measured by the entropy of future outputs conditional on the present effective state, and the future inputs, $H[\overrightarrow{Y}^{L} | \mathcal{R}, \overrightarrow{X}^{L-1}]$. $\boldsymbol{\mathcal{R}}$ has more predictive power than $\boldsymbol{\mathcal{R}}'$ iff $H[\overrightarrow{Y}^{L} | \mathcal{R}, \overrightarrow{X}^{L-1}] < H[\overrightarrow{Y}^{L} | \mathcal{R}', \overrightarrow{X}^{L-1}]$*

[1] Conditioning on the output history makes a difference iff the transducer has memory *and* internal stochasticity.

I include $\overrightarrow{X}^{L-1}$ in the conditioning variables because I want to attend only to how well the effective states capture the *internal* structure of the transducer, and the relation it imposes between inputs and outputs, not how well they do that *and* predict the future of the input series.

Lemma 14 (The Old Country Lemma for Transducers) *For all* $\mathcal{R}$ *and all* L, $H[\overrightarrow{Y}^{L}|\mathcal{R},\overrightarrow{X}^{L-1}] \geq H[\overrightarrow{Y}^{L}|(\overleftarrow{X},\overleftarrow{Y}),\overrightarrow{X}^{L-1}]$.

Proof. $(\mathcal{R},\overrightarrow{X}^{L-1}) = (\eta((\overleftarrow{X},\overleftarrow{Y})),\overrightarrow{X}^{L-1})$, i.e., the former is a function of the latter. Apply Eq. A.25 and the lemma follows.

Definition 23 (Prescience) *An effective state class* $\widehat{\mathcal{R}}$ *is prescient iff*

$$H[\overrightarrow{Y}^{L}|\widehat{\mathcal{R}},\overrightarrow{X}^{L-1}] \quad = \quad H[\overrightarrow{Y}^{L}|(\overleftarrow{X},\overleftarrow{Y}),\overrightarrow{X}^{L-1}]$$

for all L.

Lemma 15 (Prescience, Sufficiency, and Conditional Independence) *If an effective state class* $\widehat{\mathcal{R}}$ *is prescient, then it is a sufficient statistic for predicting the next output from the joint history, and it makes the next output conditionally independent of the joint history.*

Proof. Prescience $\Rightarrow$ sufficiency: Since $H[\overrightarrow{Y}^{L}|\widehat{\mathcal{R}},\overrightarrow{X}^{L-1}] = H[\overrightarrow{Y}^{L}|(\overleftarrow{X},\overleftarrow{Y}),\overrightarrow{X}^{L-1}]$, it follows (setting $L = 1$) that $H[\overrightarrow{Y}^{1}|\widehat{\mathcal{R}}] = H[\overrightarrow{Y}^{1}|(\overleftarrow{X},\overleftarrow{Y})]$. Therefore $I[\overrightarrow{Y}^{1};\widehat{\mathcal{R}}] = I[\overrightarrow{Y}^{1};(\overleftarrow{X},\overleftarrow{Y})]$, and by Proposition 6, $\widehat{\mathcal{R}}$ is a sufficient statistic. Prescience $\Rightarrow$ conditional independence: directly from Lemma 37.

Remark. The argument in the proof can be reversed to show that if an effective state class is a sufficient statistic, it attains the lower bound of Lemma 14 when $L = 1$. However, this is not enough to give us prescience.

7.3.1 Determinism

Definition 24 (Determinism for State Classes) *A class of effective states* $\mathcal{R}$ *is* deterministic *if the current state and the next input and next output fix the next state. That is, there exists a function* g *such that* $\eta((\overleftarrow{x},\overleftarrow{y}) \oplus (a,b)) = g(\eta((\overleftarrow{x},\overleftarrow{y})),(a,b))$, $\forall (a,b) \in \mathcal{A} \times \mathcal{B}$.

Remark. This definition of determinism implies that transitions from one state to another happen after seeing both a new input and a new output. In the theory of finite state transducers (Booth 1967), this is a "Mealy machine", as opposed to a "Moore machine," which has a single output for each state, and makes transitions only on inputs. Translation between the two representations is always possible for non-stochastic transducers, but is sometimes very awkward. Formulating a "Moore" version of the computational mechanics of transducers is an interesting exercise, but outside the scope of this book.

Lemma 16 (Equivalent Determination Lemma) $\mathcal{R}$ *is deterministic if and only if*

$$\begin{aligned} \forall (\overleftarrow{x}_1,\overleftarrow{y}_1),(\overleftarrow{x}_2,\overleftarrow{y}_2) &\in \overleftarrow{\mathbf{X}} \times \overleftarrow{\mathbf{Y}} \text{ and} \\ \forall (a,b) &\in \mathcal{A} \times \mathcal{B} , \end{aligned}$$

$$(\overleftarrow{x}_1,\overleftarrow{y}_1) \sim_{\eta} (\overleftarrow{x}_2,\overleftarrow{y}_2) \quad \Rightarrow \quad (\overleftarrow{x}_1,\overleftarrow{y}_1) \oplus (a,b) \sim_{\eta} (\overleftarrow{x}_2,\overleftarrow{y}_2) \oplus (a,b) \ .$$

Proof. If the statement about equivalence is true, then obviously the function invoked by Definition 24 exists and the states are deterministic. I therefore only have to prove that the existence of the function implies that the equivalence. Suppose it did not. Then there would exist at least one triple $(\overleftarrow{x}_1, \overleftarrow{y}_1), (\overleftarrow{x}_2, \overleftarrow{y}_2), (a, b)$ such that

$$\begin{aligned} \eta((\overleftarrow{x}_1, \overleftarrow{y}_1)) &= \eta((\overleftarrow{x}_2, \overleftarrow{y}_2)) \text{ and} && (7.4)\\ \eta((\overleftarrow{x}_1, \overleftarrow{y}_1) \oplus (a, b)) &\neq \eta((\overleftarrow{x}_2, \overleftarrow{y}_2) \oplus (a, b)) && (7.5) \end{aligned}$$

By hypothesis,

$$\begin{aligned} \eta((\overleftarrow{x}_1, \overleftarrow{y}_1) \oplus (a, b)) &= g(\eta((\overleftarrow{x}_1, \overleftarrow{y}_1)), (a, b)) && (7.6)\\ \eta((\overleftarrow{x}_2, \overleftarrow{y}_2) \oplus (a, b)) &= g(\eta((\overleftarrow{x}_2, \overleftarrow{y}_2)), (a, b)) \ , && (7.7) \end{aligned}$$

so

$$g(\eta((\overleftarrow{x}_1, \overleftarrow{y}_1)), (a, b)) \neq g(\eta((\overleftarrow{x}_2, \overleftarrow{y}_2)), (a, b)) \ . \tag{7.8}$$

But, substituting equals for equals, that would mean

$$g(\eta((\overleftarrow{x}_1, \overleftarrow{y}_1)), (a, b)) \neq g(\eta((\overleftarrow{x}_1, \overleftarrow{y}_1)), (a, b)) \ , \tag{7.9}$$

which is absurd. Therefore there is no such triple, and the promised implication holds. QED.

Lemma 17 (Sufficiency and Determinism Imply Prescience) *If $\mathcal{R}$ is deterministic and a sufficient statistic for predicting $\overrightarrow{Y}^1$ from $(\overleftarrow{X}, \overleftarrow{Y})$, then $\boldsymbol{\mathcal{R}}$ is prescient.*

Proof. By Proposition 6,

$$\begin{aligned} I[\overrightarrow{Y}^1; (\overleftarrow{X}, \overleftarrow{Y})] &= I[\overrightarrow{Y}^1; \mathcal{R}] && (7.10)\\ H[\overrightarrow{Y}^1] - H[\overrightarrow{Y}^1 | (\overleftarrow{X}, \overleftarrow{Y})] &= H[\overrightarrow{Y}^1] - H[\overrightarrow{Y}^1 | \mathcal{R}] && (7.11)\\ H[\overrightarrow{Y}^1 | (\overleftarrow{X}, \overleftarrow{Y})] &= H[\overrightarrow{Y}^1 | \mathcal{R}] \ . && (7.12) \end{aligned}$$

Now, let us consider $H[\overrightarrow{Y}^L | \mathcal{S}, \overrightarrow{X}^{L-1}]$. Write $\mathcal{R}_1, \mathcal{R}_2$, etc., for the present, next, etc., effective states. Decompose the conditional entropy of the future outputs as follows, using the chain rule for entropy (Eq. A.17).

$$\begin{aligned} H[\overrightarrow{Y}^L | \mathcal{R}, \overrightarrow{X}^{L-1}] &= \sum_{j=1}^{L} H[\overrightarrow{Y}_j | \mathcal{R}, \overrightarrow{X}^{j-1}, \overrightarrow{Y}^{j-1}] && (7.13)\\ &= \sum_{j=1}^{L} H[\overrightarrow{Y}_j | \mathcal{R}_j] && (7.14)\\ &= \sum_{j=1}^{L} H[\overrightarrow{Y}_j | (\overleftarrow{X}, \overleftarrow{Y})_j] && (7.15)\\ &= \sum_{j=1}^{L} H[\overrightarrow{Y}_j | (\overleftarrow{X}, \overleftarrow{Y}), \overrightarrow{X}^{j-1}, \overrightarrow{Y}^{j-1}] && (7.16)\\ &= H[\overrightarrow{Y}^L | (\overleftarrow{X}, \overleftarrow{Y}), \overrightarrow{X}^L] && (7.17) \end{aligned}$$

Eq. 7.14 comes from the determinism of the effective states. The last line uses the chain rule again. QED.

7.4 Causal States

Definition 25 (Transducer Causal States) *The causal state are the range of the function*

$$\begin{aligned}
&\epsilon((\overleftarrow{x}, \overleftarrow{y})) \\
&= \Big\{ (\overleftarrow{x}', \overleftarrow{y}') | \forall (a,b) \in \mathcal{A} \times \mathcal{B}, \forall b' \in \mathcal{B} \\
&\quad \mathrm{P}(\overrightarrow{Y}^1 = b | (\overleftarrow{X}, \overleftarrow{Y}) = (\overleftarrow{x}, \overleftarrow{y})) = \mathrm{P}(\overrightarrow{Y}^1 = b | (\overleftarrow{X}, \overleftarrow{Y}) = (\overleftarrow{x}', \overleftarrow{y}')) \\
&\quad \text{and } \mathrm{P}(\overrightarrow{Y}^1 = b' | (\overleftarrow{X}, \overleftarrow{Y}) = (\overleftarrow{x}, \overleftarrow{y}) \oplus (a,b)) = \mathrm{P}(\overrightarrow{Y}^1 = b' | (\overleftarrow{X}, \overleftarrow{Y}) = (\overleftarrow{x}', \overleftarrow{y}') \oplus (a,b)) \Big\} .
\end{aligned} \tag{7.18}$$

Remark. The second clause of the definition of ϵ ensures that the causal states are deterministic, which (as will be seen) is important for much of what follows. It would be very interesting to know necessary and sufficient conditions for the second clause to be redundant. An obvious sufficient condition is that the transducer be memoryless.

Theorem 12 (Markov Property for Transducer Causal States) *Given the causal state at time t, and the values of the input and output series from time t to time $t+L$, the causal state at $t+L$, $\mathcal{S}_L$ is independent of the values of the input and output processes, and of the causal state, at times before t, for all positive L.*

$$\forall L \in \mathbb{Z}^+,\ \mathcal{S}_L \perp\!\!\!\perp (\overleftarrow{X}, \overleftarrow{Y}) | \mathcal{S}, \overrightarrow{Y}^L, \overrightarrow{X}^L \tag{7.19}$$

Proof. Invoke the determinism of the causal states L times to see that $\mathcal{S}_L$ is a function of $\mathcal{S}, \overrightarrow{Y}^L$ and $\overrightarrow{X}^L$. Hence it is trivially conditionally independent of everything else. QED.

The Markov property implies that the causal structure of the transduction process takes a particular, repetitive form, illustrated in Figure 7.1.

Lemma 18 (Sufficiency of the Transducer Causal States) *The causal states are sufficient statistics for predicting the next output from the joint history.*

Proof. It is obvious from Definition 25 that $\mathrm{P}(\overrightarrow{Y}^1 = b | \mathcal{S} = \epsilon((\overleftarrow{x}, \overleftarrow{y}))) = \mathrm{P}(\overrightarrow{Y}^1 = b | (\overleftarrow{X}, \overleftarrow{Y}) = (\overleftarrow{x}, \overleftarrow{y}))$. Hence, by Definition 66, they are sufficient. QED.

Theorem 13 (Prescience of Causal States (Transducers)) *The causal states are prescient.*

Proof. From Lemma 18, the causal states are sufficient for the next output. Also, from their definition, they are deterministic. Hence, by Lemma 17, they are prescient. QED.

Lemma 19 (Determined Refinement Lemma) *If $\widehat{\boldsymbol{\mathcal{R}}}$ is deterministic class of prescient states, then it is a refinement a.e. of $\boldsymbol{\mathcal{S}}$.*

Proof. Because $\widehat{\boldsymbol{\mathcal{R}}}$ is prescient, $H[\overrightarrow{Y}^1 | \widehat{\mathcal{R}}]$ is as small as possible. Hence each cell of the partition must be at least weakly homogeneous for $\overrightarrow{Y}^1$, otherwise (by the usual Refinement Lemma argument) it would mix distributions for $\overrightarrow{Y}^1$, raising its conditional entropy. Hence $\eta((\overleftarrow{x}_1, \overleftarrow{y}_1)) = \eta((\overleftarrow{x}_2, \overleftarrow{y}_2))$ implies that $\mathrm{P}(\overrightarrow{Y}^1 | (\overleftarrow{X}, \overleftarrow{Y}) = (\overleftarrow{x}_1, \overleftarrow{y}_1)) = \mathrm{P}(\overrightarrow{Y}^1 | (\overleftarrow{X}, \overleftarrow{Y}) = (\overleftarrow{x}_2, \overleftarrow{y}_2))$ with probability one. Because $\widehat{\boldsymbol{\mathcal{R}}}$ is (ex hypothesi) deterministic, the Equivalent Determination Lemma (16) applies. Thus, if $\eta((\overleftarrow{x}_1, \overleftarrow{y}_1)) = \eta((\overleftarrow{x}_2, \overleftarrow{y}_2))$, then $\eta((\overleftarrow{x}_1, \overleftarrow{y}_1) \oplus (a,b)) = \eta((\overleftarrow{x}_2, \overleftarrow{y}_2) \oplus (a,b))$ for all (a,b). But the conjunction of those two conditions propositions *is* the proposition that $\epsilon((\overleftarrow{x}_1, \overleftarrow{y}_1)) = \epsilon((\overleftarrow{x}_2, \overleftarrow{y}_2))$. Hence, under the hypotheses of the lemma, if $\eta((\overleftarrow{x}_1, \overleftarrow{y}_1)) = \eta((\overleftarrow{x}_2, \overleftarrow{y}_2))$, then $\epsilon((\overleftarrow{x}_1, \overleftarrow{y}_1)) = \epsilon((\overleftarrow{x}_2, \overleftarrow{y}_2))$ almost always. Hence $\widehat{\boldsymbol{\mathcal{R}}}$ is a refinement of $\boldsymbol{\mathcal{S}}$ almost everywhere. QED.

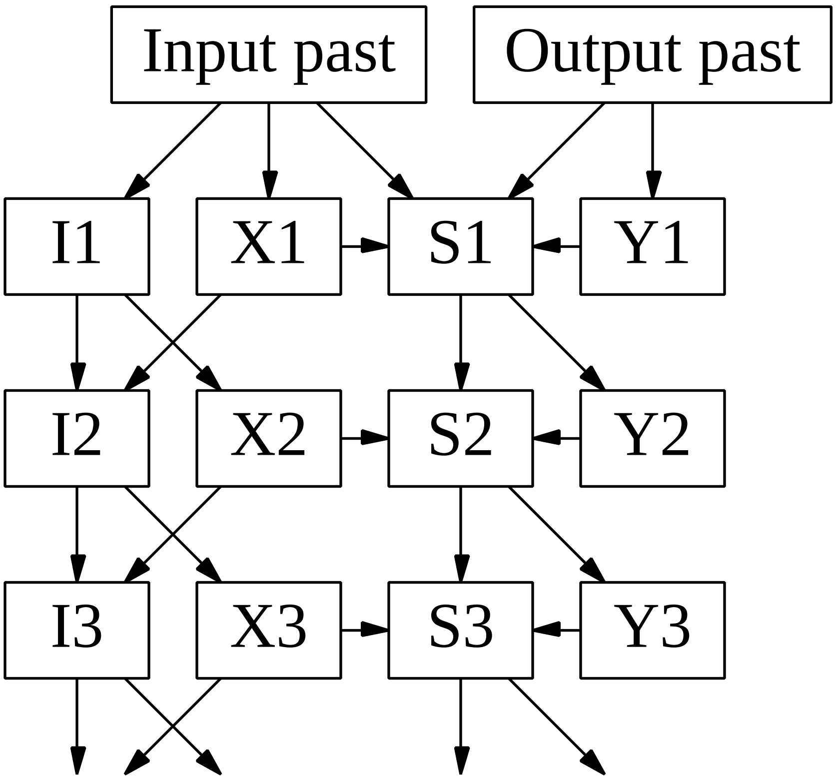

Figure 7.1: Diagram of causal influences for a portion of the time evolution of a transducer with memory but no feedback. The input process may not be Markovian, so I include the (autonomous) causal states of the input process. The absence of feedback shows up as a lack of causal paths from the output variables to future inputs.

Theorem 14 (Transducer Causal States Are Minimal) *If* $\widehat{\boldsymbol{\mathcal{R}}}$ *is deterministic class of prescient states, then* $C_\mu(\widehat{\boldsymbol{\mathcal{R}}}) \geq C_\mu(\boldsymbol{\mathcal{S}})$.

Proof. Entirely parallel to the previous minimality theorems, using the Determined Refinement Lemma in place of the other refinement lemmas.

Theorem 15 (Uniqueness of the Transducer Causal States) *If* $\widehat{\boldsymbol{\mathcal{R}}}$ *is a deterministic class of prescient states, and* $C_\mu(\widehat{\boldsymbol{\mathcal{R}}}) = C_\mu(\boldsymbol{\mathcal{S}})$*, then there exists an invertible function* f *such that* $\mathcal{S} = f(\widehat{\mathcal{R}})$ *almost always.*

Proof. Identical to the proof of the uniqueness theorem for time series.

Theorem 16 (The Transducer Causal States Are Minimally Stochastic) *For any prescient, deterministic rival class of states* $\widehat{\boldsymbol{\mathcal{R}}}$, $H[\widehat{\mathcal{R}}'|\widehat{\mathcal{R}}, \overrightarrow{X}^1] \geq H[\mathcal{S}'|\mathcal{S}, \overrightarrow{X}^1]$*, where* $\widehat{\mathcal{R}}'$ *is the next* $\hat{\eta}$*-state and* $\mathcal{S}'$ *is the next causal state.*

Proof. Begin by considering the uncertainty in $\widehat{\mathcal{R}}'$, given $\widehat{\mathcal{R}}$ and $(\overrightarrow{X}^1, \overrightarrow{Y}^1)$, remembering that $\widehat{\boldsymbol{\mathcal{R}}}$ is deterministic.

$$\begin{aligned}
H[\widehat{\mathcal{R}}'|\widehat{\mathcal{R}}, (\overrightarrow{X}^1, \overrightarrow{Y}^1)] &= 0 & (7.20)\\
&= H[\widehat{\mathcal{R}}', \overrightarrow{Y}^1|\widehat{\mathcal{R}}, \overrightarrow{X}^1] - H[\overrightarrow{Y}^1|\widehat{\mathcal{R}}, \overrightarrow{X}^1] & (7.21)\\
&= H[\overrightarrow{Y}^1|\widehat{\mathcal{R}}', \widehat{\mathcal{R}}, \overrightarrow{X}^1] + H[\widehat{\mathcal{R}}'|\widehat{\mathcal{R}}, \overrightarrow{X}^1] - H[\overrightarrow{Y}^1|\widehat{\mathcal{R}}, \overrightarrow{X}^1] & (7.22)\\
H[\widehat{\mathcal{R}}'|\widehat{\mathcal{R}}, \overrightarrow{X}^1] &= H[\overrightarrow{Y}^1|\widehat{\mathcal{R}}', \widehat{\mathcal{R}}, \overrightarrow{X}^1] - H[\overrightarrow{Y}^1|\widehat{\mathcal{R}}, \overrightarrow{X}^1] & (7.23)
\end{aligned}$$

This applies to the causal states, too:

$$H[\mathcal{S}'|\mathcal{S}, \overrightarrow{X}^1] = H[\overrightarrow{Y}^1|\mathcal{S}, \overrightarrow{X}^1] - H[\overrightarrow{Y}^1|\mathcal{S}', \mathcal{S}, \overrightarrow{X}^1] \,. \qquad (7.24)$$

Since $\overrightarrow{X}^1$ happens after $\overrightarrow{Y}^1$, the latter can depend on the former only if they are both dependent on a third variable. The only such variable available is $\overleftarrow{X}$. But conditioning on $\widehat{\mathcal{R}}$ makes $\overrightarrow{Y}^1$ independent of $\overleftarrow{X}$, so $H[\overrightarrow{Y}^1|\widehat{\mathcal{R}}, \overrightarrow{X}^1] = H[\overrightarrow{Y}^1|\widehat{\mathcal{R}}]$. And of course $H[\overrightarrow{Y}^1|\widehat{\mathcal{R}}] = H[\overrightarrow{Y}^1|\mathcal{S}]$. Bearing this in mind, subtract Eq. 7.24 from Eq. 7.23.

$$\begin{aligned}
&H[\widehat{\mathcal{R}}'|\widehat{\mathcal{R}}, \overrightarrow{X}^1] - H[\mathcal{S}'|\mathcal{S}, \overrightarrow{X}^1]\\
&= H[\overrightarrow{Y}^1|\widehat{\mathcal{R}}', \widehat{\mathcal{R}}, \overrightarrow{X}^1] - H[\overrightarrow{Y}^1|\widehat{\mathcal{R}}, \overrightarrow{X}^1] - H[\overrightarrow{Y}^1|\mathcal{S}, \overrightarrow{X}^1] + H[\overrightarrow{Y}^1|\mathcal{S}', \mathcal{S}, \overrightarrow{X}^1] & (7.25)\\
&= H[\overrightarrow{Y}^1|\mathcal{S}', \mathcal{S}, \overrightarrow{X}^1] - H[\overrightarrow{Y}^1|\widehat{\mathcal{R}}', \widehat{\mathcal{R}}, \overrightarrow{X}^1] & (7.26)
\end{aligned}$$

By the Determined Refinement Lemma, $\mathcal{S}$ and $\mathcal{S}'$ are functions of $\widehat{\mathcal{R}}$ and $\widehat{\mathcal{R}}'$, respectively. Hence $\mathcal{S}', \mathcal{S}, \overrightarrow{X}^1$ is a function of $\widehat{\mathcal{R}}', \widehat{\mathcal{R}}, \overrightarrow{X}^1$, and by Eq. A.25,

$$\begin{aligned}
H[\overrightarrow{Y}^1|\mathcal{S}', \mathcal{S}, \overrightarrow{X}^1] &\geq H[\overrightarrow{Y}^1|\widehat{\mathcal{R}}', \widehat{\mathcal{R}}, \overrightarrow{X}^1] & (7.27)\\
H[\widehat{\mathcal{R}}'|\widehat{\mathcal{R}}, \overrightarrow{X}^1] - H[\mathcal{S}'|\mathcal{S}, \overrightarrow{X}^1] &\geq 0 & (7.28)\\
H[\widehat{\mathcal{R}}'|\widehat{\mathcal{R}}, \overrightarrow{X}^1] &\geq H[\mathcal{S}'|\mathcal{S}, \overrightarrow{X}^1] \,. & (7.29)
\end{aligned}$$

QED.

Remark. In the case of time series, we looked at $H[\widehat{\mathcal{R}}'|\widehat{\mathcal{R}}]$ to gauge the internal stochasticity of a class of effective states. Here, however, that quantity $= H[\widehat{\mathcal{R}}'|\widehat{\mathcal{R}}, \overrightarrow{X}^1] + H[\overrightarrow{X}^1|\widehat{\mathcal{R}}] - H[\overrightarrow{X}^1|\widehat{\mathcal{R}}', \widehat{\mathcal{R}}]$. That is, it involves the degree of randomness in the input process, as well as whatever randomness is in the internal dynamics of the transducer. But it would be rather much to expect that the states which predict the behavior of the transducer nicely are also good predictors of the behavior of the input process.[2]

7.5 Transduction with Feedback

I assumed above that the output has no influence on the input. This is often true, and it's the classic transducer problem, but there is no logical necessity for this to be so. If the $\overleftarrow{Y}$ does influence $\overrightarrow{X}$, there's feedback (and the labels "input" and "output" are dubious). $\boldsymbol{\mathcal{S}}$ remains the unique, optimal, minimal class of states for predicting the future of the output on the basis of the joint history. But we can go through an entirely parallel construction for predicting the input on the basis of the joint history; call the resulting class of states $\mathcal{F}$. The causal structure which results is that of Figure 7.2.

Transducers without feedback are simply a special case of this situation, represented in the diagram by erasing the arrows from Y_i to $\mathcal{F}_i$.

Now, if we consider the input and the output jointly, we have simply another discrete time-series, as in Chapter 4, so the theory developed there applies. That is, we can construct a class of causal states (call it $\mathcal{J}$) for the joint input-output process. This raises the question of how $\mathcal{J}$ is related to $\boldsymbol{\mathcal{S}}$ and $\mathcal{F}$, bearing in mind that all three are partitions on $\overleftarrow{\mathbf{X}} \times \overleftarrow{\mathbf{Y}}$.

We know that

$$\overrightarrow{Y}^1 \perp\!\!\!\perp (\overleftarrow{X}, \overleftarrow{Y}) | \mathcal{S} \tag{7.30}$$

$$\overrightarrow{X}^1 \perp\!\!\!\perp (\overleftarrow{X}, \overleftarrow{Y}) | \mathcal{F} \ . \tag{7.31}$$

Since $\mathcal{S}$ and $\mathcal{F}$ are both functions of $(\overleftarrow{X}, \overleftarrow{Y})$, we have (Eq. A.38)

$$\overrightarrow{Y}^1 \perp\!\!\!\perp (\overleftarrow{X}, \overleftarrow{Y}), \mathcal{F} | \mathcal{S} \tag{7.32}$$

$$\overrightarrow{X}^1 \perp\!\!\!\perp (\overleftarrow{X}, \overleftarrow{Y}), \mathcal{S} | \mathcal{F} \ . \tag{7.33}$$

Applying Eq. A.34,

$$\overrightarrow{Y}^1 \perp\!\!\!\perp (\overleftarrow{X}, \overleftarrow{Y}) | \mathcal{S}, \mathcal{F} \tag{7.34}$$

$$\overrightarrow{X}^1 \perp\!\!\!\perp (\overleftarrow{X}, \overleftarrow{Y}) | \mathcal{S}, \mathcal{F} \ . \tag{7.35}$$

Furthermore, it's certainly true that

$$\overrightarrow{Y}^1 \perp\!\!\!\perp (\overleftarrow{X}, \overleftarrow{Y}), \overrightarrow{X}^1 | \mathcal{S}, \mathcal{F} \tag{7.36}$$

$$\overrightarrow{X}^1 \perp\!\!\!\perp (\overleftarrow{X}, \overleftarrow{Y}), \overrightarrow{Y}^1 | \mathcal{S}, \mathcal{F} \tag{7.37}$$

since $\overrightarrow{X}^1$ has no direct causal effect on $\overrightarrow{Y}^1$, and any probabilistic dependency there may be is screened off by $\mathcal{S}$ and $\mathcal{F}$ together. Now Eq. A.33 tells us that

$$(\overrightarrow{X}^1, \overrightarrow{Y}^1) \perp\!\!\!\perp (\overleftarrow{X}, \overleftarrow{Y}) | \mathcal{S}, \mathcal{F} \ , \tag{7.38}$$

which is to say, the combination of $\mathcal{S}$ and $\mathcal{F}$ is a sufficient statistic for joint futures of length 1. Since it is also deterministic, by Theorem 6, it is a prescient class of states. But then by the Refinement Lemma for Time Series (Lemma 12), there is a mapping from $\mathcal{S}, \mathcal{F}$ to $\mathcal{J}$.

[2]Note that $H[\overrightarrow{X}^1|\widehat{\mathcal{R}}] - H[\overrightarrow{X}^1|\widehat{\mathcal{R}}', \widehat{\mathcal{R}}] = I[\overrightarrow{X}^1; \widehat{\mathcal{R}}'|\widehat{\mathcal{R}}]$, the mutual information between $\overrightarrow{X}^1$ and $\widehat{\mathcal{R}}'$ conditional on $\widehat{\mathcal{R}}$.

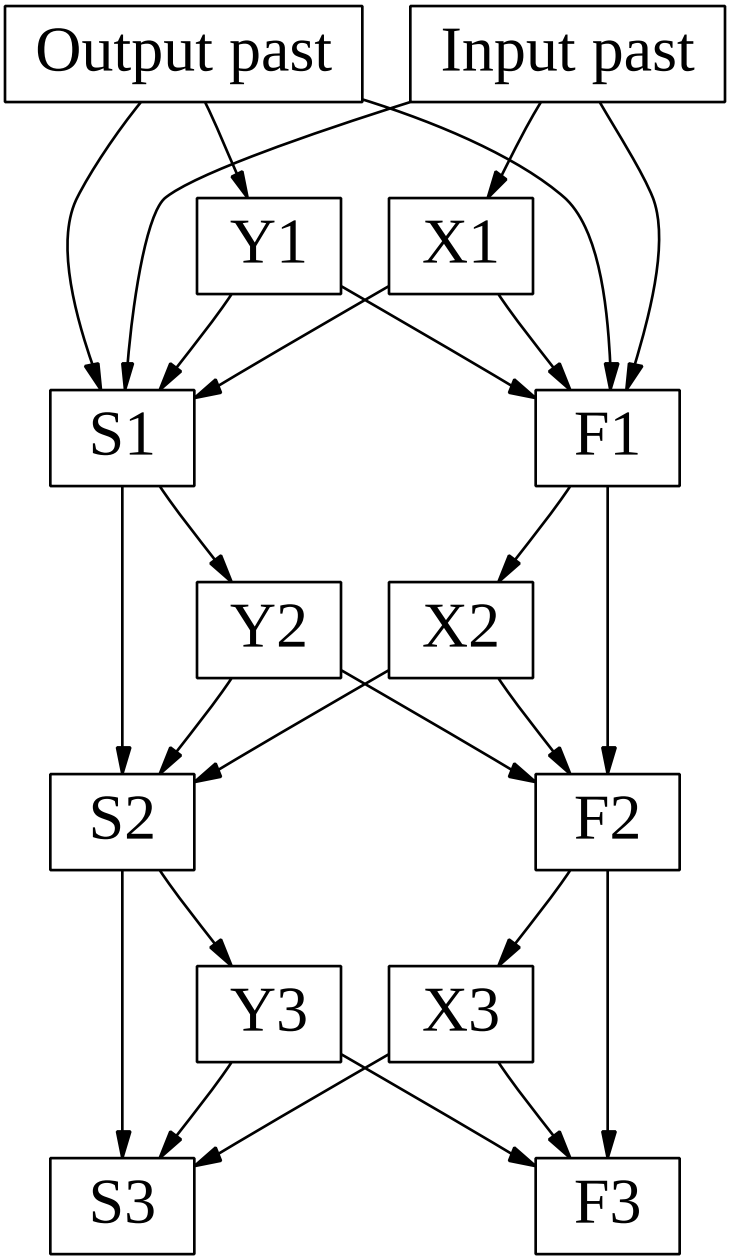

Figure 7.2: Diagram of causal effects for a transducer with memory and feedback. Observe that all paths from time i to time $i+1$ run through $\mathcal{S}_i$ or $\mathcal{F}_i$.

7.5.1 An Aside: The Elimination of Dialectics in Favor of Mechanics

The notion of a dialectical relationship between two entities is a famously murky one (for an unusually lucid historical account, see Kolakowski (1978, vol. I)). The only reasonably clear account I have found is Phil Agre's.

> A dialectical relationship between two entities, called *moments*, has three properties: (1) the moments are engaged in a time-extended interaction, (2) they influence each other through this interaction, and (3) this influence has grown sufficiently large that it becomes impossible to define either moment except in terms of its relationship to the other. The moments are typically, though not necessarily, thought of as being in conflict with one another; the interaction between them and their mutual influence are products of this conflict. If this seems overly metaphysical ... think of it in the following way. Make a list of the states or properties of that the two entities possess at a given moment. Then take each of the lists in isolation from the other and ask whether it is possible to find any rhyme or reason for that set of states or properties, except by reference to the interaction and cumulative influence that the entity has gone through. If not, i.e., if the only reasonable explanation for each entity's list makes reference to its past history of interaction with the other entity, then the relationship between the two entities is dialectical in nature. (Agre 1997, pp. 318–319)

Put in the language of computational mechanics, this says that $I[\overrightarrow{Y}^L; \overleftarrow{Y}]$ and $I[\overrightarrow{X}^L; \overleftarrow{X}]$ are negligible, while $I[\overrightarrow{Y}^L; (\overleftarrow{X}, \overleftarrow{Y})]$ and $I[\overrightarrow{Y}^L; (\overleftarrow{X}, \overleftarrow{Y})]$ are substantial. There is nothing implausible about that, and in fact it's just when we're likely to think of the processes as showing feedback. We may, of course, construct the joint causal state for the dialectical pair in the usual way. But now something amusing happens.

Suppose that the moments of the dialectical relationship are ordinary pieces of matter. (An insistence on this point is very much a part of what was historically the most influential school of dialectical thinking.) That being the case, they should obey ordinary statistical mechanics. Then, applying the techniques of Section 4.5, we can go from the causal partition of the joint histories, to a partition of the joint phase space of the two systems. That partition has the following properties:

1. The partition corresponds to a single observable macroscopic variable.
2. The dynamics of that variable are Markovian.
3. The current value of the variable is a sufficient statistic for the entire future of both of the moments.

The evolution of this macrovariable shows no signs of history or of interaction.

The upshot is that, even when it's most reasonable to talk about dialectical relationships, we can always replace the dialectical representation with a purely (statistical) mechanical one, without any loss of information.

7.6 Kindred Approaches

As I said at the beginning of the chapter, the ϵ-transducer is analogous to what computer scientists call a "finite state transducer", (Definition 64). For several decades at least, however, most treatments of these objects have been entirely nonstochastic. (The last detailed treatment of stochastic FSTs I know of is that of Booth (1967).) So far as I have been able to learn, nothing like this construction of deterministic states for stochastic transducers exists in the FST literature. And, again, what I have done in this chapter does *not* assume that only a finite number of states are needed, or that the memory of the transducer extends only a finite distance into the past.

There *has* been a burst of work on stochastic models of discrete transduction in the last few years, driven by the demands of bioinformatics (Singer 1997; Apostolico and Bejerano 2000; Eskin, Grundy and Singer 2000). Many of these models even have very nice determinism and Markov properties. The ϵ-transducer approach formally incorporates them all, and has the extra advantages of not having to *postulate* the Markov properties, nor of having to *guess* the internal architecture of the states, the one being proved and the other inferred.

Perhaps the best-known theory of nonlinear transducers is that of Norbert Wiener[3]. This expresses a nonlinear, deterministic relationship between continuous input and output signals by means of a "power series" of functionals. The n^{th} term in the series is the convolution of a kernel with n copies of the input. The kernels are chosen, with extreme cleverness, so that they can be calculated from the cross-correlation of the input and the output, and moreover so that, when the input is white noise, all the kernels are statistically independent. This theory has actually been applied to biological systems with considerable success (Rieke, Warland, de Ruyter van Steveninck and Bialek 1997), and can be expanded to accommodate stochastic transducers (Victor and Johannesma 1986).

While Wiener's theory is very elegant, the fact that it uses a series expansion has its own drawbacks. The calculation of the higher-order kernels from data, while certainly possible, is not easy, and most applications truncate the series at the first or at most the second term. There is, however, no reason to think that the series converges *quickly*, that the first two terms are a good approximation to the whole. In fact, it would be nice if we didn't have to use any sort of series at all, and simply calculate all effects, linear and nonlinear, at once. The ϵ-transducer does this, much as the ϵ-machine for a time series does.

In information theory, one of our transducers is a *channel with memory*. This is, in a way, unfortunate, because the vastly overwhelming majority of information theory is about memoryless channels, and what little there is on channels with memory has concentrated on the channel capacity, the rate at which a signal can be transmitted without error (Verdu 1994, sec. 3). In all modesty, the theory in this chapter may be of some use to people working on channels with memory!

7.7 Reconstruction

The state-splitting algorithm of Chapter 5 can easily be adapted to deal with transducers without feedback, simply by considering the joint history, and splitting joint histories when they produce significantly different distributions for the next output. The reconstruction of the feedback state would go in the same way. The reliability analysis proceeds on exactly the same lines as for time series, so I won't redo it here.

[3]Wiener (1958), the original source, is rewarding but mathematically demanding and full of misprints. A much easier introduction is to be had from Rieke, Warland, de Ruyter van Steveninck and Bialek (1997, App. A3), while Schetzen (1989) covers developments up to about 1980.

Chapter 8

A Very Brief Introduction to Cellular Automata

The chess-board is the world; the pieces are the phenomena of the universe; the rules of the game are what we call the laws of Nature.
— T. H. Huxley

8.1 An Extremely Informal Description

Take a board, and divide it up into squares, like a chess-board or checker-board. These are the cells. Each cell has one of a finite number of distinct colors — red and black, say, or (to be patriotic) red, white and blue. (We don't allow continuous shading, and every cell has just one color.) Now we come to the "automaton" part. Sitting somewhere to one side of the board is a clock, and every time the clock ticks the colors of the cells change. Each cell looks at the colors of the nearby cells, and its own color, and then applies a definite rule, the *transition rule*, specified in advance, to decide its color in the next clock-tick; and all the cells change at the same time. (The rule can sometimes tell the cell to stay the same.) Each cell is a sort of very stupid computer — in the jargon, a finite-state automaton — and so the whole board is called a cellular automaton, or CA. To run it, you color the cells in your favorite pattern, start the clock, and stand back.

Now that (I hope) you have a concrete picture, I can get a bit more technical, and more abstract. The cells don't have to be colored, of course; all that's important is that each cell is in one of a finite number of states at any given time. By custom they're written as the integers, starting from 0, but any finite alphabet will do. Usually the number of states is small, under ten, but in principle any finite number is OK. What counts as the "nearby cells", the neighborhood, varies from automaton to automaton; sometimes just the four cells on the principle directions (the *von Neumann neighborhood*), sometimes the corner cells (the *Moore neighborhood*), sometimes a block or diamond of larger size; in principle any arbitrary shape. You don't need to stick to a chess-board; you can use any pattern of cells which will fill the plane (or "tessellate" it; an old name for cellular automata is "tessellation structures"). And you don't have to stick to the plane; any integer number of dimensions is allowed. You do need to stick to discrete time, to clock-ticks; but CAs have cousins in which time is continuous. There are various tricks for handling the edges of the board; the most common, both of which have "all the advantages of theft over honest toil" are to have the edges "wrap around" to touch each other, and to assume an infinite board.

One important use of CAs is to mimic bits and pieces of the real world. CAs are fully discretized classical field theories, so they're good at the same things classical field theories are, provided continuity isn't so important, and much better at things like messy boundary conditions (Manneville, Boccara, Vichniac and Bidaux 1990; Chopard and Droz 1998). Their domain of application includes fluid flow (Rothman and Zaleski 1997), excitable media (Winfree 1987), many other sorts of pattern formation (Cross and Hohenberg

1993; D'Souza and Margolus 1999), ecosystems (Levin, Powell and Steele 1993; Tilman and Kareiva 1997), highway traffic, even the development of cities (White and Engelen 1993; Clarke, Gaydos and Hoppen 1996; Manrubia, Zanette and Sole 1999; cf. Anas, Arnott and Small 1998). There's a modest industry of seeing which types of CAs have various properties of interest to theoretical physicists — time-reversibility, various sorts of symmetry, etc. (Gutowitz 1991; Smith 1994). There's even a current of thought pushing the idea that CAs capture something really fundamental about physics, that they are more physical than the differential equations we have come to know and love these last three hundred years (Toffoli 1984; Margolus 1999). I can't say I buy this myself, but some of its believers are very smart indeed, and anyway it makes for excellent science fiction (Egan 1994).

8.1.1 General References on CAs

The best non-technical introduction to cellular automata is the book by Poundstone (1984), which describes in detail the most famous CA of all, Conway's Game of Life (Berlekamp, Conway and Guy 1982). Flake (1998) provides a bit more math, and some fun programming projects. Burks (1970) collects foundational papers from the misty, heroic age of CA theory, before they could be readily simulated and *seen* on computers. The standard modern reference is Gutowitz (1991), but it will probably be superseded by Griffeath and Moore (forthcoming), if that ever appears.

Cellular automata were introduced by John von Neumann and Stanislaw Ulam in the 1950s to study the possibility of mechanical self-reproduction (von Neumann 1966; Burks 1970). There is no adequate study of the history of cellular automata.

8.1.2 A More Formal Description

A CA starts with a d-dimensional regular lattice Ξ of sites or cells.[1] Each cell x has a *neighborhood* $n(x)$ of other cells, definitely including those it is connected to in the lattice, but possibly including others which are connected to those; neighborhoods are connected components of the lattice containing the original cell. Every cell has the same size and shape neighborhood as every other cell, i.e., $\mathbf{T}n(x) = n(\mathbf{T}x)$, where $\mathbf{T}$ is any spatial translation operator. The standard neighborhoods consist of all cells within a certain distance r of x; r is the *rule radius.*

A *configuration* of the lattice (or of the CA) assigns to every cell a value from a finite alphabet $\mathcal{A}$ of size k. We write the value at x as s^x. The configuration in the neighborhood of x is $s^{n(x)}$. Time is discrete and goes in the subscript: s^x_t is the value of the cell x at time t. The global configuration at time t is $\mathbf{s}_t$.

The *CA rule* is a function ϕ from a neighborhood configuration to a new cell-value.[2] The CA's equation of motion is given by applying the rule to each point separately:

$$s^x_{t+1} \quad = \quad \phi(s^{n(x)}_t) \ . \tag{8.1}$$

The simultaneous application of ϕ to all cells defines the *global update rule* Φ, a mapping from $\mathcal{A}^\Xi$ into itself.

Binary ($k = 2$), $r = 1$, one-dimensional CAs are called *elementary* CAs (ECAs) (Wolfram 1983).

An ensemble operator $\mathbf{\Phi}$ can be defined (Hanson and Crutchfield 1992; Wolfram 1984a) that operates on sets of lattice configurations $\Omega_t = \{\mathbf{s}_t\}$:

$$\Omega_{t+1} = \mathbf{\Phi}\Omega_t \ , \tag{8.2}$$

such that

$$\Omega_{t+1} = \{\mathbf{s}_{t+1} : \mathbf{s}_{t+1} = \Phi(\mathbf{s}_t), \ \mathbf{s}_t \in \Omega_t\} \ . \tag{8.3}$$

[1]Sometimes $\Xi = \mathbb{Z}^d$, sometimes just a finite chunk of it.

[2]If ϕ is a random function, then we have a *stochastic cellular automaton.*

8.2 CA as Dynamical Systems

CA are dynamical systems with discrete time, i.e., maps. They are a little peculiar, owing to the very discrete nature of the space in they live, but many of the familiar concepts of dynamical systems theory apply just fine.

Definition 26 (Invariant Set, Fixed Point, Transient) *A set of global configurations* Ω *is* invariant *iff* $\mathbf{\Phi}\Omega = \Omega$*. If* Ω *consists of a single configuration* $\mathbf{s}^*$*, then* $\mathbf{s}^*$ *is a* fixed point*. If* $\mathbf{s}$ *is not part of any invariant set, then it is* transient*.*

Definition 27 (Attractor) *A set of configurations A is an* attractor *iff*

1. *A is invariant; and*
2. *there is a non-empty set of configurations U such that* $U \cap A = \emptyset$ *but* $\mathbf{\Phi} U \subseteq A$*.*

Definition 28 (Basin of Attraction) *The* basin of attraction $\mathcal{B}_A$ *of an attractor A is the largest set of configurations which are eventually mapped into A, i.e., the collection of all configurations* $\mathbf{b}$ *such that* $\mathbf{b} \notin A$ *but* $\Phi^k \mathbf{b} \in A$ *for some positive integer k.*

For explicit computations of the attractor basins of a large number of one-dimensional CAs see Wuensche and Lesser (1992). Some of the pictures are quite pretty, and make nice T-shirts.

There is one charmingly-named concept which is, so far as I know, only applied to CAs (among dynamical systems!).

Definition 29 (Garden of Eden) *A configuration that can only occur as an initial condition is a* Garden of Eden*. That is,* $\mathbf{s}$ *is a Garden of Eden iff,* $\forall \mathbf{s}'$*,* $\Phi \mathbf{s}' \neq \mathbf{s}$*.*

The existence of Gardens of Eden has important implications for the computational capacities of cellular automata, including their ability to support self-reproduction (Moore 1970).

All the concepts I've defined treat each configuration as a point in the CA's state space. CA dynamics, thus defined, does not represent spatial structure in any explicit or even comprehensibly-implicit way. There is an alternative way of treating a CA as a dynamical system which does, where the state space consists, somewhat paradoxically, not of individual configurations but of sets of configurations (Hanson and Crutchfield 1992; Hanson 1993; Crutchfield and Hanson 1993b). This alternative CA dynamics is, at is happens, the spatial version of computational mechanics.

Chapter 9

Domains and Particles: Spatial Computational Mechanics

People notice patterns when they look at CAs, though whether this says more about what CA are apt to do, or what people like to look at, is a nice question. Two very common kinds of patterns noted in CAs are *domains* — patches of space-time where everything looks the same, where some "texture" is repeated over and over — and *particles*, localized blobules which propagate across the lattice (Hanson and Crutchfield 1992). A review of the literature indicates that particles are generally felt to be about the most interesting things in CAs [1]. Part of the reason for this is that propagating blobules are observed in real physical systems, where they can be very important (Manneville 1990; Cross and Hohenberg 1993; Winfree 1980; Winfree 1987; Fedorova and Zeitlin 2000; Infeld and Rowlands 1990). Sometimes, especially in condensed matter physics, they are called "defects," but some people (and fields) prefer more P.C. names, like "coherent structures," "solitons" or "organizing centers". An analogy with Conway's Game of Life (Poundstone 1984) gives them the name "gliders," which I'll avoid[2]. Many people have long suspected that particles and domains are emergent structures. A general theoretical analysis (Hanson 1993), supplemented by a comparatively small number of explicit calculations in particular cases (Hanson and Crutchfield 1997) shows that this is true.

The burden of this chapter is to expound the theory I just mentioned, the "pure-space" computational mechanics of cellular automata of Hanson and Crutchfield. This is a method for analyzing particles and domains in one-dimensional CAs in terms of regular languages and the states of machines associated with them.[3] The theory employs causal states that specify only the spatial structure of CA configurations, obtained by treating one axis of the CA lattice as though it were the time axis. Problems about CA dynamics can be posed in the theory, and indeed it has some very powerful tools for solving such problems, but dynamics are described by the CA ensemble evolution operator Φ of the previous chapter, and further objects constructed from it, and not in terms of causal states. It also only works in one dimension, since both the automata theory and the machine-reconstruction techniques it employs apply only to well-ordered sequences of symbols. Within these limits, however, spatial computational mechanics is extremely powerful, and proved essential, for instance, in understanding how computation can be embedded in cellular automata, and even evolve Darwinianly in them (Das, Mitchell and Crutchfield 1994; Das 1996; Crutchfield, Hordijk and Mitchell 2000b; Crutchfield and Mitchell 1995; Hordijk, Mitchell and Crutchfield 1998; Hordijk 1999).

[1] A very partial list would include: Burks 1970; Berlekamp, Conway and Guy 1982; Peyrard and Kruskal 1984; Grassberger 1983; Boccara, Nasser and Roger 1991; Boccara and Roger 1991; Boccara 1993; Aizawa, Nishikawa and Kaneko 1991; Park, Steiglitz and Thurston 1986; Wolfram 1986; Wolfram 1994; Lindgren and Nordahl 1990; Crutchfield and Mitchell 1995; Yunes 1994; Eloranta 1993; Eloranta 1994; Eloranta and Nummelin 1992; Manneville, Boccara, Vichniac and Bidaux 1990; Andre, Bennett and Koza 1997; Hanson and Crutchfield 1992; Hanson 1993; Hanson and Crutchfield 1997; Eppstein ongoing.

[2] A particle, in this sense, is *not* the same as a particle in the sense of interacting particle systems (IPSs) (Griffeath 1979; Liggett 1985) or lattice gases (Rothman and Zaleski 1997). The particles of an IPS or the coherent structures that emerge in lattice gases *may* be particles this sense, however.

[3] Regular languages and automata are explained in Appendix A.4.

The next chapter constructs a fully spatio-temporal, multi-dimensional computational mechanics, containing one-dimensional spatial computational mechanics as a special case; the rest of this chapter expounds the basics, and a particular application, of the latter theory, to give an idea of what can be accomplished even without the full dynamical treatment.

9.1 Domains

Definition 30 (Domain) *(Hanson and Crutchfield 1992) A regular domain Λ of a CA Φ is a process language, representing a set of spatial lattice configurations, with the following properties:*

1. *Temporal invariance (or periodicity): Λ is mapped onto itself by the CA dynamic; i.e., $\mathbf{\Phi}^p\Lambda = \Lambda$ for some finite p. (Recall that Φ takes sets of lattice configurations into sets of configurations and that a formal language, such as Λ, is a set of configurations.)*

2. *Spatial homogeneity: The* process graph *of each temporal iterate of Λ is strongly connected. That is, there is a path between every pair of states in $M(\Phi^l\Lambda)$ for all l. (Recall that $M(\mathcal{L})$ is the minimal DFA which recognizes the language $\mathcal{L}$.)*

The set of all domains of a CA Φ is denoted $\mathbf{\Lambda} = \{\Lambda^0, \Lambda^1, \ldots, \Lambda^{m-1}\}$, where $m = |\mathbf{\Lambda}|$.

According to the first property — temporal invariance or periodicity — a particular domain Λ^i consists of p temporal phases for some $p \geq 1$; i.e., $\Lambda^i = \{\Lambda^i_0, \Lambda^i_1, \ldots, \Lambda^i_{p-1}\}$, such that $\mathbf{\Phi}^l\Lambda^i_j = \Lambda^i_{(j+l) \bmod p}$. Here p is the *temporal periodicity* of the domain Λ^i, denoted $T(\Lambda^i)$.

Each of the temporal phases Λ^i_j of a domain Λ^i is represented by a process graph $M(\Lambda^i_j)$ which, according to the second property (spatial homogeneity), is strongly connected. Each of these process graphs consists of a finite number of states. Denote the k^{th} state of the j^{th} phase of Λ^i by $\Lambda^i_{j,k}$, suppressing the $M(\cdot)$ notation for conciseness. Write the number of states in a given phase as $S(\Lambda^i_j)$.

The process graphs of all temporal phases Λ^i_j of all domains Λ^i can be connected together and transformed into a finite-state transducer, called the *domain transducer*, that reads in a spatial configuration and outputs various kinds of information about the sites. (The construction is given in, for example, Crutchfield and Hanson (1993b).) Variations on this transducer can do useful recognition tasks. For example, all transitions that were in domain Λ^i_j's process graph are assigned output symbol D, indicating that the input symbol being read is "participating" in a domain. All other transitions in the transducer indicate deviations from the sites being in a domain. They can be assigned a unique output ("wall") symbol $w \in \{W^i_j\}$ that labels the kind of domain violation that has occurred. The resulting domain transducer can now be used to *filter* CA lattice configuration, mapping all domain *regularities* to D and mapping all domain *violations* to output symbols w that indicate *domain walls* of various kinds.

I'll call a phase of a domain (spatially) *periodic* when the process graph consists of a periodic chain of states, with a single transition between successive states in the chain. That is, as one moves from state to state, an exactly periodic sequence of states is encountered and an exactly periodic sequence of symbols from Σ is encountered on the transitions. The *spatial periodicity* of a periodic phase is simply $S(\Lambda^i)$. I'll call a domain periodic when all its phases are periodic. We'll only deal with periodic domains here, for the following reason. It turns out that for such domains all of the spatial periodicities $S(\Lambda^i_j)$ at each temporal phase are equal. Thus, we can speak of *the* spatial periodicity $S(\Lambda^i)$ of a periodic domain Λ^i. This property, in turn, is central to the proof of the upper bound on the number of particle interaction products.

Lemma 20 (Periodic Phase Implies Periodic Domain) *If a domain Λ^i has a periodic phase, then the domain is periodic, and the spatial periodicities $S(\Lambda^i_j)$ of all its phases $\Lambda^i_j, j = 0, \ldots, p-1$, are equal.*

Proof. See the Appendix.

Thus, the number of states in the process graph representing a particular temporal phase Λ^i_j is the same for all $j \in \{1, \ldots, T(\Lambda^i)\}$, and it is, in fact, $S(\Lambda^i)$.

Finally, there is a larger class of *cyclic domains* whose process graphs consist of a periodic chain of states: as one moves from state to state an exactly periodic sequence of *states* is seen. Note that this class includes more than periodic domains, which are obviously cyclic. It includes domains in which between two successive states in the chain there are multiple transitions over Σ. (See Crutchfield and Hanson (1993b) for a CA exhibiting two such cyclic domains.) Based on our experience we conjecture that Lemma 20 also holds for cyclic domains. If this is so, most of the following results, and in particular the upper bound theorem, would hold for this larger class.

Conjecture 1 (Spatial Periodicities of Cyclic Domains) *For any cyclic domain Λ^i, the spatial periodicities $S(\Lambda^i_j)$ of all its phases $\Lambda^i_j, j = 0, \ldots, p-1$, are equal.*

9.2 Particles

When domain violations form a spatially localized (finite width), temporally periodic boundary between two adjacent domains, they are called *particles.*

Definition 31 *A particle α is a set $\{\alpha^0, \alpha^1, \ldots, \alpha^{p-1}\}$ of finite-width words α^j over Σ^*, called* wedges*, such that*

$$\mathbf{\Phi}(\Lambda\alpha^i\Lambda') = \Lambda\alpha^{(i+1) \bmod p}\Lambda' \ , \tag{9.1}$$

for some finite p and Λ and $\Lambda' \in \mathbf{\Lambda}$.

Since a particle is a bounded structure, it does not have a spatial periodicity. "Periodicity of a particle" therefore always means temporal periodicity.

Since these particles are temporally periodic, we can view the appearance of wedge α^j as the particle being in it's jth *phase.* The kth symbol in the wedge's word is denoted α^j_k. The state in which the domain transducer finds itself after reading the kth symbol α^j_k in the wedge α^j is denoted $q(\alpha^j_k)$.

Now I'll introduce an important but subtle distinction. The particle period p referred to above — the *surface periodicity* — is associated with the repetition over time of the wedge words as observed in the raw space-time behavior $\boldsymbol{s}_0, \boldsymbol{s}_1, \boldsymbol{s}_2, \ldots$. It turns out, as will become clear, that particles have an internal periodicity that may be some multiple of the surface periodicity p. The internal periodicity — the one of actual interest here — though, is the periodicity seen by the various phases of the bordering domains.

Definition 32 *A particle α's* intrinsic periodicity *$P(\alpha)$ is the periodicity of the set of transducer-state sequences generated when reading a particle's wedges. For wedge $\alpha^j = \alpha^j_0 \ldots \alpha^j_n$ the state sequence $q(\alpha^j_0) \ldots q(\alpha^j_n)$ is generated in the transducer. Denote this state sequence by $q(\alpha^j)$. $P(\alpha)$, then, is the number of iterations over which the sequence $q(\alpha^j)$ reappears.*

Remark 1. $P(\alpha)$ is an integer multiple of α's apparent periodicity.

Remark 2. A simple illustration of the need for intrinsic, as opposed to merely surface, periodicity is provided by the γ particles of ECA 54. See Figure 9.4(b) and the accompanying text in Section 9.5.1.

After one period $P(\alpha)$, a particle α will have moved a number d_α of sites in the CA lattice. This shift d_α in space after one period is called the particle's *displacement.* d_α is negative for displacements to the left and positive for displacements to the right. From the particle's periodicity $P(\alpha)$ and displacement d_α, its average velocity is simply $v_\alpha = d_\alpha / P(\alpha)$.

It doesn't matter whether you look at the wedges, or at the transducer-state labeled wedges, the velocity is the same.

The set of all particles $\alpha, \beta, \ldots$ of a CA Φ is denoted by $\mathbf{P}$.

Remark 3. We've just defined temporally periodic particles. There are particles in CAs, such as in ECA 18, which are temporally aperiodic. In this case, one replaces the periodicity condition Eq. 9.1 by one using the ensemble operator; viz.,

$$\mathbf{\Phi}^p(\Lambda\alpha\Lambda') = \Lambda\alpha\Lambda' \ . \tag{9.2}$$

9.2.1 Structural Complexity of a Particle

The preceding definitions and discussion suggest that one can think of particles as having an internal clock or, in the more general case that includes aperiodic particles, an internal state, much as the solitary-wave solutions of continuum envelope equations have internal states (Infeld and Rowlands 1990). One can ask about how much information a particle stores in its states. This is the amount of information that a particle transports across space and time and brings to interactions. These considerations lead one to a natural measure of the amount of structural complexity associated with individual particles.

Definition 33 *The* structural complexity $C(\alpha)$ *of a particle* α *is defined to be*

$$C(\alpha) = -\sum_{j=0}^{p-1} \Pr(q(\alpha^j)) \log_2 \Pr(q(\alpha^j)) \ , \tag{9.3}$$

where p is α's period and $\Pr(q(\alpha^j))$ is the probability of α being in phase α^j with the state-sequence $q(\alpha^j)$.

Remark 1. For the straightforward case of periodic particles, in which the wedges and so their associated state sequences are equally probable,

$$C(\alpha) = \log_2 P(\alpha) \ . \tag{9.4}$$

Remark 2. The information available to be processed in particle interactions is upper-bounded by the sum of the individual particle complexities, since this sum assumes independence of the particles. As we'll see shortly, the information in one particle, conditioned on the other's phase (via the constraints imposed by the mediating domain) and suitably averaged, determines the information available for processing by interactions.

9.2.2 Domain Transducer View of Particle Phases

A particle is bounded on either side by two patches of domain. (They could be patches of the same or different domains.) Consider what happens to the domain transducer as it scans across the part of the lattice containing the bounding domains (Λ^i and $\Lambda^{i'}$) and the particle (α). It begins by cycling through the states of the process graph of a phase (j) of the first bounding domain (Λ^i). It then encounters a symbol that does not belong to the language of that domain phase, and this then causes a transition out of that process graph. Each successive symbol of the particle wedge leads to additional transitions in the transducer. Finally, the transducer reaches cells at the beginning of the other bounding domain ($\Lambda^{i'}$), whereupon it begins to follow the process graph of $\Lambda^{i'}_{j'}$ at some appropriate phase j'. In this way, a particle wedge α^j corresponds to a sequence $q(\alpha^j)$ of transducer states.

More formally, the transducer maps a particle wedge α^j, bordered by Λ^i_j and $\Lambda^{i'}_{j'}$, to an ordered $n-$tuple ($n = |\alpha^j| + 2$) of states

$$Q(\alpha^j) \quad = \quad \left\langle q(\Lambda^i_{j,k}), q(\alpha^j), q(\Lambda^{i'}_{j',k'}) \right\rangle \ , \tag{9.5}$$

where $q(\Lambda^i_{j,k})$ is the transducer state reach on reading symbol $\Lambda^i_{j,k}$. Since the transducer-state sequence is determined by the bounding domain phases and the actual wedge α^j, it follows that the mapping from particle wedges to state sequences is 1-1. If two particle wedges correspond to the same sequence of states, then they are the same phase of the same particle, and vice versa.

This representation of particle phases will prove very handy below.

9.3 Interactions

In many CAs, when two or more particles collide they create another set of particles or mutually annihilate. Such *particle interactions* are denoted $\alpha + \beta \to \gamma$, for example. This means that the collision of an α particle

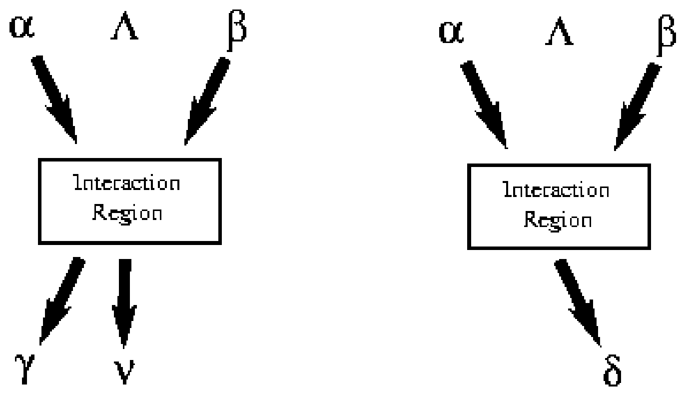

Figure 9.1: Interactions between an α and a β particle with domain Λ lying between.

on the left and a β particle on the right leads to the creation of a γ particle. Particle annihilation is denoted $\alpha + \beta \rightarrow \emptyset$. There are also *unstable* walls that can spontaneously decay into particles. This is denoted $\alpha \rightarrow \beta + \gamma$, for example.

Often, the actual product of a particle interaction depends on the phases α^j and β^k in which the interacting particles are at the time of collision. In such a case, there can be more than one interaction product for a particular collision: e.g., both $\alpha + \beta \rightarrow \gamma$ and $\alpha + \beta \rightarrow \emptyset$ can be observed.

The set of a CA's possible particle interactions is denoted **I**. The complete information about a CA's domains $\mathbf{\Lambda}$, particles **P**, and particle interactions **I** can be summarized in a *particle catalog.* The catalog forms a high-level description of the CA's dynamics. It is high-level in the sense of capturing the dynamics of emergent structures. The latter are objects on a more abstract level than the original equations of motion and raw (uninterpreted) spatial configurations of site values.

9.4 Bounding the Number of Interaction Products

Restricting ourselves to particle interactions with just two colliding particles — α and β, say — we'll now derive an upper bound on the number $n_{\alpha,\beta}$ of possible interaction products from a collision between them. (See Figure 9.1 for the interaction geometry.) In terms of the quantities just defined, the upper bound, stated as Theorem 17 below, is:

$$n_{\alpha,\beta} \leq \frac{P(\alpha)P(\beta)\Delta v}{T(\Lambda^i)S(\Lambda^i)} \;, \tag{9.6}$$

where $\Delta v = v_\alpha - v_\beta > 0$ and Λ^i is the domain in between the two particles before they collide. Note that if $\Delta v = 0$, then $n_{\alpha,\beta} = 0$ trivially.

For simplicity, let's assume that $\Delta v = v_\alpha - v_\beta \geq 0$. This simply means that particle α lies to the left of β and they move closer to each other over time, as in Figure 9.1.

This section proves that Eq. 9.6 is indeed a proper upper bound. The next section gives a number of examples, of both simple and complicated CAs, that show the bound is and is not attained. These highlight an important distinction between the number of possible interactions (i.e., what can enter the interaction region) and the number of unique interaction products (i.e., what actually leaves the interaction region).

To establish the bound, we'll need some intermediate results. The first three come from elementary number theory. Recall that the *least common multiple* $\text{lcm}(a, b)$ of two integers a and b is the smallest number c that is a multiple of both a and b. Similarly, the *greatest common divisor* $\gcd(a, b)$ of two integers a and b is the largest number c that divides both a and b.

Proposition 1 *(Burton 1976, Theorem 2.7)* $\gcd(ca, cb) = c\ \gcd(a, b),\ c > 0.$

Proposition 2 *(Burton 1976, Theorem 2.8)* $\gcd(a, b)\ \mathrm{lcm}(a, b) = ab.$

Lemma 21 $\mathrm{lcm}(ca, cb) = c\, \mathrm{lcm}(a, b),\ c > 0.$

Proof. Using Propositions 1 and 2, it follows that

$$\begin{aligned} \mathrm{lcm}(ca, cb) &= \frac{cacb}{\gcd(ca, cb)} \\ &= c\frac{ab}{\gcd(a, b)} \\ &= c\, \mathrm{lcm}(a, b)\ . \end{aligned} \tag{9.7}$$

QED.

Now we can start talking about particles.

Lemma 22 ***(Particle Periodicity Is a Multiple of Domain Periodicity)*** *The intrinsic periodicity $P(\alpha)$ of a particle α is a multiple of the temporal periodicity $T(\Lambda^i)$ of either domain Λ^i for which α is a boundary. That is,*

$$P(\alpha) = m_{\alpha i} T(\Lambda^i)\ , \tag{9.8}$$

for some positive integer $m_{\alpha i}$ that depends on α and Λ^i.

Proof. At any given time, a configuration containing the particle α consists of a patch of the domain Λ^i, a wedge belonging to α, and then a patch of $\Lambda^{i'}$, in that order from left to right. (Or right to left, if that is the chosen scan direction.) Fix the phase of α to be whatever you like — α^l, say. This determines the phases of Λ^i, for the following reason. Recall that, being a phase of a particle, α^l corresponds to a unique sequence $Q(\alpha^l)$ of transitions in the domain transducer. That sequence starts in a particular domain-phase state $\Lambda^i_{j,k}$ and ends in another domain-phase state $\Lambda^{i'}_{j',k'}$. So, the particle phase α^l occurs only at those times when Λ^i is in its j^{th} phase. Thus, the temporal periodicity of α must be an integer multiple of the temporal periodicity of Λ^i. By symmetry, the same is also true for the domain $\Lambda^{i'}$ to the right of the wedge. QED.

Corollary 5 (Phase Restriction) *Given that the domain Λ^i is in phase Λ^i_j at some time step, a particle α forming a boundary of Λ^i can only be in a fraction $1/T(\Lambda^i)$ of its $P(\alpha)$ phases at that time.*

Proof. This follows directly from Lemma 22.

Remark. Here is the first part of the promised restriction on the information in multiple particles. Consider two particles α and β, separated by a domain Λ^0. Naively, we expect α to contain $\log_2 P(\alpha)$ bits of information and β, $\log_2 P(\beta)$ bits. Given the phase of α, however, the phase of Λ^0 is fixed, and therefore the number of possible phases for β is reduced by a factor of $1/T(\Lambda^0)$. Thus the number of bits of information in the α-β pair is at most

$$\log_2 P(\alpha) + \log_2 P(\beta) - \log_2 T(\Lambda^0) = \log_2 \frac{P(\alpha)P(\beta)}{T(\Lambda^0)}\ . \tag{9.9}$$

The argument works equally well starting from β.

Lemma 23 *For any two particles α and β, the quantity $\mathrm{lcm}(P(\alpha), P(\beta))\Delta v$ is a non-negative integer.*

Proof. We know that the quantity is non-negative, since the least common multiple always is and Δv is so by construction. It remains to show that their product is an integer. Let $k_\alpha = \mathrm{lcm}(P(\alpha), P(\beta))/P(\alpha)$ and $k_\beta = \mathrm{lcm}(P(\alpha), P(\beta))/P(\beta)$; these are integers. Then

$$\begin{aligned}\Delta v &\equiv \frac{d_\alpha}{P(\alpha)} - \frac{d_\beta}{P(\beta)} \\ &= \frac{k_\alpha d_\alpha - k_\beta d_\beta}{\mathrm{lcm}(P(\alpha), P(\beta))} .\end{aligned}$$

When multiplied by $\mathrm{lcm}(P(\alpha), P(\beta))$ this is just $k_\alpha d_\alpha - k_\beta d_\beta$, which is an integer. QED.

Lemma 24 (Displacements Preserving Relative Phase) *When the distance d between two approaching particles α and β, in phases α^j and $\beta^{j'}$, respectively, is increased by* $\mathrm{lcm}(P(\alpha), P(\beta))\Delta v$ *sites, the original configuration — distance d and phases α^j and $\beta^{j'}$ — recurs after* $\mathrm{lcm}(P(\alpha), P(\beta))$ *time steps.*

Proof. From the definition of $\mathrm{lcm}(a, b)$ it follows directly that $\mathrm{lcm}(P(\alpha), P(\beta))$ is a multiple of $P(\alpha)$. Thus,

$$\alpha^{(j+\mathrm{lcm}(P(\alpha),P(\beta))) \bmod P(\alpha)} = \alpha^j , \tag{9.10}$$

and the α particle has returned to its original phase. Exactly parallel reasoning holds for the β particle. So, after $\mathrm{lcm}(P(\alpha), P(\beta))$ time steps both α and β are in the same phases α^j and $\beta^{j'}$ again. Furthermore, in the same amount of time the distance between the two particles has decreased by $\mathrm{lcm}(p_\alpha, p_\beta)\Delta v$, which is the amount by which the original distance d was increased. (By Lemma 23, that distance is an integer, and so we can meaningfully increase the particles' separation by this amount.) Thus, after $\mathrm{lcm}(P(\alpha), P(\beta))$ time steps the original configuration is restored. QED.

Lemma 25 (Phase-Preserving Displacements and Spatial Periodicity) *If Λ^i is the domain lying between two particles α and β, then the ratio*

$$r = \frac{\mathrm{lcm}(P(\alpha), P(\beta))\Delta v}{S(\Lambda^i)} \tag{9.11}$$

is an integer.

Proof. Suppose, without loss of generality, that the particles begin in phases α^0 and β^0, at some substantial distance from each other. We know from the previous lemma that after a time $\mathrm{lcm}(P(\alpha), P(\beta))$ they will have returned to those phases and narrowed the distance between each other by $\mathrm{lcm}(P(\alpha), P(\beta))\Delta v$ cells. What the lemma asserts is that this displacement is some integer multiple of the spatial periodicity of the intervening domain Λ^i. Call the final distance between the particles d. Note that the following does not depend on what d happens to be.

Each phase of each particle corresponds to a particular sequence of transducer states — those associated with reading the particle's wedge for that phase. Reading this wedge from left to right (say), we know that $Q(\alpha^0)$ must end in some phase-state of the domain Λ^i; call it $\Lambda^i_{0,0}$. Similarly, $Q(\beta^0)$ must *begin* with a phase-state of Λ^i, but, since every part of the intervening domain is in the same phase, this must be a state of the *same* phase Λ^i_0; call it $\Lambda^i_{0,k}$. In particular, consistency requires that k be the distance between the particles modulo $S(\Lambda^i)$. But this is true both in the final configuration, when the separation between the particles is d, and in the initial configuration, when it is $d + \mathrm{lcm}(P(\alpha), P(\beta))\Delta v$. Therefore

$$\begin{aligned} d + \mathrm{lcm}(P(\alpha), P(\beta))\Delta v &= d \pmod{S(\Lambda^i)} \\ \mathrm{lcm}(P(\alpha), P(\beta))\Delta v &= 0 \pmod{S(\Lambda^i)} .\end{aligned}$$

Thus, $\mathrm{lcm}(P(\alpha), P(\beta))\Delta v$ is an integer multiple of the spatial period $S(\Lambda^i)$ of the intervening domain Λ^i. QED.

Remark. It is possible that $\mathrm{lcm}(P(\alpha), P(\beta))\Delta v = 0$, but this does not affect the subsequent argument. Note that if this is the case, then, since the least common multiple of the periods is at least 1, $\Delta v = 0$. This, in turn, implies that the particles do not, in fact, collide and interact, and so the number of interaction products is simply zero. The formula gives the proper result in this case.

The next result follows easily from Proposition 1 and Lemma 22.

Lemma 26 (Relation of the Periods of Particles Bounding a Common Domain) *If Λ^i is the domain lying between particles α and β, then*

$$\gcd(P(\alpha), P(\beta)) = T(\Lambda^i)\gcd(m_{\alpha i}, m_{\beta i}) \ . \tag{9.12}$$

Proof. Apply Lemma 21:

$$\begin{aligned}\gcd(P(\alpha), P(\beta)) &= \gcd(m_{\alpha i}T(\Lambda^i), m_{\beta i}T(\Lambda^i)) \\ &= T(\Lambda^i)\gcd(m_{\alpha i}, m_{\beta i}).\end{aligned}$$

QED.

With the above lemmas the following theorem can be proved, establishing an upper bound on the number of possible particle interaction products.

Theorem 17 (Hordijk's Rule) *The number $n_{\alpha,\beta}$ of products of an interaction between two approaching particles α and β with a domain Λ^i lying between is at most*

$$n_{\alpha,\beta} \leq \frac{P(\alpha)P(\beta)\Delta v}{T(\Lambda^i)S(\Lambda^i)} \ . \tag{9.13}$$

Proof. First, let's show that this quantity is an integer. Use Proposition 2 to get

$$\frac{P(\alpha)P(\beta)\Delta v}{T(\Lambda^i)S(\Lambda^i)} = \frac{\gcd(P(\alpha), P(\beta))\mathrm{lcm}(P(\alpha), P(\beta))\Delta v}{T(\Lambda^i)S(\Lambda^i)} \ , \tag{9.14}$$

and then Lemma 25 to find that

$$\frac{P(\alpha)P(\beta)\Delta v}{T(\Lambda^i)S(\Lambda^i)} = \frac{\gcd(P(\alpha), P(\beta))r}{T(\Lambda^i)} \ , \tag{9.15}$$

and finally Lemma 26 to show that

$$\begin{aligned}\frac{P(\alpha)P(\beta)\Delta v}{T(\Lambda^i)S(\Lambda^i)} &= \frac{T(\Lambda^i)\gcd(m_{\alpha i}, m_{\beta i})r}{T(\Lambda^i)} \\ &= r\gcd(m_{\alpha i}, m_{\beta i}) \ ,\end{aligned} \tag{9.16}$$

which is an integer.

Second, assume that, at some initial time t, the two particles are in some arbitrary phases α^j and $\beta^{j'}$, respectively, and that the distance between them is d cells. This configuration gives rise to a particular particle-phase combination at the time of collision. Since the global update function is deterministic, the combination, in turn, gives one and only one interaction result. Now, increase the distance between the two particles, at time t, by one cell, while keeping their phases fixed. This gives rise to a different particle-phase combination at the time of collision and, thus, possibly to a different interaction result. We can repeat this operation of increasing the distance by one cell $\mathrm{lcm}(P(\alpha), P(\beta))\Delta v$ times. At that point, however, we know from Lemma 24 that after $\mathrm{lcm}(P(\alpha), P(\beta))$ time steps the particles find themselves again in phases α^j and

$\beta^{j'}$ at a separation of d. That is, they are in exactly the original configuration and their interaction will therefore also produce the original product, whatever it was.

Starting the two particles in phases α^j and $\beta^{j'}$, the particles go through a fraction $1/\gcd(P(\alpha), P(\beta))$ of the possible $P(\alpha)P(\beta)$ phase combinations, over $\mathrm{lcm}(p_\alpha, p_\beta)$ time steps, before they start repeating their phases again. So, the operation of increasing the distance between the two particles by one cell at a time needs to be repeated for $\gcd(P(\alpha), P(\beta))$ different initial phase combinations. This way all possible phase combinations with all possible distances (modulo $\mathrm{lcm}(P(\alpha), P(\beta))\Delta v$) are encountered. Each of these can give rise to a different interaction result.

From this one sees that there are at most

$$\gcd(P(\alpha), P(\beta))\mathrm{lcm}(P(\alpha), P(\beta))\Delta v = P(\alpha)P(\beta)\Delta v \tag{9.17}$$

unique particle-domain-particle configurations. And so, there are at most this many different particle interaction products, given that Φ is many-to-one. (Restricted to the homogeneous, quiescent ($\Lambda = 0^*$) domain which has $T(\Lambda) = 1$ and $S(\Lambda) = 1$, this is the result, though not the argument, of Park, Steiglitz and Thurston (1986).)

However, given the phases α^j and $\beta^{j'}$, the distance between the two particles cannot always be increased by an arbitrary number of cells. Keeping the particle phases α^j and $\beta^{j'}$ fixed, the amount Δd by which the distance between the two particles can be increased or decreased is a multiple of the spatial periodicity $S(\Lambda^i)$ of the intervening domain. The argument for this is similar to that in the proof of Lemma 25. Consequently, of the $\mathrm{lcm}(P(\alpha), P(\beta))\Delta v$ increases in distance between the two particles, only a fraction $1/S(\Lambda^i)$ are actually possible.

Furthermore, and similarly, not all arbitrary particle-phase combinations are allowed. Choosing a phase α^j for the α particle subsequently determines the phase Λ^i_j of the domain Λ^i for which α forms one boundary. From Corollary 5 it then follows that only a fraction $1/T(\Lambda^i)$ of the $P(\beta)$ phases are possible for the β particle which forms the other boundary of Λ^i.

Adjusting the number of possible particle-domain-particle configurations that can give rise to different interaction products according to the above two observations results in a total number

$$\frac{P(\alpha)P(\beta)\Delta v}{T(\Lambda^i)S(\Lambda^i)} \tag{9.18}$$

of different particle-phase combinations and distances between two particles α and β. Putting the pieces together, then, this number is an upper bound on the number $n_{\alpha,\beta}$ of different interaction products. QED.

Remark 1. As we'll see in the examples, on the one hand, the upper bound is strict, since it is saturated by some interactions. On the other hand, there are also interactions that do not saturate it.

Remark 2. We saw (Corollary 5, Remark) that the information in a pair of particles α and β, separated by a patch of domain Λ^i, is at most

$$\log_2 \frac{P(\alpha)P(\beta)}{T(\Lambda^i)} \tag{9.19}$$

bits. In fact, Hordijk's Rule implies a stronger restriction. The amount of information the interaction carries about its inputs is, at most, $\log_2 n_{\alpha,\beta}$ bits, since there are only $n_{\alpha,\beta}$ configurations of the particles that can lead to distinct outcomes. If the number of outcomes is less than $n_{\alpha,\beta}$, the interaction effectively performs an irreversible logical operation on the information contained in the input particle phases.

Remark 3. This is "Hordijk's Rule" because Wim Hordijk was the first person to notice that, empirically, it was valid, and to use it in analyzing cellular automata. This proof is joint work with Wim and Jim Crutchfield.

9.5 Examples

9.5.1 ECA 54 and Intrinsic Periodicity

Figure 9.2 shows the raw and domain-transducer filtered space-time diagrams of ECA 54, starting from a random initial configuration. First, let's review the results of Hanson and Crutchfield (1997) for ECA 54's particle dynamics.

Figure 9.3 shows a space-time patch of ECA 54's dominant domain Λ, along with the domain transducer constructed to recognize and filter it out, as was done to produce Figure 9.2(b).

Examining Figure 9.2 shows that there are four particles, called α, β, γ^+, and γ^-. The first two have zero velocity; they are the larger particles seen in Figure 9.2(b). The γ particles have velocities 1 and -1, respectively. They are seen in the figure as the diagonally moving "light" particles that mediate between the "heavy" α and β particles.

The analysis in Hanson and Crutchfield (1997) identified 7 dominant two- and three-particle interactions. Let's now analyze just one: the $\gamma^+ + \gamma^- \to \beta$ interaction to illustrate the importance of a particle's intrinsic periodicity.

Naive analysis would simply look at the space-time diagram, either the raw or filtered ones in Figure 9.2, and conclude that these particles had periodicities $P(\gamma^+) = P(\gamma^-) = 1$. Plugging this and the other data — $T(\Lambda) = 2$, $S(\Lambda) = 4$, and $\Delta v = 2$ — leads to upper bound $n_{\alpha,\beta} = 1/4$! This is patently wrong; it's not even an integer.

Figure 9.4 gives the transducer-filtered space-time diagram for the γ^+ and γ^- particles. The domain Λ is filtered out, as above. In the filtered diagrams the transducer state reached on scanning the particle wedge cells is indicated.

From the space-time diagrams of Figure 9.4(b) one notes that the transducer-state labeled wedges for each particle indicate that their intrinsic periodicities are $P(\gamma^+) = 2$ and $P(\gamma^-) = 2$. Then, from Theorem 17, $n_{\alpha,\beta} = 1$. That is, there is at most one product of these particles' interaction.

Figure 9.5 gives the transducer-filtered space-time diagram for the $\gamma^+ + \gamma^- \to \beta$ interaction. A complete survey of all possible $\gamma^+\Lambda\gamma^-$ initial particle configurations shows that this is the only interaction for these particles. Thus, the upper bound is saturated.

9.5.2 An Evolved CA

The second example for which we test the upper bound is a CA that was evolved by a genetic algorithm to perform a class of spatial computations: from all random initial configurations, synchronize within a specified number of iterations. This CA is ϕ_{sync_1} of Hordijk, Mitchell and Crutchfield (1998): a binary, radius-3 CA. The 128-bit look-up table for ϕ_{sync_1} is given in Table 9.1.

Here we're only interested in locally analyzing the various pairwise particle interactions observed in ϕ_{sync_1}. It turned out that this CA used a relatively simple set of domains, particles, and interactions. Its particle catalog is given in Table 9.2.

As one example, the two particles α and β and the intervening domain Λ have the properties given in Table 9.2. From this data, Theorem 17 tells us that there is at most one interaction product:

$$n_{\alpha,\beta} = \frac{4 \cdot 2 \cdot \frac{1}{4}}{2 \cdot 1} = 1 \ . \tag{9.20}$$

The single observed interaction between the α and β particles is shown in Figure 9.6. As this space-time diagram shows, the interaction creates another β particle, i.e., $\alpha + \beta \to \beta$. An exhaustive survey of the 8 ($= 4 \times 2$) possible particle-phase configurations shows that this is the only interaction for these two particles. Thus, in this case, Hordijk's Rule again gives a tight bound; it cannot be reduced.

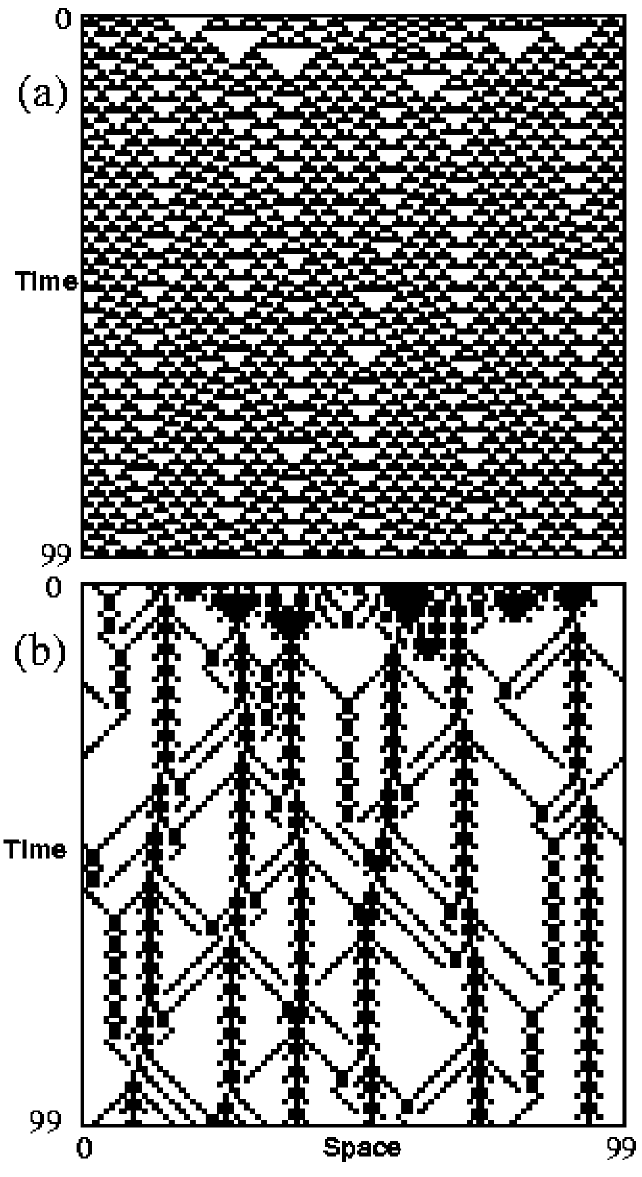

Figure 9.2: (a) Raw space-time diagram and (b) filtered space-time diagram of ECA 54 behavior starting from an arbitrary initial configuration. After Hanson and Crutchfield (1997).

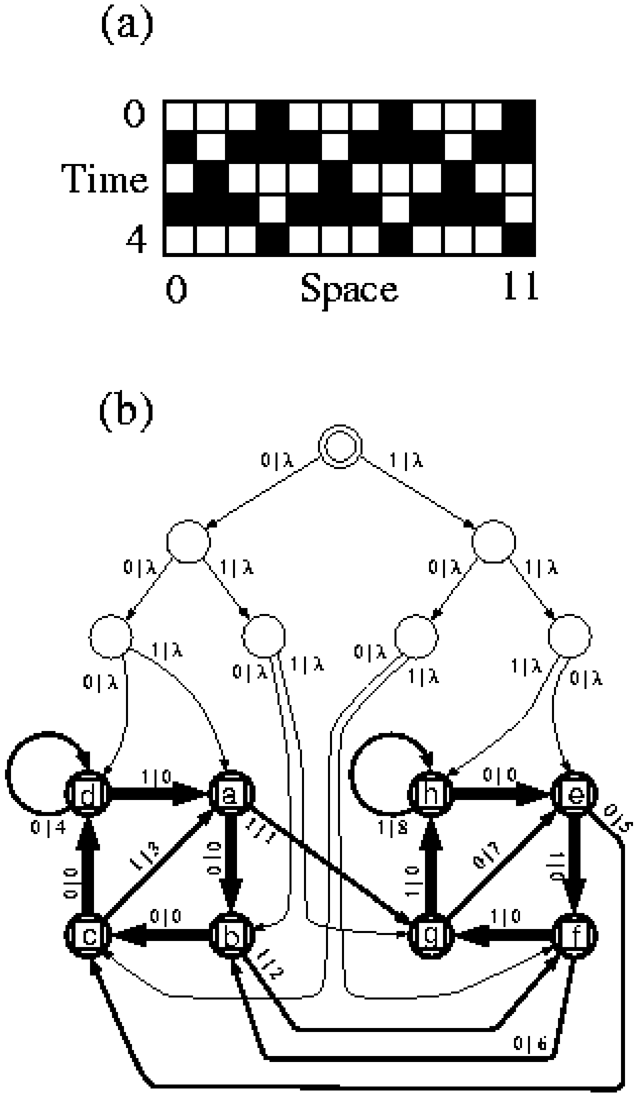

Figure 9.3: (a) Space-time patch of ECA54's primary domain Λ. (b) The transducer that recognizes Λ and deviations from it. After Hanson and Crutchfield (1997).

ϕ	Look-up Table (hexadecimal)
ϕ_{sync_1}	F8A19CE6B65848EA
	D26CB24AEB51C4A0
ϕ_{parent}	CEB2EF28C68D2A04
	E341FAE2E7187AE8

Table 9.1: Lookup tables (in hexadecimal) for ϕ_{sync_1} and ϕ_{parent}. To recover the 128-bit string giving the CA look-up table output bits s_{t+1}, expand each hexadecimal digit (the first row followed by the second row) to binary. The output bits s_{t+1} are then given in lexicographic order starting from the all-0s neighborhood at the leftmost bit in the 128-bit string.

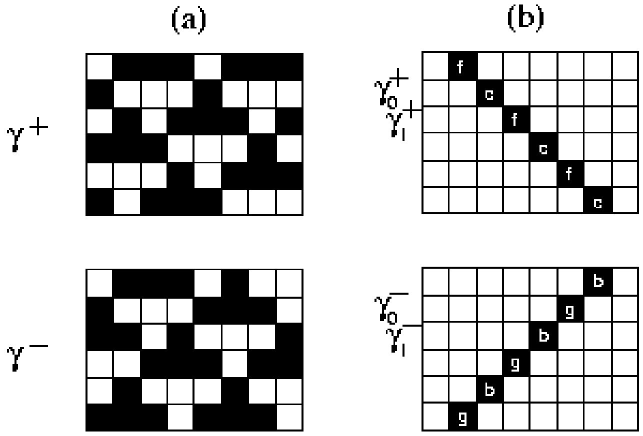

Figure 9.4: The transducer-filtered space-time diagrams for the γ^+ and γ^- particles. (a) The raw space-time patches containing the particles. (b) The same patches with the Λ filtered out. The cells not in Λ are denoted in black; those in Λ in white. In the filtered diagrams the transducer state reached on scanning the particle wedge cells is indicated. After Hanson and Crutchfield (1997).

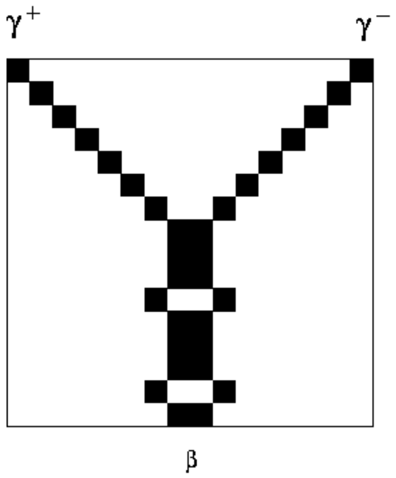

Figure 9.5: The transducer-filtered space-time diagrams for the $\gamma^+ + \gamma^- \to \beta$ interaction. After Hanson and Crutchfield (1997).

ϕ_{sync_1} **Particle Catalog**				
Domains Λ				
Name	Regular language	$T(\Lambda)$	$S(\Lambda)$	
Λ	$0^4 0^*$, $1^4 1^*$	2	1	
Particles P				
Name	Wall	P	d	v
α	$\Lambda_j \Lambda_j$	4	-1	-1/4
β	$\Lambda_j \Lambda_{1-j}$	2	-1	-1/2
γ	$\Lambda_j \Lambda_j$	8	-1	-1/8
δ	$\Lambda_j \Lambda_j$	2	0	0
Interactions I				
Type	Interaction	Interaction		
React	$\alpha + \beta \to \beta$	$\gamma + \beta \to \beta$		
React	$\delta + \beta \to \beta$	$\gamma + \alpha \to \alpha$		
React	$\delta + \alpha \to \alpha$	$\delta + \gamma \to \alpha$		

Table 9.2: The particle catalog of ϕ_{sync_1}. Λ_j, $j \in \{0, 1\}$, indicates the two temporal phases of domain Λ.

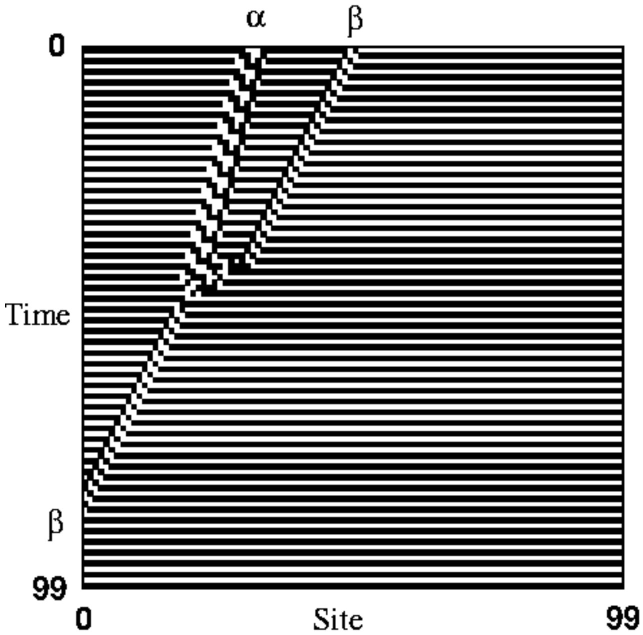

Figure 9.6: The interaction between an α and a β particle in ϕ_{sync_1}.

ϕ_{parent} **Particle Properties**			
Domain	T	S	
Λ	2	1	
Particle	P	d	v
α	8	2	1/4
β	2	-3	-3/2

Table 9.3: Properties of two of ϕ_{parent}'s particles.

9.5.3 Another Evolved CA

The third, more complicated example is also a CA that was evolved by a genetic algorithm to synchronize. This CA is ϕ_{parent} of Crutchfield, Hordijk and Mitchell (2000b). It too is a binary radius-3 CA. The 128-bit look-up table for ϕ_{parent} was given in Table 9.1.

Here the two particles α and β and the intervening domain Λ have the properties given in Table 9.3. Note that this is the same domain as in the preceding example.

From this data, Theorem 17 now says that there are at most:

$$n_{\alpha,\beta} = \frac{8 \cdot 2 \cdot \frac{7}{4}}{2 \cdot 1} = 14 \tag{9.21}$$

interactions.

Of these 14 input configurations, it turns out several give rise to the same products. From a complete survey of α-Λ-β configurations, the result is that there are actually only 4 different products from the $\alpha + \beta$ interaction; these are:

$$\begin{aligned} \alpha + \beta &\rightarrow \emptyset \\ \alpha + \beta &\rightarrow \gamma \\ \alpha + \beta &\rightarrow 2\beta \\ \alpha + \beta &\rightarrow \beta + \alpha \end{aligned}$$

They are shown in Figure 9.7.

This example serves to highlight the distinction between the maximum number of interaction configurations, as bounded by Theorem 17, and the actual number of unique products of the interaction. We'll come back to this.

9.5.4 ECA 110

In the next example, we test Theorem 17 on one of the long-appreciated "complex" CA, elementary CA 110. As long ago as 1986, Wolfram (Wolfram 1986, Appendix 15) conjectured that this rule is able to support universal, Turing-equivalent computation (replacing an earlier dictum (Wolfram 1984b, p. 31) that all elementary CA are "too simple to support universal computation"). While this conjecture initially excited little interest, in the last few years it has won increasing acceptance in the CA research community. Though to date there is no published proof of universality, there are studies of its unusually rich variety of domains and particles, one of the most noteworthy of which is McIntosh's work on their tiling and tessellation properties (McIntosh 2000). Because of this CA's behavioral richness, I won't present its complete particle catalog and computational-mechanical analysis here; rather see Crutchfield and Shalizi (2001). Instead, I'll look at a single type of reaction where the utility of Hordijk's Rule is particularly notable.

Consider one domain, labeled Λ^0, and two particles that move through it, called β and κ (Crutchfield and Shalizi 2001). (This β particle is not to be confused with the β of the previous examples.) Λ^0 is ECA 110's

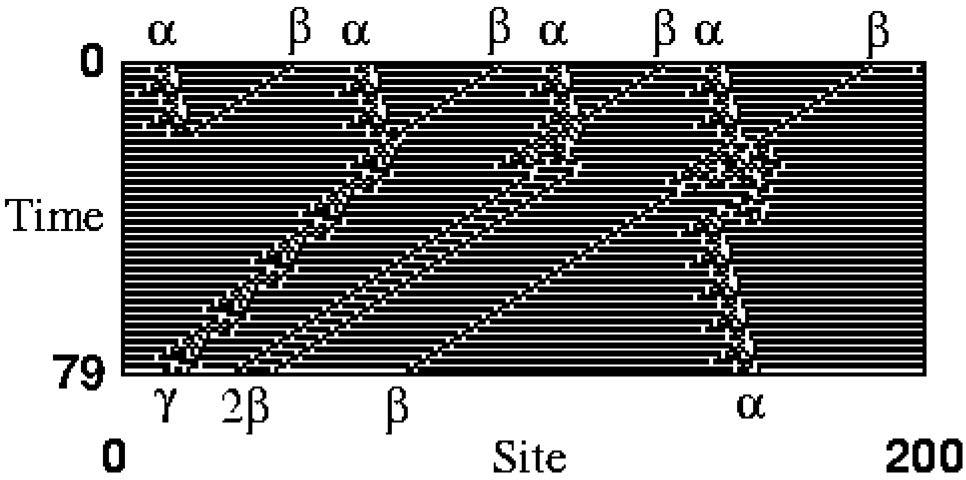

Figure 9.7: The four different (out of 14 possible) interaction products for the $\alpha + \beta$ interaction.

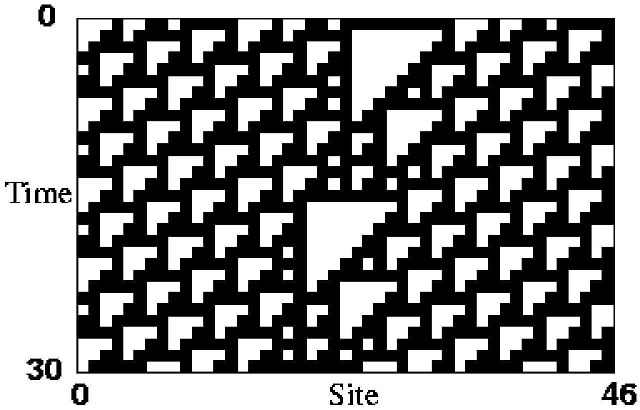

Figure 9.8: The particle β of ECA 110: The space-time patch shows two complete cycles of particle phase.

"true vacuum": the domain that is stable and overwhelmingly the most prominent in space-time diagrams generated from random samples of initial configurations. It has a temporal period $T(\Lambda^0) = 1$, but a spatial period $S(\Lambda^0) = 14$. The β particle has a period $P(\beta) = 15$, during the course of which it moves four steps to the left: $d_\beta = 4$. The κ particle, finally, has a period $P(\kappa) = 42$, and moves $d_\kappa = 14$ steps to the left during its cycle. This data gives the β particle a velocity of $v_\beta = -4/15$ and the κ particle $v_\kappa = -1/3$.

Naively, one would expect to have to examine 630 ($= P(\beta)P(\kappa) = 15 \times 42$) different particle-phase configurations to exhaust all possible interactions. Theorem 17, however, tells us that all but

$$\frac{(15)(42)(\frac{-4}{15} - \frac{-1}{3})}{(14)(1)} = 3 \tag{9.22}$$

of those initial configurations are redundant. In fact, an exhaustive search shows that there are exactly three

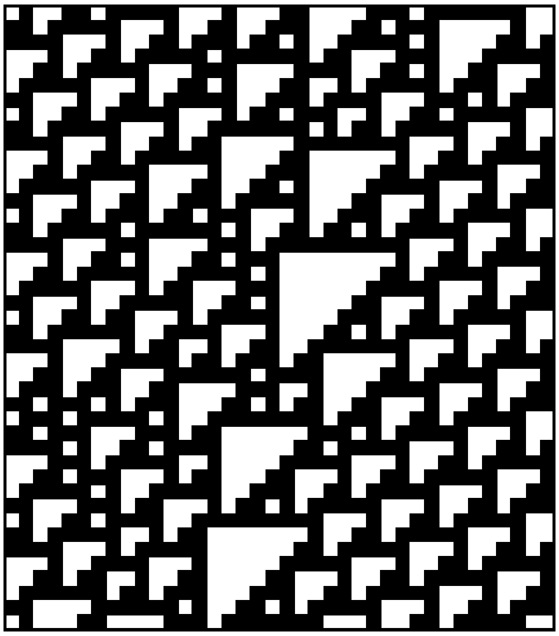

Figure 9.9: The particle κ of ECA 110: The space-time diagram shows one complete cycle of particle phase.

distinct interactions:

$$\begin{aligned} \beta + \kappa &\rightarrow \alpha + 3w_{right} \ , \\ \beta + \kappa &\rightarrow \beta + 4w_{right} \ , \\ \beta + \kappa &\rightarrow \eta \ . \end{aligned}$$

Here, α, w_{right}, and η are additional particles generated by ECA 110. These interactions are depicted, respectively, in Figures 9.11, 9.10, and 9.12.

The w_{right} particle is somewhat unusual in that several can propagate side by side, or even constitute a domain of their own. There are a number of such "extensible" particle families in ECA 110 (Crutchfield and Shalizi 2001).

Finally, observe that, though all these particles are wide and have long periods, and move through a complicated background domain, Hordijk's Rule is not just obeyed, but gives the exact number of interaction products. I'll come back to what significance this might have in the conclusion to this chapter.

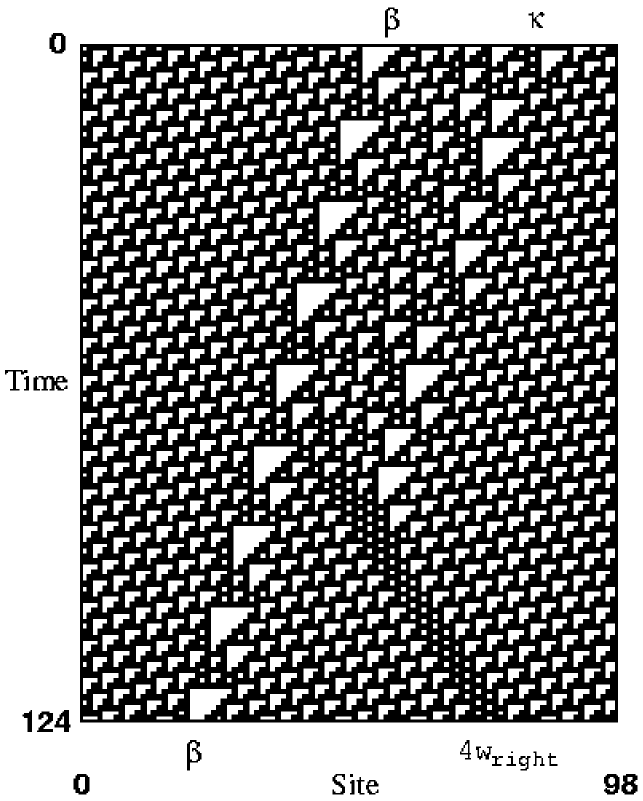

Figure 9.10: The reaction $\beta + \kappa \rightarrow \beta + 4w_{right}$ in ECA 110.

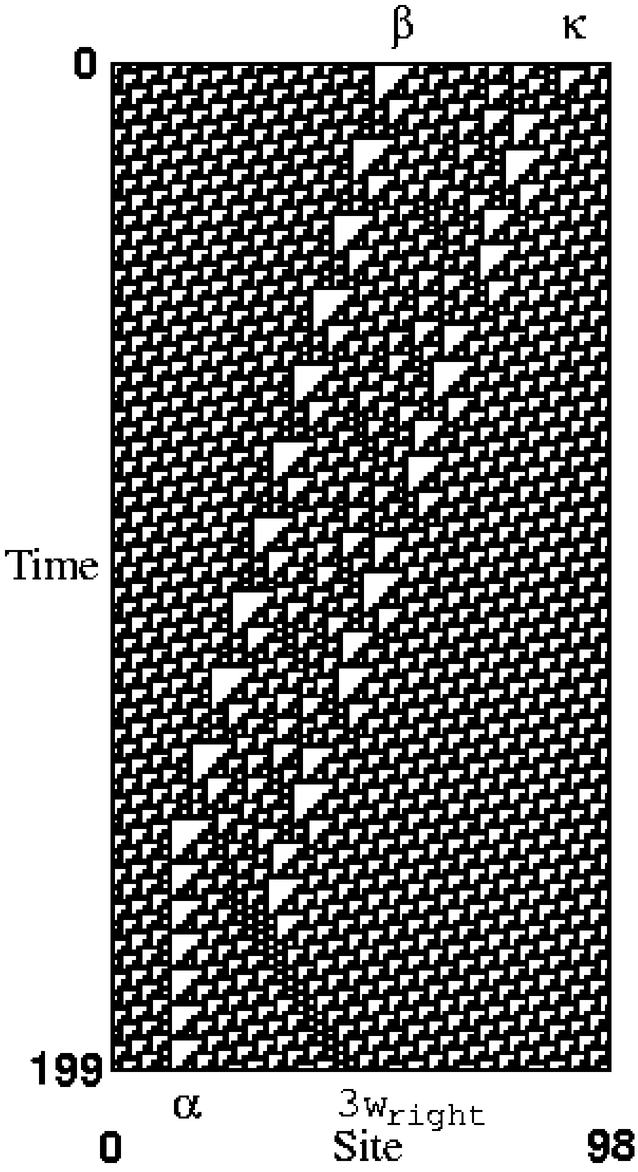

Figure 9.11: The reaction $\beta + \kappa \rightarrow \alpha + 3w_{right}$ in ECA 110.

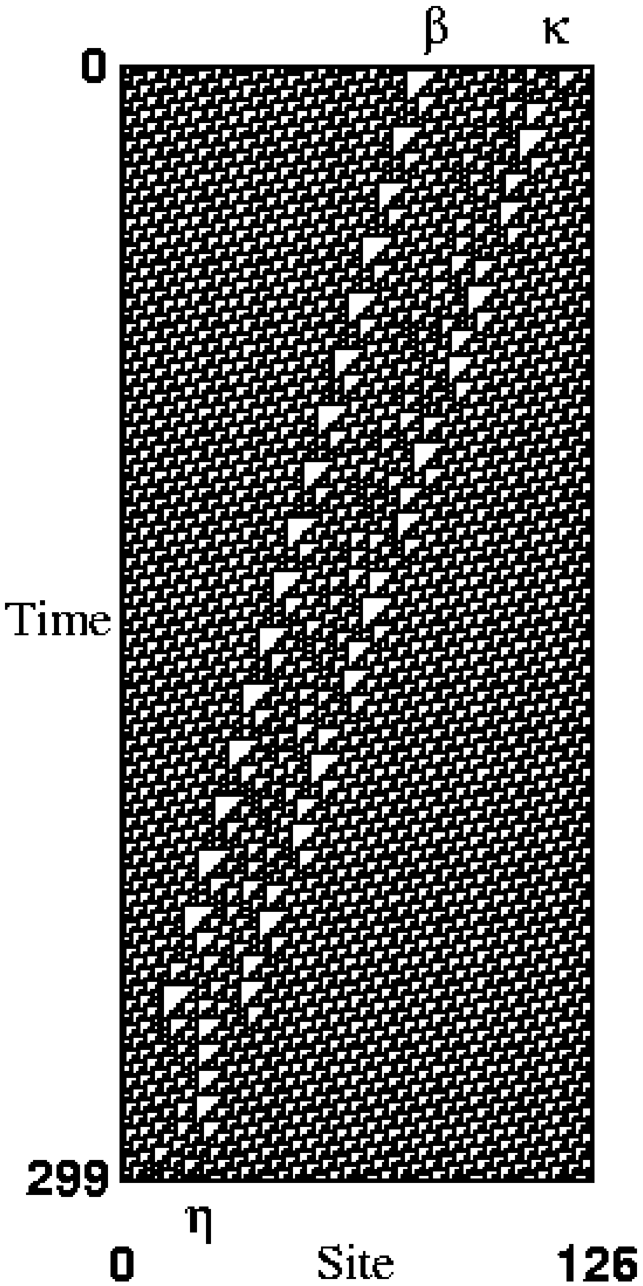

Figure 9.12: The reaction $\beta + \kappa \rightarrow \eta$ in ECA 110.

9.6 Conclusion

9.6.1 Summary

The original interaction product formula of Park, Steiglitz and Thurston (1986) is limited to particles propagating in a completely uniform background; i.e., to a domain whose spatial and temporal periods are both 1. When compared to the rich diversity of domains generated by CAs, this is a considerable restriction, and so the formula does not help in analyzing many CAs. We've generalized their result and along the way established a number of properties of domains and particles — structures defined by CA computational mechanics. The examples showed that the upper bound is tight and that, in complex CAs, particle interactions are substantially less complicated than they look at first blush. Moreover, in developing the bound for complex domains, the analysis elucidated the somewhat subtle notion of a particle's intrinsic periodicity — a property not apparent from the CA's raw space-time behavior: it requires rather an explicit representation of the bordering domains' structure.

Understanding the detailed structure of particles and their interactions moves us closer to an engineering discipline that would tell one how to design CA to perform a wide range of spatial computations using various particle types, interactions, and geometries. In a complementary way, it also brings us closer to scientific methods for analyzing the intrinsic computation of spatially extended systems (Chapter 10).

9.6.2 Open Problems

The foregoing analysis merely scratches the surface of a detailed analytical approach to CA particle "physics": Each CA update rule specifies a microphysics of local (cell-to-cell) space and time interactions for its universe; the goal is to discover and analyze those emergent structures that control the macroscopic behavior. We'll return to that problem in the next chapter, but first I'll list a few of questions raised by these results.

It would be preferable to directly calculate the number of products coming out of the interaction region, rather than (as here) the number of distinct particle-domain-particle configurations coming into the interaction region. We believe this is eminently achievable, given the detailed representations of domain and particles that are entailed by a computational mechanics analysis of CAs.

Two very desirable extensions of these results suggest themselves. The first is to go from strictly periodic domains to cyclic (periodic and "chaotic") domains and then to general domains. The principle difficulty here is that Proposition 20 plays a crucial role in the current proof, but we do not yet see how to generalize its proof to chaotic (positive entropy density) domains. The second extension would be to incorporate aperiodic particles, such as the simple one exhibited by ECA 18 (Crutchfield and Hanson 1993a). We suspect this will prove considerably more difficult than the extension to cyclic domains: it is not obvious how to apply notions like "particle period" and "velocity" to these defects. A third extension, perhaps more tractable than the last, is to interactions of more than two particles. The geometry and combinatorics will be more complicated than in the two-particle case, but we conjecture that it will be possible to establish an upper bound on the number of interaction products for $n-$particle interactions via induction.

Does there exist an analogous lower bound on the number of interactions? If so, when do the upper and lower bounds coincide?

In solitonic interactions the particle number is preserved (Peyrard and Kruskal 1984; Aizawa, Nishikawa and Kaneko 1991; Park, Steiglitz and Thurston 1986; Steiglitz, Kamal and Watson 1988; Ablowitz, Kruskal and Ladik 1979). What are the conditions on the interaction structure that characterize solitonic interactions? The class of soliton-like particles studied in Park, Steiglitz and Thurston (1986) possess a rich "thermodynamics" closely analogous to ordinary thermodynamics, explored in detailed in Goldberg (1988). Do these results generalize to the broader class of domains and particles, as the original upper bound of Park, Steiglitz and Thurston (1986) does?

While the particle catalog for ECA 110 is not yet *provably* complete, for every known pair of particles the number of distinct interaction products is exactly equal to the upper bound given by Hordijk's Rule. This is not generally true of most of the CAs we have analyzed and is especially suggestive in light of the widely-accepted conjecture that the rule is computation universal. We suspect that ECA 110's fullness or behavioral flexibility is connected to its computational power. (Cf. Remark 2 to Theorem 17.) However, we have yet to examine other, computation universal CA to see whether they, too, saturate the Hordijk's Rule bound. One approach to this question would be to characterize the computational power of systems employing different kinds of interactions, as is done in Jakubowski, Steiglitz and Squier (1997) for computers built from interacting (continuum) solitary waves.

Chapter 10

Spatio-temporal Computational Mechanics

Not in the spaces we know, but *between* them, They walk serene and primal, undimensioned and to us unseen.
— Abdul Alhazred (c. 750)

10.1 The Difficulties of Higher Dimensions

It may not have escaped the reader's attention that, while Chapter 9 spoke freely of "hidden states" in a CA, I was quite vague about just what those states were, and how they related to the hidden states constructed in the previous chapters. It would be *nice* to say that those states are causal states, in some sense, but it's not clear what that might be. It's hard to see in what sense the value at one point in the lattice is caused by the values of its neighbors *at that time*, for instance. So the justification of the domain-and-particle methods has been somewhat pragmatic — you can think of the CAs as doing things to regular languages, and those languages can be represented by machines with states, and the results are fruitful — but also unsatisfactory.

Now, of course, CA are dynamical systems (we've been over that at some length in Section 8.2), so we could apply computational mechanics to them, at the global level, in a very straightforward manner. The causal states we'd derive in this way would capture all the patterns relevant to the global evolution of the CA, and so in some sense all the information about their spatial structure would be encoded into the global ϵ-machine. On the other hand, that encoding would be very hairy indeed, and we'd really like something where the spatial structure was *transparent*, just as the ϵ-machine makes the causal structure transparent. So what we're looking for are spatially-localized states, which we can somehow link up with the states from spatial computational mechanics of CAs. Our desired states should also be causal, in some reasonable sense, and have the optimality properties to which we've become accustomed.

An obvious first step is to turn to information theory and automata theory, since they served us so well with time-series. Unfortunately, nobody really knows how to work either of those theories in higher dimensions. Automata theory, in particular, gets really ugly (Lindgren, Moore and Nordahl 1998), and information theory isn't much better. (See Eriksson and Lindgren 1987; Lempel and Ziv 1986; Andrienko, Brilliantov and Kurths 2000; Feixas, del Acebo, Bekaert and Sbert 1999; Garncarek and Piasecki 1999; Piasecki 2000 for attempts to extend information theory to fields in two or more dimensions.) That avenue being blocked, a natural second step is to look to statistics. The statistical analysis of spatial data is notoriously difficult in all but the most trivial cases (Ripley 1981; Ripley 1988; Cressie 1993; Grenander 1996). In part this is because the proper treatment of spatial stochastic processes is also notoriously difficult (see Schinazi (1999) and Guttorp (1995, ch. 4) for gentle introductions to spatial processes and spatial statistics, respectively. Griffeath (1979), Liggett (1985) and Guyon (1995) are more advanced treatments).

In fact, we are in the uncomfortable position of having to strike out more or less on our own.

10.2 Global Causal States for Spatial Processes

In what follows, we'll consider, not spatial processes in general, but those whose space, time and state are all discrete, and space is a regular lattice. While this is a very broad class of processes — it includes all cellular automata, for instance — it is still a limited one. At the end of this chapter and in the conclusion, I'll talk about how these assumptions might be relaxed.

We suppose that at each point in space can be in one of a finite number of states, drawn from a finite alphabet $\mathcal{A}$. Thus, a global configuration is an element of $\mathcal{A}^S$, where S is the lattice — either a finite number, for a finite lattice, or $\mathbb{Z}^d$ for a d-dimensional infinite lattice. The global spatial process is a distribution over sequence of global configurations, an ensemble of elements of $\mathcal{A}^S \times T$, where $T = \mathbb{N}$ if there is a definite starting-time, or $= \mathbb{Z}$ if time extends in both directions.

We write the random variable for the sequence of all configurations up to and including time t as $\overleftarrow{G}(t)$; we call its particular values *pasts* or *histories*. The future sequence is $\overrightarrow{G}(t)$.

For simplicity, we assume that the process is invariant under spatial translations (but not necessarily any other element of the space group).

Definition 34 (Global Causal State) *The global causal state of a history is the set of all histories which yield the same conditional distribution of futures. We write the random variable for the global causal state as $\mathcal{G}$ (realizations γ), and the function from history to causal state as ϵ. That is, $\mathcal{G} = \epsilon(\overleftarrow{G})$, and*

$$\epsilon(\overleftarrow{G}) \quad = \quad \left\{ \overleftarrow{g}' | \forall \overrightarrow{g},\ \mathrm{P}(\overrightarrow{G}{=}\overrightarrow{g} \mid \overleftarrow{G}{=}\overleftarrow{g}) = \mathrm{P}(\overrightarrow{G}{=}\overrightarrow{g} \mid \overleftarrow{G}{=}\overleftarrow{g}') \right\} \tag{10.1}$$

Note that while global causal states are defined in a time-invariant manner, they do not necessarily have a time-invariant distribution. In particular, if the global process is non-stationary, the distribution over global causal states will be non-stationary.

All the properties of normal causal states, from chapter 4 are inherited.

10.2.1 Why Global States Are not Enough

While in a sense knowing the global causal state tells us all there is to know about the future of the spatial process, it is not the ideal representation of the process's structure, for three reasons.

First, there is no *explicit* representation of the spatial structure. It is encoded, to be sure, in the global causal state, but generally not in any way which is easily comprehended by human beings. In many CA, for instance, the presence or absence of phase-defects makes a great deal of difference to the dynamics. This would be reflected by differing global causal states, but not in any way which made it particularly clear what made the difference.

Second, the number of global causal states is apt to be very large — in the case of deterministic CA, roughly on the order of the number of lattice configurations. (The exact number would depend on the degree of irreversibility of the update rule.) This is not a particularly compact representation of the spatial process's structure, nor one which lends itself easily to calculation.

Third, getting adequate statistics to empirically estimate the global causal states is simply not practical.

For all these reasons, the global causal states approach to spatial processes, while valid, is useless. What we would like, instead, is some way of factoring or distributing the information contained in the global causal state across the lattice — of finding *local* causal states. It is to this question that we now turn.

10.3 Local Causal States

A number of different ideas have been advanced for how to define "localities" for purposes of higher-dimensional information theory and automata theory. Some involve looking at larger and larger "blocks" of the same basic shape (Eriksson and Lindgren 1987); others attempt to "scan over" paths in the lattice, reducing the problem to that of well-ordered series (Feldman 1998). Neither of these quite works (Feldman 1998), so here is another, which seems to.

10.3.1 Light Cones and Their Equivalence Classes

Let $\mathbf{x} = (\vec{x}, t)$ be a single cell at a single time, or a *point-instant.* We define the *past light-cone* of $\mathbf{x}$, denoted $\overleftarrow{\mathrm{L}}(\mathbf{x})$, as all other point-instants $(\vec{y}, s)$ such that $s \leq t$ and $||\vec{y} - \vec{x}|| < c(t-s)$, where c is a positive constant for the system, the maximum speed at which disturbances can propagate — the "speed of light," as it were.[1] We denote the random variable for the configuration in the past light cone by $\overleftarrow{L}$, and its realization by $\overleftarrow{l}$. The *future light-cone*, $\overrightarrow{\mathrm{L}}(\mathbf{x})$, is similarly defined as all those point-instants $(\vec{y}, s)$ such that $s > t$ and $||\vec{y} - \vec{x}|| < c(s-t)$. It is the set of point-instants at which changes at $\mathbf{x}$ might have an effect. (See Figure 10.1 for a schematic.)

As with classical stochastic processes, any function on past light-cones partitions the set of them into equivalence classes, and so assigns them to effective states. The obvious analogs of the definitions and lemmas about effective states for time-series all apply straight-forwardly, since none of the information-theoretic arguments we made rely on having a well-ordered sequence as the input to the effective-state function η. We wish to point out, however, a subtlety in the appropriate definition of prescience.

Definition 35 (Local Effective States) *Any partition* $\boldsymbol{\mathcal{R}}$ *of* $\overleftarrow{\mathbf{L}}$ *is an* effective state class*; a cell* $\rho \in \boldsymbol{\mathcal{R}}$ *is an* effective state. *When the current past light-cone* $\overleftarrow{l}$ *is included in the set* ρ*, we will speak of the process being in state* ρ *at that point-instant. Thus, we define a function* η *from past light-cones to effective states:*

$$\eta : \overleftarrow{\mathbf{L}} \mapsto \boldsymbol{\mathcal{R}} \ . \tag{10.2}$$

A specific individual past light-cone $\overleftarrow{l} \in \overleftarrow{\mathbf{L}}$ *maps to a specific state* $\rho \in \boldsymbol{\mathcal{R}}$*; the random variable* $\overleftarrow{S}$ *for the past maps to the random variable* $\mathcal{R}$ *for the effective states.*

Remark. We have used the same notation for local effective states as for purely temporal effective states; we trust this will not cause confusion, as we shall only be dealing with the local versions from now on.

When we wish to refer to an arbitrary, finite spatio-temporal region, we shall write K. The random variable for the configuration in K is K.

Lemma 27 (Old Country Lemma) *For any local effective state, and any finite region* $\mathrm{K} \subset \overrightarrow{\mathrm{L}}$*,* $H[K|\mathcal{R}] \geq H[K|\overleftarrow{L}]$.

Proof: Entirely analogous to the time-series case.

Definition 36 (Local Prescience) *A local effective state is prescient iff, for any finite space-time region* $\mathrm{K} \subset \overrightarrow{\mathrm{L}}$*,* $H[K|\widehat{\mathcal{R}}] = H[K|\overleftarrow{L}]$.

Remark. As with the definition of prescience in the case of classical stochastic processes, this definition avoids having to invoke an entropy which may well be infinite, namely $H[\overrightarrow{L} \mid \overleftarrow{S}]$.

[1] By this definition, a point-instant is in its own past light-cone. This is a slight departure from the standard usage in relativity, essentially to accommodate discrete time.

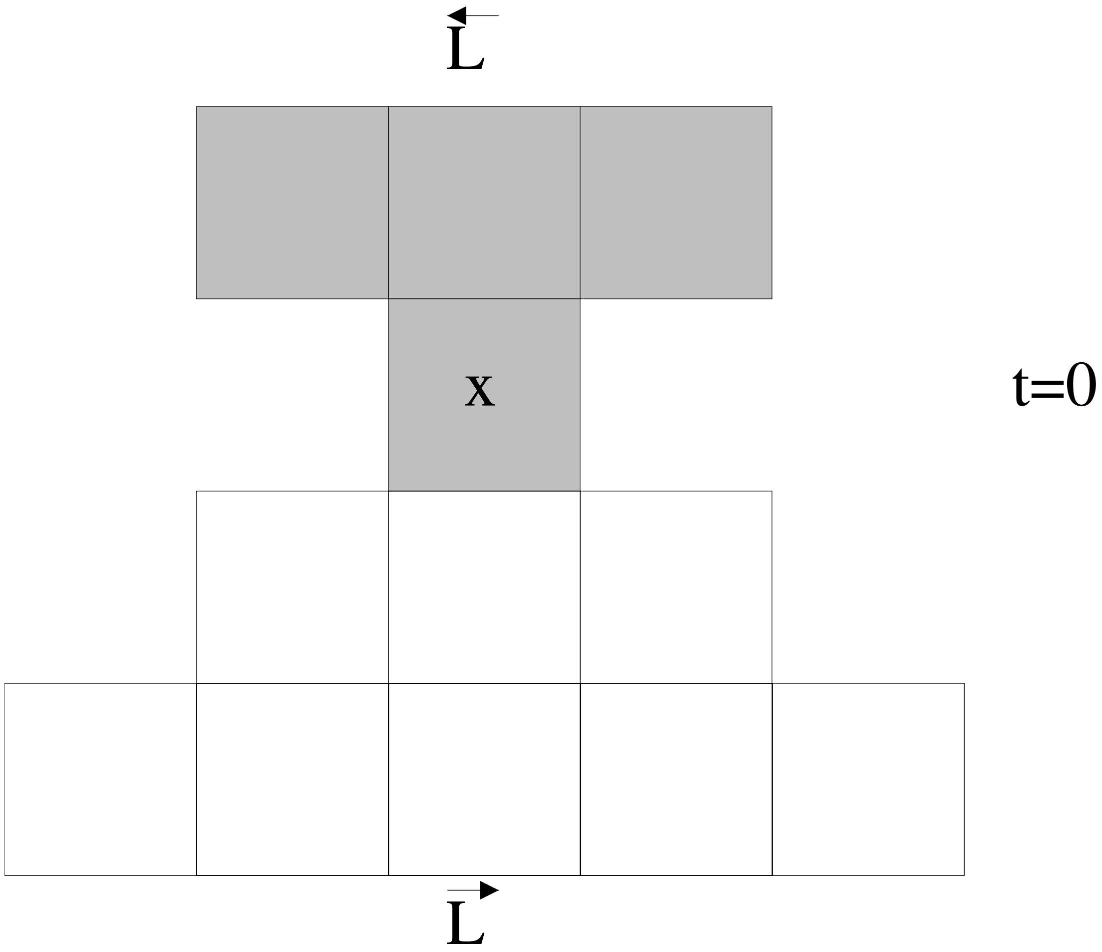

Figure 10.1: Schematic of the light-cones of a single point-instant, $\mathbf{x}$. Following convention for CAs, time runs vertically downward, and the spatial coordinate(s) are horizontal. The grey squares denote $\overleftarrow{L}(\mathbf{x})$, the past light-cone of $\mathbf{x}$. The white squares are its future light-cone, $\overrightarrow{L}(\mathbf{x})$. Note that we include $\mathbf{x}$ within its own past light-cone, resulting in a slight asymmetry between the two cones.

Lemma 28 (Local Prescience and Conditional Independence) *For any set of effective states* $\boldsymbol{\mathcal{R}}$*, if* $\overrightarrow{L} \perp\!\!\!\perp \overleftarrow{L} | \mathcal{R}$*, then* $\boldsymbol{\mathcal{R}}$ *is prescient.*

Proof: By Lemma 37, since $\mathcal{R} = \eta(\overleftarrow{L})$, the conditional independence implies that $\mathrm{P}(\overrightarrow{L} \mid \overleftarrow{L} = \overleftarrow{l}) = \mathrm{P}(\overrightarrow{L} | \mathcal{R} = \eta(\overleftarrow{l}))$ for all $\overleftarrow{l}$. Therefore this is true for any K as well. Hence the entropy of K conditional on $\overleftarrow{L}$ is equal to its entropy conditional on $\mathcal{R}$. But this is the definition of local prescience. QED.

Definition 37 (Local Excess Entropy) *The local excess entropy at* $\mathbf{x}$ *is*

$$\mathbf{E}^{loc}(\mathbf{x}) \quad = \quad \log_2 \frac{\mathrm{P}(\overrightarrow{L} = \overrightarrow{l}(\mathbf{x}), \overleftarrow{L} = \overleftarrow{l}(\mathbf{x}))}{\mathrm{P}(\overrightarrow{L} = \overrightarrow{l}(\mathbf{x}))\mathrm{P}(\overleftarrow{L} = \overleftarrow{l}(\mathbf{x}))} \tag{10.3}$$

Definition 38 (Excess Entropy Density) *The excess entropy density,* $\overline{\mathbf{E}^{loc}}$*, is*

$$\overline{\mathbf{E}^{loc}} \quad \equiv \quad I(\overleftarrow{L}; \overrightarrow{L}) \ . \tag{10.4}$$

It is the expectation value of the local excess entropy, $\mathbf{E}^{loc}(\mathbf{x})$*.*

Remark 1. The proof of the assertion follows directly from the definition of mutual information.

Remark 2. Both $\mathbf{E}^{loc}(\mathbf{x})$ and $\overline{\mathbf{E}^{loc}}$ can vary over time, and generally do in non-stationary processes. Note that $\overline{\mathbf{E}^{loc}}$ is not the average of $\mathbf{E}^{loc}(\mathbf{x})$ over the lattice in any particular realization of the process. Rather, it is the average over the ensemble of all realizations. If the two averages coincide on large lattices, then the process has a kind of spatial ergodicity. (Cf. the notion of "broken ergodicity" in spin glasses (Fischer and Hertz 1988; Palmer 1989).)

Definition 39 (Local Statistical Complexity) *For a set of effective states* $\boldsymbol{\mathcal{R}}$*, the local statistical complexity a* $\mathbf{x}$*, written* ${C_\mu}^{loc}(\boldsymbol{\mathcal{R}}, \mathbf{x})$ *is*

$$ {C_\mu}^{loc}(\boldsymbol{\mathcal{R}}, \mathbf{x}) \quad \equiv \quad -\log_2 \mathrm{P}(\mathcal{R}(\mathbf{x}) = \eta(\overleftarrow{l}(\mathbf{x}))) \tag{10.5}$$

Definition 40 (Statistical Complexity Density) *The statistical complexity density of a set of local effective states:*

$$\overline{{C_\mu}^{loc}}(\boldsymbol{\mathcal{R}}) \quad \equiv \quad H[\mathcal{R}] \ . \tag{10.6}$$

It is the expectation value of the local statistical complexity.

Remark. See the remarks on the excess entropy density.

10.3.2 The Local Causal States

We adapt the definition of causal states to the use of light cones in the obvious way.

Definition 41 (Local Causal State) *The local causal state at* $\mathbf{x}$*, written* $\mathcal{L}(\mathbf{x})$*, is the set of all past light-cones whose conditional distribution of future light-cones is the same as that of the past light-cone at* $\mathbf{x}$*. That is,*

$$\epsilon(\overleftarrow{l}(\mathbf{x})) \quad = \quad \left\{ \overleftarrow{l}' \middle| \mathrm{P}(K = k | \overleftarrow{L} = \overleftarrow{l}') = \mathrm{P}(K = k | \overleftarrow{L} = \overleftarrow{l}(\mathbf{x})) \right. \tag{10.7}$$

$$\left. \forall \mathrm{K} \subset \overrightarrow{\mathrm{L}}, \forall k \right\} \tag{10.8}$$

Lemma 29 ***(Conditional Independence of Past and Future Light-Cones)*** *The past and future light-cones are independent given the local causal state.*

Proof: By the construction of the local causal state,

$$\mathrm{P}(\overrightarrow{L}=\overrightarrow{l} \mid \overleftarrow{L}=\overleftarrow{l}) = \mathrm{P}(\overrightarrow{L}=\overrightarrow{l} \mid \mathcal{L}=\lambda) \tag{10.9}$$

But $\mathcal{L} = \epsilon(\overleftarrow{L})$, so, by Lemma 37, $\overrightarrow{L} \perp\!\!\!\perp \overleftarrow{L} \mid \mathcal{L}$. QED.

Theorem 18 (Prescience of Local Causal States) *The local causal states are prescient:* $\forall \mathrm{K}$,

$$H[K|\mathcal{L}] = H[K|\overleftarrow{L}] \tag{10.10}$$

Proof: Follows immediately from the combination of Lemmas 29 and 28. QED.

Lemma 30 (Local Refinement Lemma) *If $\widehat{\boldsymbol{\mathcal{R}}}$ is a prescient set of local effective states, then $\widehat{\boldsymbol{\mathcal{R}}}$ is a refinement of $\mathcal{L}$, and there is a function h such that $\mathcal{L} = h(\widehat{\mathcal{R}})$ almost always.*

Proof: Identical to the global lemma.

Theorem 19 (Minimality of the Local Causal States) *For any prescient rival set of states, $\widehat{\boldsymbol{\mathcal{R}}}$,*

$$\overline{C_\mu{}^{loc}}(\widehat{\boldsymbol{\mathcal{R}}}) \geq \overline{C_\mu{}^{loc}}(\mathcal{L}) \tag{10.11}$$

Proof: Identical to the global theorem.

Theorem 20 (Uniqueness of the Local Causal States) *If $\widehat{\boldsymbol{\mathcal{R}}}$ is prescient, and $\overline{C_\mu{}^{loc}}(\widehat{\boldsymbol{\mathcal{R}}}) = \overline{C_\mu{}^{loc}}(\mathcal{L})$, then there is a function g such that $\widehat{\mathcal{R}} = g(\mathcal{L})$ almost always.*

Proof: Identical to the global theorem, substituting $\overline{C_\mu{}^{loc}}$ for C_μ.

Theorem 21 (Local Statistical Complexity and Excess Entropy)

$$\mathbf{E}^{loc}(\mathbf{x}) \leq C_\mu{}^{loc}(\mathcal{L},\mathbf{x}) \tag{10.12}$$

Proof: Recall the definition of $\mathbf{E}^{loc}(\mathbf{x})$:

$$\begin{aligned}
\mathbf{E}^{loc}(\mathbf{x}) &= \log_2 \frac{\mathrm{P}(\overrightarrow{L}=\overrightarrow{l}(\mathbf{x}), \overleftarrow{L}=\overleftarrow{l}(\mathbf{x}))}{\mathrm{P}(\overrightarrow{L}=\overrightarrow{l}(\mathbf{x}))\mathrm{P}(\overleftarrow{L}=\overleftarrow{l}(\mathbf{x}))} && (10.13)\\
&= \log_2 \frac{\mathrm{P}(\overrightarrow{L}=\overrightarrow{l}(\mathbf{x})| \overleftarrow{L}=\overleftarrow{l}(\mathbf{x}))}{\mathrm{P}(\overrightarrow{L}=\overrightarrow{l}(\mathbf{x}))} && (10.14)\\
&= \log_2 \frac{\mathrm{P}(\overrightarrow{L}=\overrightarrow{l}(\mathbf{x})|\mathcal{L}=\epsilon(\overleftarrow{l}(\mathbf{x})))}{\mathrm{P}(\overrightarrow{L}=\overrightarrow{l}(\mathbf{x}))} && (10.15)\\
&= \log_2 \frac{\mathrm{P}(\overrightarrow{L}=\overrightarrow{l}(\mathbf{x}), \mathcal{L}=\epsilon(\overleftarrow{l}(\mathbf{x})))}{\mathrm{P}(\overrightarrow{L}=\overrightarrow{l}(\mathbf{x}))\mathrm{P}(\mathcal{L}=\epsilon(\overleftarrow{l}(\mathbf{x})))} && (10.16)\\
&= \log_2 \frac{\mathrm{P}(\mathcal{L}=\epsilon(\overleftarrow{l}(\mathbf{x}))| \overrightarrow{L}=\overrightarrow{l}(\mathbf{x}))}{\mathrm{P}(\mathcal{L}=\epsilon(\overleftarrow{l}(\mathbf{x})))} && (10.17)\\
&= \log_2 \mathrm{P}(\mathcal{L}=\epsilon(\overleftarrow{l}(\mathbf{x}))| \overrightarrow{L}=\overrightarrow{l}(\mathbf{x})) - \log_2 \mathrm{P}(\mathcal{L}=\epsilon(\overleftarrow{l}(\mathbf{x}))) && (10.18)\\
&= C_\mu{}^{loc}(\mathcal{L},\mathbf{x}) + \log_2 \mathrm{P}(\mathcal{L}=\epsilon(\overleftarrow{l}(\mathbf{x}))| \overrightarrow{L}=\overrightarrow{l}(\mathbf{x})) && (10.19)\\
&\leq C_\mu{}^{loc}(\mathcal{L},\mathbf{x}) && (10.20)
\end{aligned}$$

Theorem 22 (Bounds of Excess (Densities)) $\overline{\mathbf{E}^{loc}} \leq \overline{C_{\mu}{}^{loc}}(\mathcal{L})$.

Proof: Identical to the global theorem. Alternately, simply take the expectation value of both sides of the previous theorem.

10.3.3 Composition of the Global State from Local States

10.3.3.1 Extended or Patch Causal States

Rather than considering the past and future light-cones of a single point-instant, we can consider those of a *patch* of points at the same time. It will be convenient to only consider connected patches. The past and future cones of the patch are simply the unions of the cones of the patch's constituent cells. We denote the patch's past light-cone by $\overleftarrow{P}$, its future by $\overrightarrow{P}$. We define prescience and the patch causal state exactly as for the local case; the patch causal state is written $\mathcal{P}$.

Transparently, the patch causal states are prescient (for the patch future light-cone), minimal among the prescient patch states, and render the patch's future light cone conditionally independent of its past light cone. Moreover, Lemma 28 holds good for them, too; we shall make much use of this in what follows.

10.3.3.2 Composing Local States into Patch and Global States

Lemma 31 (Patch Composition Lemma) *The causal state of a patch at one time $\mathcal{P}$ is uniquely determined by the composition of all the local causal states within the patch at that time.*

Proof: We will show that the composition of local causal states within the patch is a prescient "effective state" *of the patch*, and then apply minimality.

Consider first a patch consisting of two (spatially) adjacent cells, $\mathbf{x}_1$ and $\mathbf{x}_2$. Define the following regions:

$$\begin{aligned} \overleftarrow{\mathrm{L}}_c &= \overleftarrow{\mathrm{L}}(\mathbf{x}_1) \cap \overleftarrow{\mathrm{L}}(\mathbf{x}_2) \\ \overleftarrow{\mathrm{L}}_1 &= \overleftarrow{\mathrm{L}}(\mathbf{x}_1) \setminus \overleftarrow{\mathrm{L}}_c \\ \overleftarrow{\mathrm{L}}_2 &= \overleftarrow{\mathrm{L}}(\mathbf{x}_2) \setminus \overleftarrow{\mathrm{L}}_c \end{aligned}$$

Thus $\overleftarrow{\mathrm{L}}(\mathbf{x}_1) = \overleftarrow{\mathrm{L}}_1 \cup \overleftarrow{\mathrm{L}}_c$, and likewise for $\overleftarrow{\mathrm{L}}(\mathbf{x}_2)$. Define $\overrightarrow{\mathrm{L}}_1$, $\overrightarrow{\mathrm{L}}_c$ and $\overrightarrow{\mathrm{L}}_2$ similarly. (See Figure 10.2 for a picture of these regions.) Now consider the configurations in these regions. We may draw a diagram of causal effects (Figure 10.3).

Lemma 29 tells us that every path from $\overleftarrow{L}_1$ or $\overleftarrow{L}_c$ to $\overrightarrow{L}_1$ must go through $\mathcal{L}_1$. By the very definition of light-cones, there cannot be a path linking $\overleftarrow{L}_2$ to $\overrightarrow{L}_1$. Therefore there cannot be a link from $\mathcal{L}_2$ to $\overrightarrow{L}_1$. (Such a link would in any case indicate that $\overrightarrow{L}_1$ had a dependence on $\overleftarrow{L}_c$ which was not mediated by $\mathcal{L}_1$, which is false.) All of this is true, *mutatis mutandis*, for $\overrightarrow{L}_2$ as well.

Now notice that every path from variables in the top row — the variables which collectively constitute $\overleftarrow{P}$ — to the variables in the bottom row — which collectively are $\overrightarrow{P}$ — must pass through either $\mathcal{L}_1$ or $\mathcal{L}_2$. The set $Z = \{\mathcal{L}_1, \mathcal{L}_2\}$ thus "blocks" those paths. In the terminology of graphical studies of causation, Z *d-separates* $\overleftarrow{P}$ and $\overrightarrow{P}$. But d-separation implies conditional independence (Pearl 2000, p. 18). Thus $\overleftarrow{P}$ and $\overrightarrow{P}$ are independent given the composition of $\mathcal{L}_1$ and $\mathcal{L}_2$. But that combination is a function of $\overleftarrow{P}$, so Lemma 28 applies, telling us that the composition of local states is prescient. Then Lemma 30 tells us that there is a function from the composition of local states to the patch causal state.

Now, the reader may verify that this argument would work if one of the two "cells" above was really itself a patch. That is, if we break a patch down into a single cell and a sub-patch, and we know their causal states, the causal state of the larger patch is fixed. Hence, by mathematical induction, if we know all the local causal states of the cells within a patch, we have fixed the patch causal state uniquely. QED.

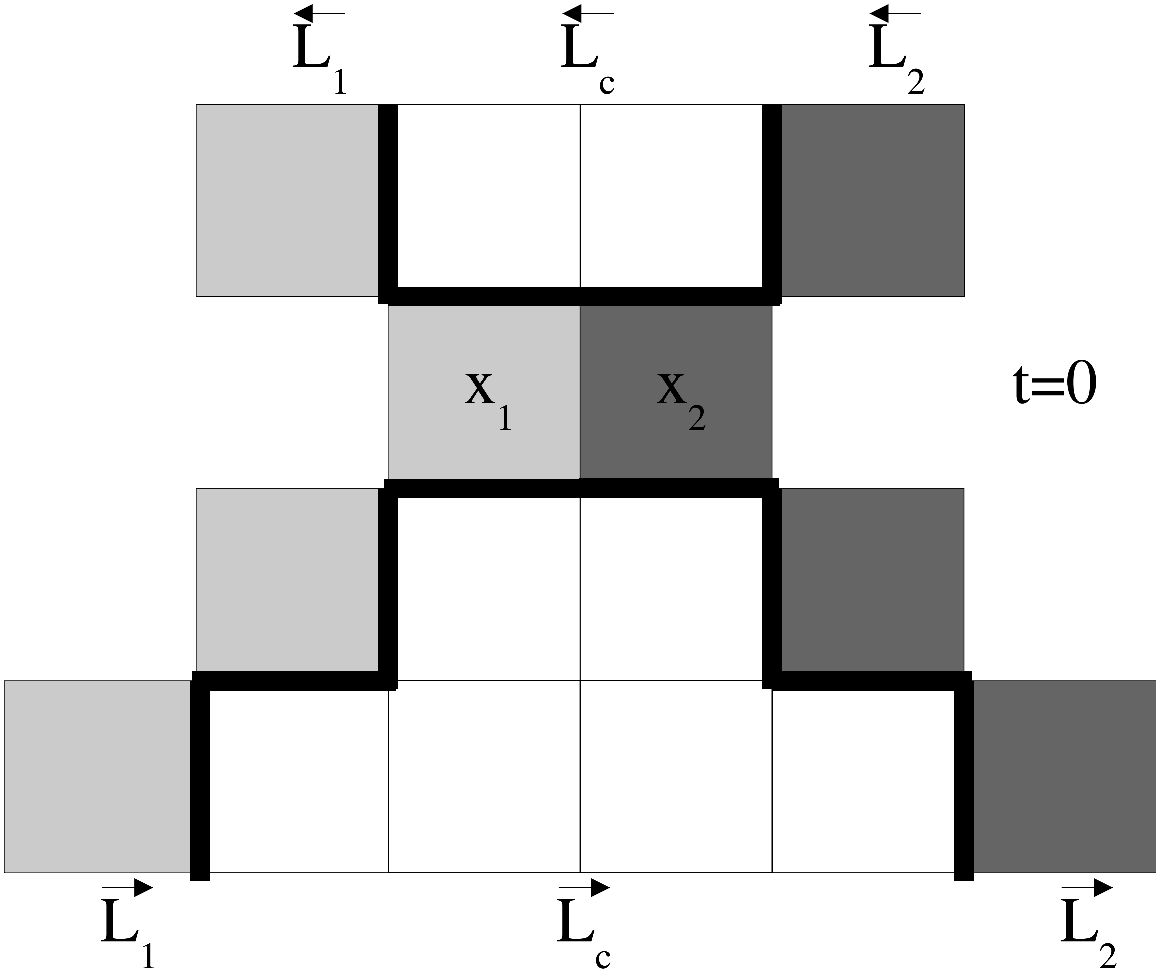

Figure 10.2: The space-time regions for a patch of two cells. point-instants which belong exclusively to the light-cones of the cell on the left ($\mathbf{x}_1$) are shaded light grey; those which belong exclusively to the light-cones of the other cell ($\mathbf{x}_2$) are shaded dark grey. The areas of overlap ($\overleftarrow{\mathrm{L}}_c$ and $\overrightarrow{\mathrm{L}}_c$) are white, with heavy borders. Note that, by the definition of light-cones, the configuration in $\overleftarrow{\mathrm{L}}_1$ can have no effect on that in $\overrightarrow{\mathrm{L}}_2$ or vice versa.

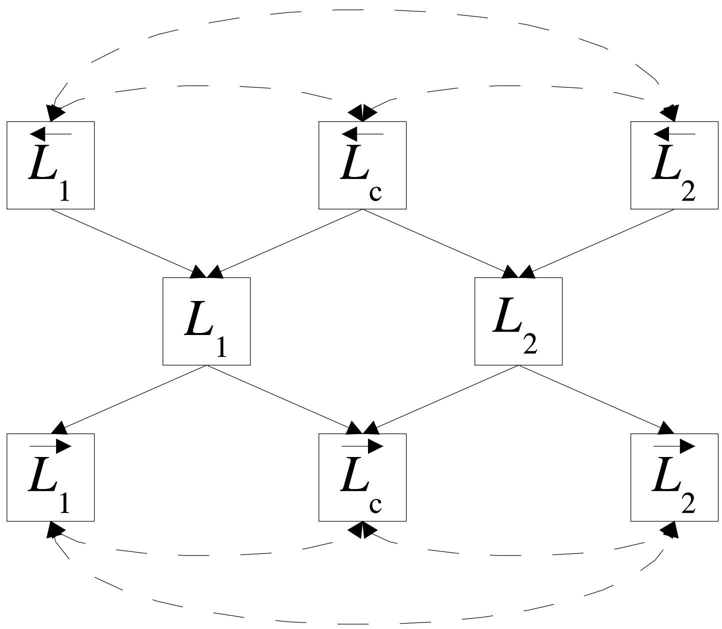

Figure 10.3: Diagram of causal effects for the two-cell patch. Arrows flow from causes to effects; the absence of a variable between two nodes indicates an absence of direct causal influence. Dashed lines indicate possible correlations.

Theorem 23 (Global Composition Theorem) *The global causal state at one time $\mathcal{G}$ is uniquely determined by the composition of all the local causal states at that time.*

Proof: Apply Lemma 31 to the "patch" of the entire lattice. The proof of the lemma goes through, because it in no way depends on the size or the shape of the past, or even on the patch being finite in extent. Since the patch causal state for this patch is identical with the global causal state, it follows that the latter is uniquely fixed by the composition of the local causal states at all points on the lattice. QED.

Remark 1. We have thus shown that the global causal state can be decomposed into local causal states, as we have defined them, *without* losing its global properties or indeed any information.

Remark 2. Conceivably, we could define local causal states with reference not to light-cones but to regions of other shapes, and some of our formal results would still hold. It is not clear, however, whether we could then recover the global causal state through composition, since the properties of light-cones per se played an important role in our proof. This topic deserves further investigation.

10.4 Connections Among Local States; the ϵ-Machine

Just as in the case of time series or of transducers, causal states succeed each other, with transitions between states being accompanied by observational symbols. In the case of spatial processes, there are two complications.

First, transitions can be made, not just forward in time, but also laterally, from a cell to any of its neighbors. Thus we will need to label transitions, not just by their probabilities and their symbols, but also by their directions.

The second complication concerns those symbols. For time series, the symbols on the transitions were simply that, symbols from the alphabet $\mathcal{A}$. For transducers, we needed to label transitions by two symbols, one from the input alphabet $\mathcal{A}$ and one from the output alphabet $\mathcal{B}$. In both cases, the labels consisted of all the new observations, of all the new data, observed in the course of the transition.[2] The new data obtained from a transition in a spatial process consists of the values of point-instants which are in the past light cone of the new point-instant, but were inaccessible from the old one.

More formally, define the *fringe* of the past light cone of $\mathbf{x}$, when moving to the neighboring point-instant $\mathbf{x}'$, as all point-instants in $\overleftarrow{\mathrm{L}}(\mathbf{x}')$ that were not in $\overleftarrow{\mathrm{L}}(\mathbf{x})$. (See Figures 10.2 and 10.4.) Then the new data consists of the configuration in the fringe.

That we should consider the new data to be the fringe configurations is not at all obvious (at least not to me); therefore it needs to be proved. The proof will take the form of showing that the old local causal state, plus the fringe, determines the new local causal state. There are two cases to consider moving forward in time, and moving sideways in space.

10.4.1 Temporal Transitions

We want to move forward in time one step, while staying in place. Call the point-instant we start at $\mathbf{x}$, and its successor $\mathbf{x}^+$. A little thought will convince you that the whole of the new future light cone is contained inside the old future light cone, and vice versa for the past cones. So let's define

$$
\begin{aligned}
\overleftarrow{\mathrm{L}}_n &= \overleftarrow{\mathrm{L}}(\mathbf{x}^+) \setminus \overleftarrow{\mathrm{L}}(\mathbf{x}) \\
\overrightarrow{\mathrm{L}}_o &= \overrightarrow{\mathrm{L}}(\mathbf{x}) \setminus \overrightarrow{\mathrm{L}}(\mathbf{x}^+) ;
\end{aligned}
$$

$\overleftarrow{\mathrm{L}}_n$ is the fringe. (See Figure 10.4 for a picture of these regions.)

Lemma 32 (Determinism of Temporal Transitions) *The local causal state at $\mathbf{x}^+$ is a function of the local causal state at $\mathbf{x}$ and the time-forward fringe $\overleftarrow{L}_n$.*

[2] Cf. the idea of the "innovation" in filtering theory (Bucy 1994).

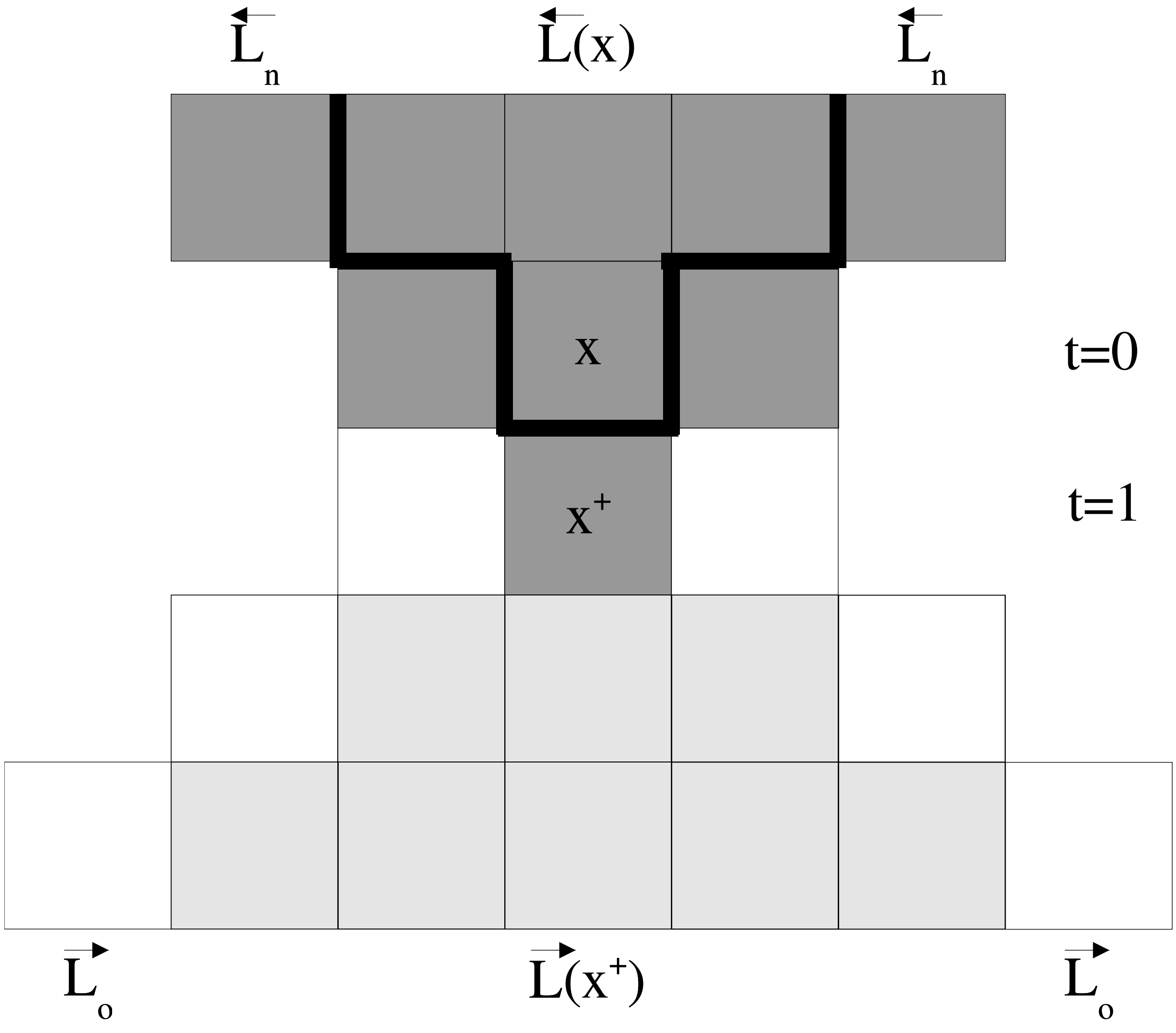

Figure 10.4: Space-time regions for the time-forward transition from $\mathbf{x}$ to $\mathbf{x}^+$. Region inside heavy borders: $\overleftarrow{\mathrm{L}}(\mathbf{x})$, the past light-cone of $\mathbf{x}$. Dark grey: $\overleftarrow{\mathrm{L}}(\mathbf{x}^+)$, the past light-cone of $\mathbf{x}^+$. Light grey: $\overrightarrow{\mathrm{L}}(\mathbf{x}^+)$, the future light-cone of $\mathbf{x}^+$. White: $\overrightarrow{\mathrm{L}}(\mathbf{x})$, the future light-cone of $\mathbf{x}$. Note that $\overleftarrow{\mathrm{L}}(\mathbf{x}) \subset \overleftarrow{\mathrm{L}}(\mathbf{x}^+)$ and $\overrightarrow{\mathrm{L}}(\mathbf{x}^+) \subset \overrightarrow{\mathrm{L}}(\mathbf{x})$. $\overleftarrow{\mathrm{L}}_n$ consists of the dark grey cells outside of the heavy lines; $\overrightarrow{\mathrm{L}}_o$ consists of white cells (not the light grey ones).

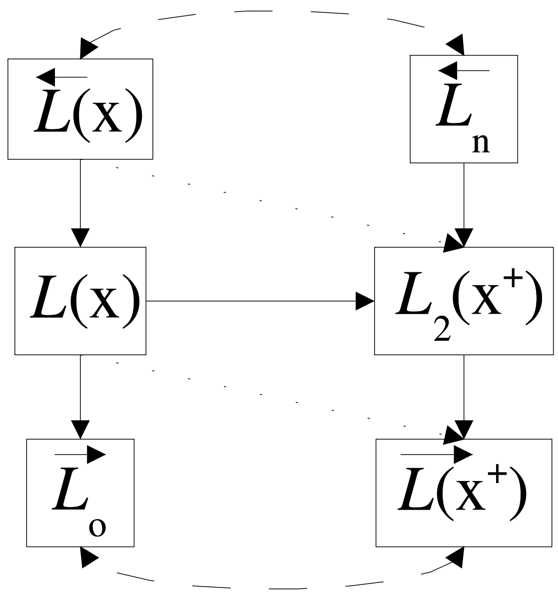

Figure 10.5: Diagram of causal effects for the configurational variables involved in a time-forward transition. Dashed arrows indicate possible non-causal correlations. Dotted arrows indicate indirect effects, mediated by paths consisting entirely of solid arrows.

Proof. Start by drawing the diagram of causal effects (Figure 10.5).

$\overrightarrow{L}_o$ and $\overrightarrow{L}(\mathbf{x}^+)$ jointly constitute $\overrightarrow{L}(\mathbf{x})$, so there must be paths from $\mathcal{L}(\mathbf{x})$ to both of them. Now, $\mathcal{L}(\mathbf{x}^+)$ renders $\overleftarrow{L}(\mathbf{x}^+)$ and $\overrightarrow{L}(\mathbf{x}^+)$ conditionally independent. Hence it should d-separate them in the graph of effects. But $\overleftarrow{L}(\mathbf{x})$ is part of $\overleftarrow{L}(\mathbf{x}^+)$ and has a direct path to $\mathcal{L}(\mathbf{x})$. This means that there cannot be a direct path from $\mathcal{L}(\mathbf{x})$ to $\overrightarrow{L}(\mathbf{x}^+)$; rather, the causation must be mediated by $\mathcal{L}(\mathbf{x}^+)$. (We indicate this in the graph by a dotted arrow from $\mathcal{L}(\mathbf{x})$ to $\overrightarrow{L}(\mathbf{x}^+)$. Similarly, $\overleftarrow{L}(\mathbf{x})$ certainly helps determine $\mathcal{L}(\mathbf{x}^+)$, but it need not do so directly. In fact, it cannot: $\mathcal{L}(\mathbf{x})$ must d-separate $\overleftarrow{L}(\mathbf{x})$ and $\overrightarrow{L}(\mathbf{x})$, i.e., must d-separate $\overleftarrow{L}(\mathbf{x})$ from $\overrightarrow{L}(\mathbf{x}^+)$ and $\overrightarrow{L}_o$. Hence the influence of $\overleftarrow{L}(\mathbf{x})$ on $\mathcal{L}(\mathbf{x}^+)$ must run through $\mathcal{L}(\mathbf{x})$. (We indicate this, too, by a dotted arrow from $\overleftarrow{L}(\mathbf{x})$ to $\mathcal{L}(\mathbf{x}^+)$.)

Now it is clear that the combination of $\mathcal{L}(\mathbf{x})$ and $\overleftarrow{L}_n$ d-separates $\overleftarrow{L}(\mathbf{x}^+)$ from $\overrightarrow{L}(\mathbf{x}^+)$, and hence makes them conditionally independent. But now the usual combination of Lemmas 28 and 30 tell us that there's a function from $\mathcal{L}(\mathbf{x}), \overleftarrow{L}_n$ to $\mathcal{L}(\mathbf{x}^+)$. QED.

10.4.2 Spatial Transitions

Lemma 33 (Determinism of Spatial Transitions) *Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be simultaneous, neighboring point-instants. Then $\mathcal{L}(\mathbf{x}_2)$ is a function of $\mathcal{L}(\mathbf{x}_1)$ and the fringe in the direction from $\mathbf{x}_1$ to $\mathbf{x}_2$, $\overleftarrow{L}_2$.*

Here the breakdown of the past and future light-cone regions is the same as when we saw how to compose patch causal states out of local causal states in Section 10.3.3, as is the diagram of causal effects; we'll use the corresponding terminology, too. (See Figs.10.2 and 10.3, respectively.) What we hope to show here is that conditioning on the combination of $\mathcal{L}_1$ and $\overleftarrow{L}_2$ makes $\overrightarrow{L}(\mathbf{x}_2)$ independent of $\overleftarrow{L}_2$ and $\overleftarrow{L}_c$. Unfortunately, as the reader may verify by inspecting the diagram, our conditional variables no longer d-separate the other variables (since they have an unblocked connection through $\mathcal{L}_2$). All is not lost, however: d-separation implies conditional independence, but not conversely.

Abbreviate the pair of variables $\left\{\mathcal{L}_1, \overleftarrow{L}_2\right\}$ by Z. Now, $\mathcal{L}_2$ is a (deterministic) function of $\overleftarrow{L}_c$ and $\overleftarrow{L}_2$. Hence it is also a function of Z and $\overleftarrow{L}_c$. Thus $\mathrm{P}(\overrightarrow{L}_2 | \mathcal{L}_2, Z, \overleftarrow{L}_c) = \mathrm{P}(\overrightarrow{L}_2 | Z, \overleftarrow{L}_c)$. But this tells us that

$$\overrightarrow{L}_2 \perp\!\!\!\perp \mathcal{L}_2 | Z, \overleftarrow{L}_c \tag{10.21}$$

From d-separation, we also have

$$\overrightarrow{L}_2 \perp\!\!\!\perp \overleftarrow{L}_c | Z, \mathcal{L}_2 \tag{10.22}$$

Applying Eq. A.33,

$$\overrightarrow{L}_2 \perp\!\!\!\perp \mathcal{L}_2, \overleftarrow{L}_c | Z \tag{10.23}$$

Applying Eq. A.34,

$$\overrightarrow{L}_2 \perp\!\!\!\perp \overleftarrow{L}_c | Z \tag{10.24}$$

Since $Z = Z, \overleftarrow{L}_2$,

$$\overrightarrow{L}_2 \perp\!\!\!\perp \overleftarrow{L}_c | Z, \overleftarrow{L}_2 \tag{10.25}$$

The following conditional independence is odd-looking, but trivially true:

$$\overrightarrow{L}_2 \perp\!\!\!\perp \overleftarrow{L}_2 | Z \tag{10.26}$$

And it, along with Eq. A.35, gives us

$$\overrightarrow{L}_2 \perp\!\!\!\perp \overleftarrow{L}_c, \overleftarrow{L}_2 | Z \tag{10.27}$$

A similar train of reasoning holds for $\overrightarrow{L}_c$. Thus, the entire future light cone of $\mathbf{x}_2$ is independent of that point-instant's past light cone, given $\mathcal{L}_1$ and $\overleftarrow{L}_2$. This tells us that $\left\{\mathcal{L}_1, \overleftarrow{L}_2\right\}$ is prescient for $\overrightarrow{L}(\mathbf{x}_2)$, hence $\mathcal{L}_2$ is a function of it.

QED.

10.4.3 Arbitrary Transitions

Lemma 34 (Determinism along Paths) *Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be two point-instants, such that $\mathbf{x}_2$ is at the same time or later than $\mathbf{x}_1$. Let Γ be a spatio-temporal path connecting the two point-instants, arbitrary except that it can never go backwards in time. Let F_Γ be the succession of fringes encountered along Γ. Then $\mathcal{L}(\mathbf{x}_2)$ is a function of $\mathcal{L}(\mathbf{x}_1)$, Γ and F_Γ,*

$$\mathcal{L}(\mathbf{x}_2) = g(\mathcal{L}(\mathbf{x}_1), \Gamma, F_\Gamma) \tag{10.28}$$

for some function g.

Proof. Apply Lemma 32 or 33 at each step of Γ. QED.

Lemma 35 (Path-Independence of Transitions) *Let* $\mathbf{x}_1$ *and* $\mathbf{x}_2$ *be two point-instants as in the previous lemma, and let* Γ_1, Γ_2 *be two paths connecting them, and* F_{Γ_1} *and* F_{Γ_2} *their fringes, all as in the previous lemma. Then the state at* $\mathbf{x}_2$ *is independent of which path was taken to reach it,*

$$g(\mathcal{L}(\mathbf{x}_1), \Gamma_1, F_{\Gamma_1}) \quad = \quad g(\mathcal{L}(\mathbf{x}_1), \Gamma_2, F_{\Gamma_2}) \ . \tag{10.29}$$

Proof. Suppose otherwise. Then either the state we get by going along Γ_1 is wrong, i.e., isn't $\mathcal{L}(\mathbf{x}_2)$, or the state we get by going along Γ_2 is wrong, or both are.

$$\mathcal{L}(\mathbf{x}_2) \neq g(\mathcal{L}(\mathbf{x}_1), \Gamma_1, F_{\Gamma_1}) \quad \vee \quad \mathcal{L}(\mathbf{x}_2) \neq g(\mathcal{L}(\mathbf{x}_1), \Gamma_2, F_{\Gamma_2}) \tag{10.30}$$

$$\overrightarrow{L}(\mathbf{x}_2) \not\perp\!\!\!\perp \overleftarrow{L}(\mathbf{x}_2) | \mathcal{L}(\mathbf{x}_1), \Gamma_1, F_{\Gamma_1} \quad \vee \quad \overrightarrow{L}(\mathbf{x}_2) \not\perp\!\!\!\perp \overleftarrow{L}(\mathbf{x}_2) | \mathcal{L}(\mathbf{x}_1), \Gamma_2, F_{\Gamma_2} \tag{10.31}$$

$$\neg(\overrightarrow{L}(\mathbf{x}_2) \perp\!\!\!\perp \overleftarrow{L}(\mathbf{x}_2) | \mathcal{L}(\mathbf{x}_1), \Gamma_1, F_{\Gamma_1} \quad \wedge \quad \overrightarrow{L}(\mathbf{x}_2) \perp\!\!\!\perp \overleftarrow{L}(\mathbf{x}_2) | \mathcal{L}(\mathbf{x}_1), \Gamma_2, F_{\Gamma_2}) \tag{10.32}$$

But, by the path determinism lemma 34, $\overrightarrow{L}(\mathbf{x}_2) \perp\!\!\!\perp \overleftarrow{L}(\mathbf{x}_2) | \mathcal{L}(\mathbf{x}_1), \Gamma_1, F_{\Gamma_1}$ and $\overrightarrow{L}(\mathbf{x}_2) \perp\!\!\!\perp \overleftarrow{L}(\mathbf{x}_2) | \mathcal{L}(\mathbf{x}_1), \Gamma_2, F_{\Gamma_2}$. Hence transitions must be path-independent. QED.

10.4.4 The Labeled Transition Probabilities

Just as in the case of time series, we can construct labeled transition probability functions, $\mathbf{T}$, which take as arguments the current state, the direction of transition and the fringe seen on transition (regarded as a string over $\mathcal{A}$), and returns the state arrived at. For time series, we also include the probability of emitting that symbol and so of arriving at that state. That is licit, because each state is associated with a unique morph, and so a unique distribution for the next symbol. Here the local causal states have morphs, but only over their future light cones, which include little if any of the relevant fringes. So it's not immediately obvious that those transition probabilities are well-defined, stationary objects.

In practice, every spatial system we have examined *does* have well-defined transition probabilities between its local states. I am led to the following.

Definition 42 (Causal Parents of a Local Causal State) *The causal parents of the local causal state at* $\mathbf{x}$ *are the causal states at all point-instants which are one time-step before* $\mathbf{x}$ *and inside its past light-cone:*

$$A(\mathbf{x}) \quad \equiv \quad \{\mathcal{L}(\vec{y}, t-1) \,|\, ||\vec{y} - \vec{x}|| \leq c\} \tag{10.33}$$

Lemma 36 (Screening-off of Past Light Cone by Causal Parents) *The local causal state at a point-instant,* $\mathcal{L}(\mathbf{x})$*, is independent of the configuration in its past light cone, given its causal parents:*

$$\mathcal{L}(\mathbf{x}) \perp\!\!\!\perp \overleftarrow{L}(\mathbf{x}) | A(\mathbf{x}) \tag{10.34}$$

Proof. $\mathbf{x}$ is in the intersection of the future light cones of all the cells in the patch at $t-1$. Hence, by the arguments given in the proof of the composition theorem, it is affected by the local states of all those cells, *and by no others*. In particular, previous values of the configuration in $\overleftarrow{L}(\mathbf{x})$ have no direct effect; all causation is mediated through those cells. Hence, by d-separation, $\mathcal{L}(\mathbf{x})$ is independent of $\overleftarrow{L}(\mathbf{x})$. QED.

Theorem 24 ***(Temporal Markov Property for Local Causal States)*** *The local causal state at a point-instant,* $\mathcal{L}(\mathbf{x})$*, is independent of the local causal states of point-instants in its past light cone, given its causal parents.*

Proof. By the previous lemma, $\mathcal{L}(\mathbf{x})$ is conditionally independent of $\overleftarrow{L}(\mathbf{x})$. But the local causal states in its past light cone are a function of $\overleftarrow{L}(\mathbf{x})$. Hence by Lemma A.38, $\mathcal{L}(\mathbf{x})$ is also independent of those local states. QED.

Comforting though that is, we would really like a stronger Markov property, namely the following.

Conjecture 2 ***(Markov Field Property for Local Causal States)*** *The local causal states form a Markov random field in space-time.*

Argument for why this is plausible. We've seen that, temporally speaking, a Markov property holds: given a patch of cells at one time, they are independent of their past light cone, given their causal parents. What we need to add for the Markov field property is that, if we condition on *present* neighbors of the patch, as well as the parents of the patch, then we get independence of the states of all point-instants at time t or earlier. It's plausible that the simultaneous neighbors are informative, since they are also causal descendants of causal parents of the patch. But if we consider any more remote cell at time t, its last common causal ancestor with any cell in the patch must have been before the immediate parents of the patch, and the effects of any such local causal state are screened off by the parents.

10.4.5 ϵ-Machine Reconstruction

We have designed and implemented an algorithm for the reconstruction of local causal states from simulation data for lattice dynamical systems. (It could, in principle, be used on experimental data as well.) The procedure is as follows. We examine the empirical joint distribution for configurations of past light-cones of depth L and future light-cones of depth K. That is, we gather statistics on the joint distribution of past and future cones. If we have seen N light-cone pairs, then we estimate $\mathrm{P}(\overleftarrow{L}^L = \overleftarrow{l}^L, \overrightarrow{L}^K = \overrightarrow{l}^K)$ by

$$\hat{\mathrm{P}}_N(\overleftarrow{L}^L = \overleftarrow{l}^L, \overrightarrow{L}^K = \overrightarrow{l}^K) = \frac{\nu(\overleftarrow{l}^L, \overrightarrow{l}^K)}{N} \tag{10.35}$$

where $\nu(\overleftarrow{l}^L, \overrightarrow{l}^K)$ simply counts the number of times we have seen that pair of light-cones. This is known as the *joint empirical distribution.* Then we calculate the empirical conditional distribution of futures for each past, $\hat{\mathrm{P}}_N(\overrightarrow{L}^K = \overrightarrow{l}^K | \overleftarrow{L}^L = \overleftarrow{l}^L)$, for each $\overleftarrow{l}^L$ and each $\overrightarrow{l}^K$, as

$$\hat{\mathrm{P}}_N(\overrightarrow{L}^K = \overrightarrow{l}^K | \overleftarrow{L}^L = \overleftarrow{l}^L) = \frac{\hat{\mathrm{P}}_N(\overleftarrow{L}^L = \overleftarrow{l}^L, \overrightarrow{L}^K = \overrightarrow{l}^K)}{\hat{\mathrm{P}}_N(\overleftarrow{L}^L = \overleftarrow{l}^L)} \tag{10.36}$$

where the denominator is obtained, in the normal way, by summing the joint distribution over all future light-cone configurations. Finally, we group the past light-cones into classes or effective states. We list the pasts in some order, and start by assigning the first past to the first class. Now consider past $\overleftarrow{l}^L$, which is at least the second past in our order. We go through all the existing classes in order, and check whether $\overleftarrow{l}^L$ is compatible with all the pasts in that class. Compatibility between two pasts is defined by the Euclidean distance between their empirical conditional distributions of futures being less than a pre-chosen tolerance parameter δ. That is, $\overleftarrow{l}^L$ and $\overleftarrow{l}^{L\prime}$ are compatible when

$$\sum_{\overrightarrow{l}^K} \left(\hat{\mathrm{P}}_N(\overrightarrow{L}^K = \overrightarrow{l}^K | \overleftarrow{L}^L = \overleftarrow{l}^L) - \hat{\mathrm{P}}_N(\overrightarrow{L}^K = \overrightarrow{l}^K | \overleftarrow{L}^L = \overleftarrow{l}^{L\prime}) \right)^2 \leq \delta \tag{10.37}$$

If $\overleftarrow{l}^L$ is compatible with all the pasts already in state i, it is compatible with state i. We add $\overleftarrow{l}^L$ to the first state in our enumeration with which it is compatible. If it is not compatible with any existing state, we create a new one for it. This procedure is repeated until all pasts have been assigned to states.

Clearly, compatibility between histories is not a true equivalence relation (it is not transitive), so the order in which pasts are checked for membership in states, and in which states are created, does matter. This can be effectively randomized, however, and in any case does not effect the reliability of the procedure, which we now address.

10.4.5.1 Reliability of Reconstruction

Suppose that L and K are sufficiently large that they suffice to distinguish the causal states, i.e., that if we had the exact distribution over past and future light-cones of those respective depths, and partitioned according to the definition of local causal states, we would recover the true local causal states. Then conditioning on pasts of depth L makes futures independent of the further past. Indeed, every time we examine the future of a certain past configuration of depth L, it is independent of all the other futures of that same configuration. Thus, the strong law of large numbers tells us that our estimate of the conditional probability of any future configuration of depth K will almost surely converge on the true probability:

$$\left| \mathrm{P}(\overrightarrow{L}^K = \overrightarrow{l}^K \mid \overleftarrow{L}^L = \overleftarrow{l}^L) - \hat{\mathrm{P}}_N(\overrightarrow{L}^K = \overrightarrow{l}^K \mid \overleftarrow{L}^L = \overleftarrow{l}^L) \right| \quad \overrightarrow{N \to \infty} \quad 0 \tag{10.38}$$

Hence the squared errors also converge to zero. Since there are only a finite number of such configurations, it follows that the sum of the sum of such squared errors will also converge to zero, with probability one.

Now, under the assumptions we have made about being able to recover the causal states from examining only cones of finite depth, for any process there will be only a finite number of distinct conditional distributions of future light-cones. Hence there will be a strictly positive δ_0, such that all the conditional distributions have a total-variation distance of at least δ_0 from each other. Pick a $\delta \leq \delta_0/2$ for our tolerance parameter. Then two pasts will be wrongly assigned to the same state only if one of their empirical distributions differs from its true distribution by at least $\delta_0/2$. But the probability of this happening goes to zero as $N \to \infty$, as we've seen. Hence, asymptotically, the probability of any two light-cones being wrongly assigned to the same class goes to zero. Similarly, if two light-cones should be placed together, the probability that their empirical distributions differ by enough to separate them also goes to zero. Thus, asymptotically, all light-cones are assigned to the correct equivalence class, provided $\delta \leq \delta_0/2$. Indeed, all that we really need is for δ to be below $\delta_0/2$ for sufficiently large N, so it suffices that $\delta \to 0$.

To summarize: the spatial reconstruction algorithm given here is consistent, PAC, and reliable, in the same senses as the state-splitting algorithm for time series (Chapter 5). All this, recall, is under the assumption that past and future light-cones of depth L and K are sufficient to recover the causal states. If we can let $L, K \to \infty$, then the algorithm is consistent for all spatial processes with some finite speed-of-light.

Any attempt to reconstruct causal states from empirical data is necessarily an approximation. Other algorithms exist in the literature, all of which deliver the appropriate causal states in the limit of infinite data and infinite history-length (Chapter 5; Crutchfield and Young 1990; Hanson 1993; Perry and Binder 1999). That is, like the present algorithm, they are consistent estimators if given infinite histories. (If every causal state can be unambiguously identified after only a finite history, then they are simply consistent.) A number of these algorithms (Chapter 5; Perry and Binder 1999) could be adapted to light-cones; others are restricted to working with time-series. We hope to address the important question of the error statistics (Mayo 1996) of these reconstruction algorithms in future work; our conjectures about the convergence rate of the state-splitting algorithm (Chapter 5) are relevant here, too.

10.5 Emergent Structures

In Chapter 9, I claimed that domains and particles were emergent structures. Here I will show how to define domains, particles, and other common spatial emergent structures in terms of the ϵ-machine. Later, in Section 11.2.2, I'll consider the idea that emergent structures can generically be defined as sub-machines of the ϵ-machine (I'll also give a definition of "emergent").

Definition 43 (Domain) *A* domain phase *is a sub-machine of the ϵ-machine which is strongly connected for transitions in all spatial directions. A* domain *is a strongly-connected set of domain phases.*

Definition 44 (Defect) *Any point-instant in a configuration which is reached on a transition that does not belong to any domain is in a* defect.

Definition 45 (d-Brane) *A* d-brane *is a defect machine which is a strongly-connected graph in time (possibly with a translation) and $n > d \geq 1$ spatial directions. If $d = 1$, then it is a em line.*

Definition 46 (Particle) *A* particle *is a defect machine which is a strongly-connected graph in time (possibly composed with a translation), but has bounded extent in all spatial directions.*

In every case which has been checked, the particles and domains identified by hand, through spatial computational mechanics, exactly correspond to sub-machines identified in this way. That is, spatial causal states are also spatio-temporal local causal states.

Conjecture 3 (Domain and Particle Emergence) *Suppose that a spatial process has domains, branes and particles. Derive a new process from it by applying a filter which maps each domain to a distinct value, each brane-type to a distinct value, and each particle to a distinct value. Then that derived process is emergent.*

It is hard to see how domain-filtering could *lower* the efficiency of prediction, but no proof saying otherwise exists.

10.6 Examples

10.6.1 ECA Rule 54: Domain and Particles

Let us return to rule 54. Running the ϵ-machine reconstruction algorithm for spatial processes on it identifies eight equivalence classes of past light-cones; we need only go back to a depth of 2 to do this. (See Figure 10.6.) Furthermore, we can get the spatial transition structure (Figure 10.7) and the temporal structure (Figure 10.8). Compare Figure 10.7 with Figure 9.3. The two structures are manifestly the same, both in states and in transitions. (It is easy to work out the correspondence between the fringes in the former and the scanning symbols in the later.) But the domain filter was assembled by hand, and the new ϵ-machine was automatically constructed.

Observe that the probability of staying within a domain phase, once entered, is much higher than that of leaving it, so that grouping the domain states together (by filtering on the domain) will improve the efficiency of prediction. That is, the domain-filtered process is emergent.

10.6.2 ECA Rule 110: Domains and Particles

Recall that ECA 110 has one primary domain, and a large number of minor, less stable ones. The primary domain, Λ^0, has spatial period 14 and temporal period 7, so that each point in the domain follows one of two distinct time courses.

All of this was discovered by hand, and pretty painful hands at that. Here is the result of running the spatial ϵ-machine reconstruction algorithm on rule 110, starting from random initial conditions, with a past and future depth set equal to 3.

There are 31 causal states, each occupied by only a single past light cone. (See Table 10.1.) The spatial structure is given by Figure 10.9, for left-to-right transitions. The Λ^0 domain can easily be seen, as the chain of 14 states on the left. It is fairly easy to find other closed chains of states, but these are not the other domains. This becomes evident when we look at the temporal structure (Figure 10.10). Λ^0 has two sub-components, corresponding to the two time courses available to a site within the domain. Most of the chains of states outside Λ^0 are not preserved under time-evolution, therefore they are not domains.

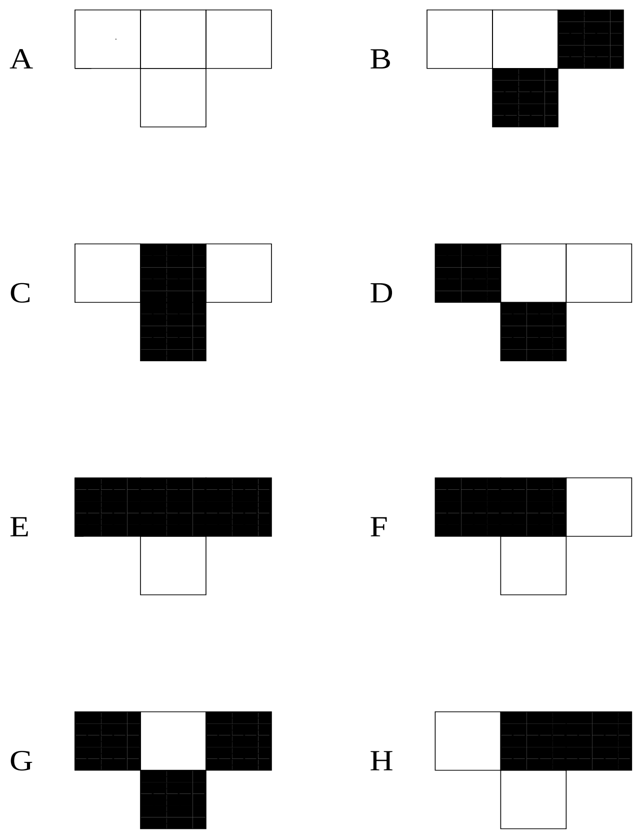

Figure 10.6: The light-cones for the local causal states of rule 54

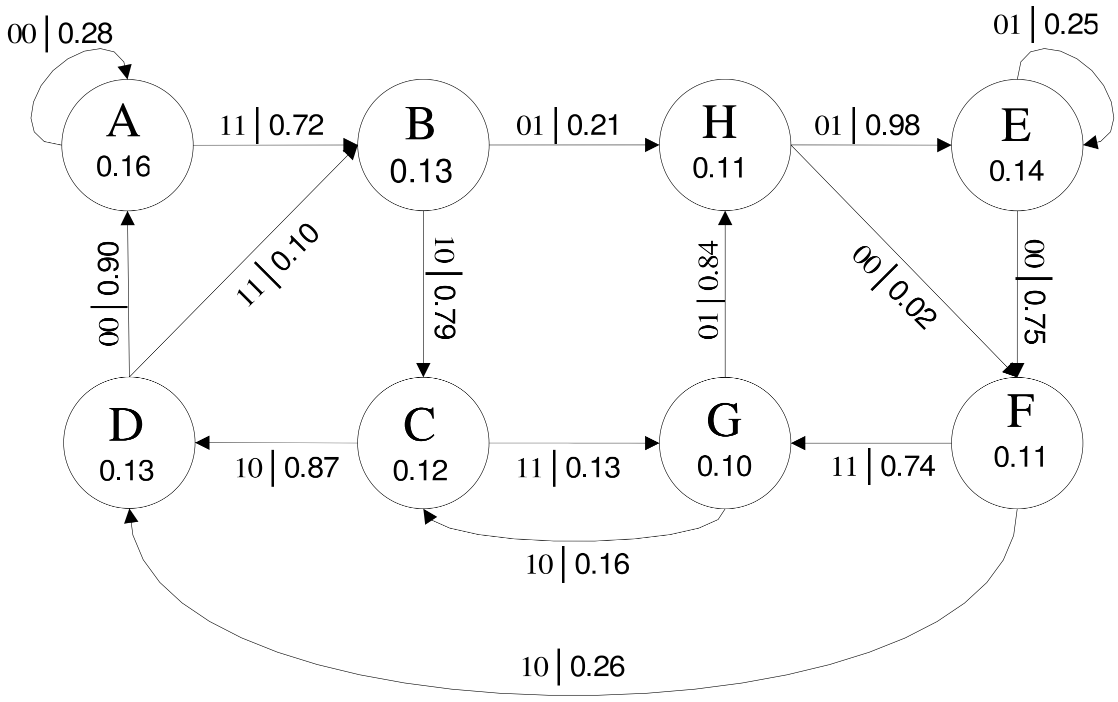

Figure 10.7: Left-to-right spatial transitions for rule 54. The fringe symbols should be read backwards in time. That is, "10" means that the cell to the right of the current one, at the current time, has a value of 1, and the cell to the right of that, one time-step previously, has a value of 0.

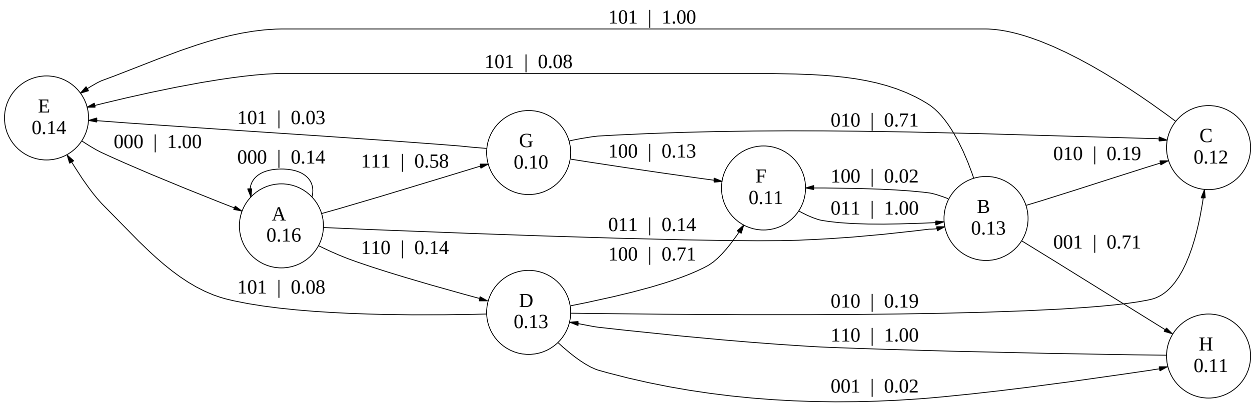

Figure 10.8: Temporal transitions for rule 54. The fringe should be read left-to-right. That is, "011" means that the cell to the left of the present cell has a value of 0 at the present time, that the present cell will have a value of 1 at the next time, and that the cell to the right of the present cell has a value of 1 currently.

State Label	Past Light Cone	State	Past	State	Past	State	Past	State	Past
A	11110 001 1	B	11100 010 1	C	11000 100 0	D	10001 001 1	E	00010 011 1
F	00100 110 1	G	01001 101 1	H	10011 011 1	I	00110 111 0	J	01101 111 0
K	11011 111 0	L	10011 110 1	M	01111 100 0	N	11111 000 0	0	00000 000 0
16	10000 000 0	34	01000 100 0	87	10100 110 1	103	01100 110 1	157	10010 011 1
190	11010 111 0	222	10110 111 0	235	01110 101 1	265	00001 001 1	315	11001 101 1
334	00101 111 0	350	10101 111 0	381	11101 011 1	397	00011 011 1	430	01011 111 0
455	00111 110 1								

Table 10.1: Local causal states of rule 110. Each state contains only a single light cone. The states which compose the primary domain are given alphabetical labels, in accordance with previous studies of the rule. The others are labeled by numbers assigned to them by the reconstruction algorithm.

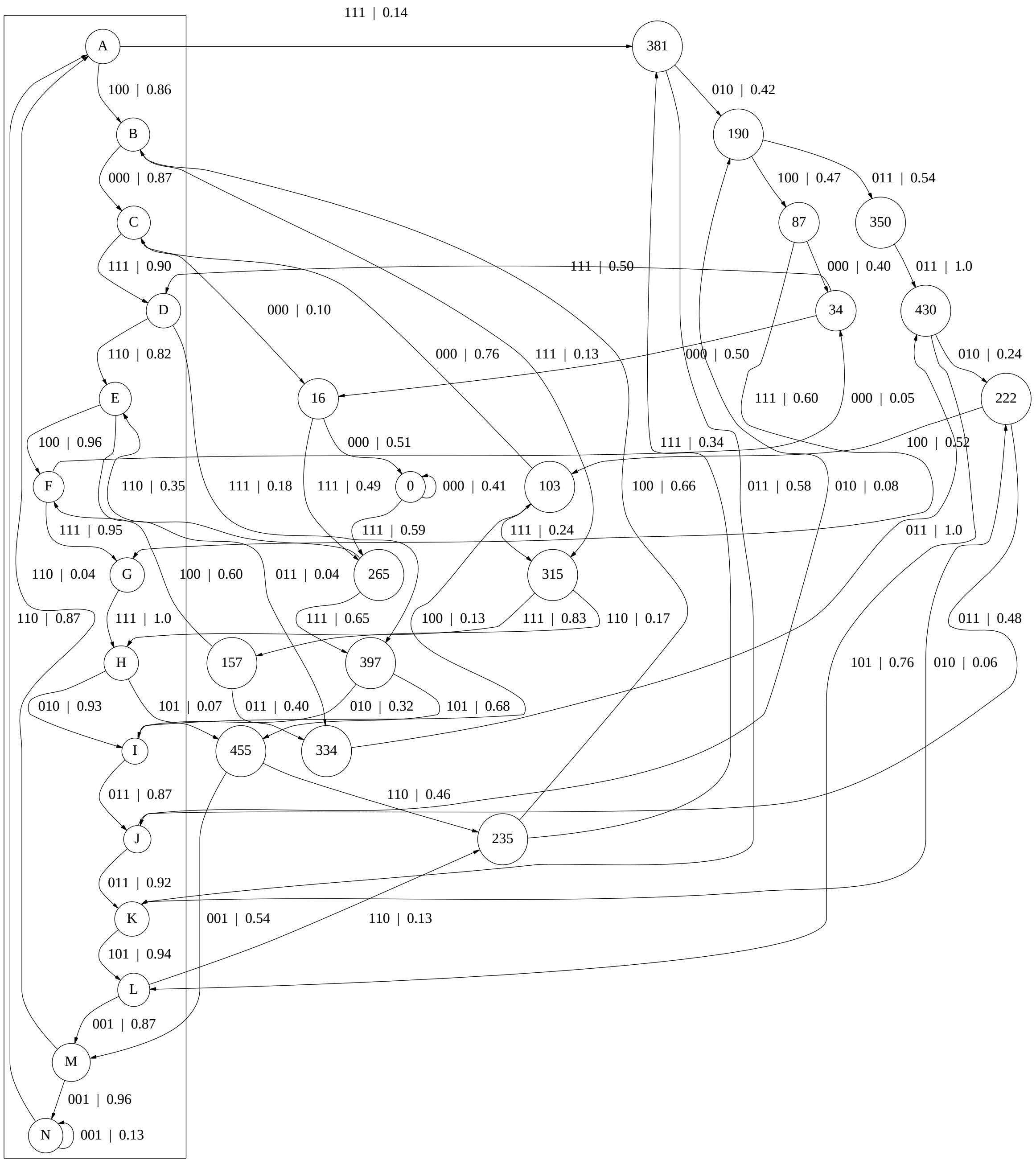

Figure 10.9: Spatial part of the ϵ-machine for rule 110, left-to-right transitions. States belonging to the Λ^0 domain, the rule's "true vacuum," are enclosed in a box on the left. The fringe labels are read left-to-right, as in Figure 10.7.

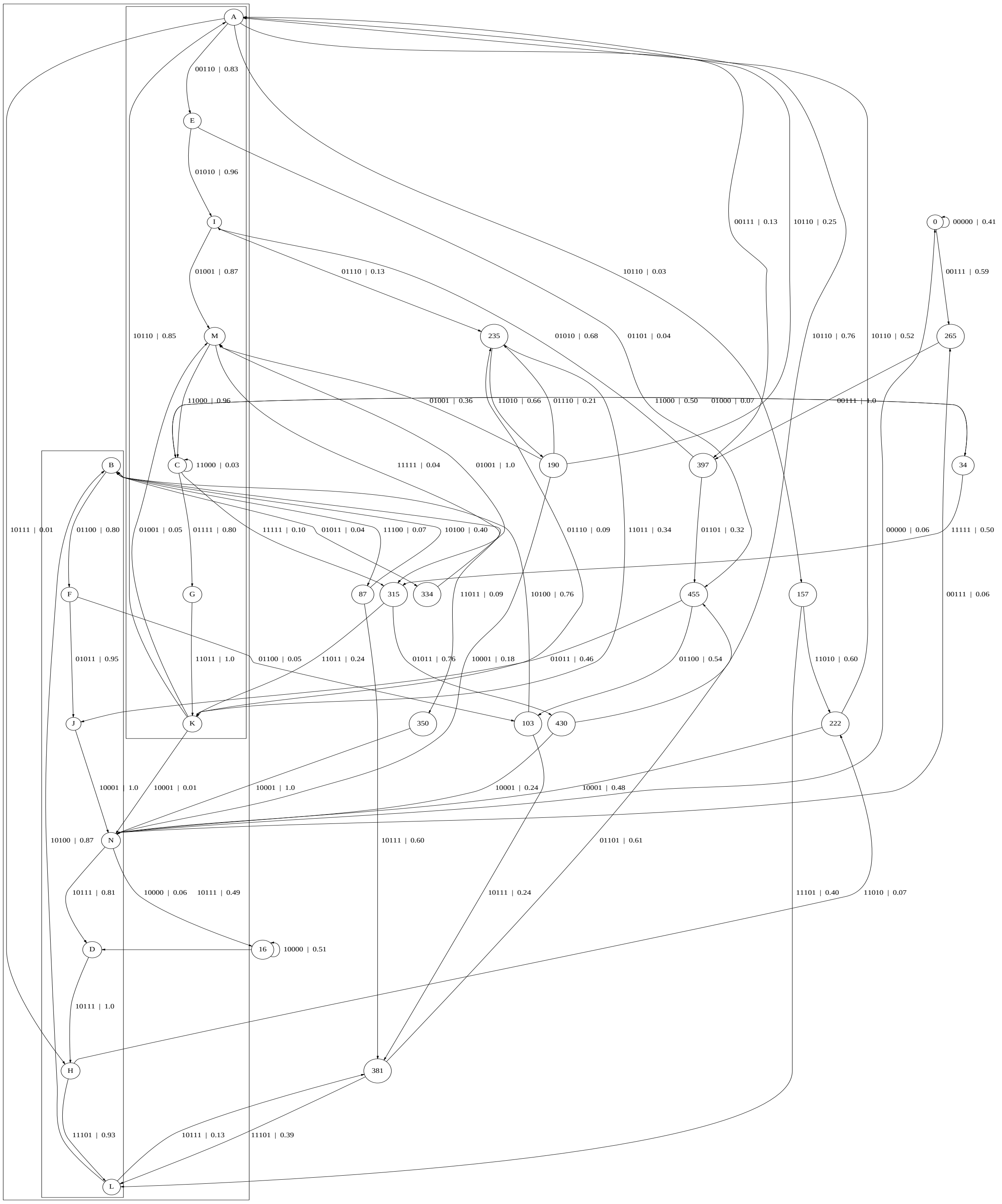

Figure 10.10: Temporal part of the ϵ-machine for rule 110. The large box on the left encloses the domain Λ^0, the sub-boxes its two phases. The fringe labels are read as in Figure 10.8.

10.7 Summary

The main point of this chapter has been to show how to define local causal states for well-behaved spatial processes. By using light cones for our histories and futures, we can assign a causal state to each point-instant, and these are the unique minimal optimal predictors, as we'd hope; indeed, almost all of the familiar, comforting properties of causal states in purely temporal processes carry over. We can also compose these local causal states into the causal states for extended regions, even the entire lattice, thereby recovering the global causal state. We can *define* the most common sorts of emergent structure (domain, particle, etc.) in terms of the ϵ-machine connecting the local causal states, and so put all the results of Chapter 9 on a much firmer footing.

If the ideas in this chapter are the right way of thinking about patterns and complexity in spatial processes, then it really doesn't make much sense to try to work out the complexity of (say) static *images*, or of individual configurations. Complexity, on this view, must be a function of the *process* which generates configurations (cf. Lloyd and Pagels 1988); we need movies, not snapshots. But this should not be distressing to physicists: we, of all people, should be very suspicious if pattern appeared *without* a causal history to back it up.

I want to close this chapter by suggesting two area for future work.

One has to do with irregular lattices. I have assumed throughout that space is a regular lattice, that every cell's connections look like every other cells. But a lot of the math developed here doesn't *depend* on that. Space could be an arbitrary graph, for instance, and we could still define past and future light cones, and so local causal states — presumably a different set of causal states for each point with a distinct set of connections. I *think* the Composition Theorem would still hold, but I don't really know. It would be interesting to find out, since there are many important dynamical systems which live on spatial lattices, but not on regular graphs. In particular, many technical, biological and social networks seem to be "small world" networks, and it would be nice to understand how they work, and particularly nice to understand their emergent structures (if any) (Watts 1999; Shalizi 2000). We might also look at these networks as so many interconnected transducers, along the lines of Chapter 7 — which may be formally equivalent! But the transducer view may be more valuable when we do not know what the network is to start with — and, after all, a network, in these sense, *is* a pattern of causal interaction, so it ought to be something we infer. So this's one area where the theory could use some work.

Another, much more abstract one, goes back to the composition theorems, which say that global properties can be built up out of local ones. This is reminiscent of a common sort of result in algebraic geometry, where a global invariant is algebraically composed out of objects which represent local properties, as the polynomial equivalent of a knot is constructed from terms representing its various parts. Since we can define algebraic structures which are related to the causal states, as in Appendix B.2, we might be able to give an algebraic version of the composition theorem, which actually stated what the composition function was, rather than just proving that it must exist. This could also open the way to a more direct and algebraic characterization of things like domains. But this is all, alas, quite speculative.[3]

[3]Thanks to Mitchell Porter for suggesting this idea.

Chapter 11

Conclusion

11.1 What Has Been Accomplished

The main line of this book has been the exposition of computational mechanics for increasingly sophisticated processes.

We started, in Chapter 3 with memoryless transducers, where we constructed causal states as equivalence classes of inputs — two inputs are causally equivalent when they have the same conditional distribution of outputs. The causal states, we saw, were optimal predictors, and the unique minimal optimal predictors. Since they are both unique and minimal, we could identify the complexity of the process with the complexity of the causal states, defined as the amount of information needed to specify the current causal state.

The rest of the book showed how the same basic idea of causal state works with different sorts of process: time series, transducers and CAs. The time series chapter introduced the idea of assigning a distinct causal state to each moment of time and of connecting them together through an ϵ-machine. The ϵ-machine's internal transitions are deterministic (in the automata-theory sense) and minimally stochastic. This work revisits the core of computational mechanics (Crutchfield and Young 1989; Crutchfield 1994a) with more rigor and new techniques of proof, which lead to some new results, such as the minimal stochasticity of the ϵ-machine, and the uniqueness of the causal states. Chapter 7 introduced the computational mechanics of interacting time series. Chapter 9, following a long tradition of spatial computational mechanics (Hanson and Crutchfield 1992; Crutchfield and Hanson 1993b; Hanson and Crutchfield 1997; Feldman and Crutchfield 1998a), assigns a causal state to each point in one-dimensional space, effectively treating the spatial coordinate as Chapter 4 treated time. Finally, Chapter 10 went beyond the older temporal and spatial computational mechanics, to a fully spatio-temporal version of the theory, with the advantage of working in any number of spatial dimensions.

Along the way, we saw how to estimate the causal states and the ϵ-machine from data, and how spatial computational mechanics lets us begin to get a handle on the computational powers of cellular automata. Now we'll see how to define emergence and self-organization.

11.2 Emergence

> Reductionism, roughly speaking, is the view that everything in this world is really something else, and that the something else is always in the end unedifying. So lucidly formulated, one can see that this is a luminously true and certain idea.
> — Ernest Gellner (1974, p. 107)

"Emergence" is an extremely slippery concept, used in an immense number of ways, generally with no attempt at precision whatsoever. It is also one with a decidedly unsavory history.[1] It is not at all clear that it is worth explicating. Nonetheless, let us try.

The strongest sense of "emergence" known to me, and also the oldest, is the following. A property of a composite object is emergent if it *cannot* be explained from the properties and interactions of the "lower level" entities composing the object. Now, we cannot know that *anything* is emergent in this sense. At best we can say that we don't yet have an explanation for a particular property, so for all we know it might be emergent. To call something emergent is therefore not to say anything about the property at all, but merely to make a confession of scientific and mathematical incompetence. (Epstein (1999) provides many examples of explanations of phenomena once taken to be exemplars of emergence, from chemical affinity on up.) Humility is all very well and good, but this is excessive.

A more moderate notion of emergence is also one which is more interesting, and potentially says something about the world, rather than our inability to interpret it. In this view, emergent properties[2] are ones which arise from the interactions of the lower-level entities, but which the latter themselves do not display. Standard examples of this sort of emergence are the laws of thermodynamics (individual molecules don't have a temperature or a pressure), or efficient allocation[3] of resources in various types of market economy (Debreu 1959; Lange and Taylor 1938; Simon 1996; Stiglitz 1994), or collective oscillations in ecosystems or economies (Krugman 1996).

A number of authors (see especially Simon 1996; Dennett 1991; Holland 1998 and Auyang 1998) have explored this sort of emergence, and while they have reached no definite conclusions or formalizations, there does seem to be a consensus on two points. First, the variables describing emergent properties must be fully determined by lower-level variables — must *supervene* on them, as the philosophers say (Kim 1998). Second, higher-level properties are worthy of being called emergent only if they are "easier to follow," or "simplify the description," or otherwise make our life, as creatures attempting to understand the world around us, at least a little easier.

Putting these two ideas together, we can actually *define* emergence. Crutchfield did so in his 1994 papers (1994a, 1994b), but I fear he was over-subtle, since very few people have picked up on it. The goal of this section is to present his views, with a few modest technical additions, in a crushingly explicit manner.

11.2.1 Emergent Processes

For the rest of this section, I'll write as though we were only dealing with time series, but everything applies, *mutatis mutandis*, to transducers (Chapter 7) as well. There are more subtle changes needed to deal with spatial processes (Chapter 10), which I'll mention as they arise. Let's start by fixing just how easy it is to predict a process.

Definition 47 (Efficiency of Prediction) *The* efficiency of prediction *of a process is the ratio between its excess entropy and its statistical complexity.*

$$e = \frac{\mathbf{E}}{C_\mu} . \tag{11.1}$$

It is clear from the Bounds of Excess Theorem (Theorem 10) that e is a real number between 0 and 1, just as an efficiency should be. We may think of it as the fraction of historical memory stored in the process which does "useful work" in the form of telling us about the future. It is straight-forward to check that, for any prescient state class $\widehat{\boldsymbol{\mathcal{R}}}$, $\mathbf{E}/C_\mu(\widehat{\boldsymbol{\mathcal{R}}}) \leq e$.

If $C_\mu = 0$, there are two possibilities. One is that the process is completely uniform and deterministic. The other is that it is IID. In neither case is any interesting prediction possible, so we set $e = 0$ in those

[1] For remarks on the association between the notion of emergence and obscurantism in biology and social science, see Epstein (1999). For the connections between holism and totalitarianism, see Popper (1945, 1960).

[2] Or emergent phenomena or behaviors or structures or what-not; all of these terms are used, if not interchangeably, then with an apparent conviction that they're close enough for government work.

[3] Is "efficient allocation" an emergent *property*, an emergent *phenomenon*, or an emergent *behavior*? Who can say?

cases[4]

For spatial processes, e is the ratio between the densities of the excess entropy and the statistical complexity, $\overline{\mathbf{E}^{loc}}/\overline{C_\mu{}^{loc}}$.

Definition 48 (Derived Process) *One process,* $\overleftrightarrow{S}'$, derives *from another,* $\overleftrightarrow{S}$*, iff* $S'_t = f(\overleftarrow{S}_t)$*, for some measurable function* f*.* $\overleftrightarrow{S}'$ *is called the* derived or filtered process*,* $\overleftrightarrow{S}$ *the* original, underlying or raw process

This definition is intended to capture the idea of "supervenience"; that is, one set of variables is on a "higher level" than another. We can think of f as a sort of filter applied to the original process, passing through only certain aspects of it.

For spatial processes, we keep the requirement that f depends only on the history, but we do not require that it be spatially local.

Definition 49 (Emergent Process) *A derived process is* emergent *if it has a greater predictive efficiency than the process it derives from. We then say the derived process* emerges from *the underlying process.*

Definition 50 (Intrinsic Emergence) *A process is* intrinsically emergent *if there exists another process which emerges from it.*

This formalizes the two intuitions we started with. And it is not trivial, because there are plenty of derived processes whose efficiency of prediction is the same or even lower than that of the process they derive from. Moreover, once we have chosen a new set of variables in which to describe a process (i.e. a filter f), whether the new process is emergent is simply a fact about the dynamics of the raw process. And so whether the underlying process is emergent is just a fact about its dynamics. Emergence is thus intrinsic and objective, and has nothing whatsoever to do with observers.

It may help to contrast this notion of emergence with what people attempt to accomplish with statistical regression. There the goal is to "explain" all of the variance in the output by accounting for the effects of all possible input variables. What we are attempting to do in looking for an emergent process, on the other hand, is to filter out everything we can — get rid of all the small-but-significant inputs — so as to simplify the relationship. We are not trying to explain everything we can measure; we are trying to find what's intrinsically important in our measurements. Emergence is anti-regression.[5]

11.2.2 Emergent Structures Are Sub-Machines of the ϵ-Machine

There is a sense in which the dynamics of a process are completely summarized by its ϵ-machine — so why can't we use it to build a filter? The following procedure, in fact, suggests itself. Divide the ϵ-machine into sub-machines, i.e., strongly connected components, and label them all. Find all the transitions between sub-machines, and give those labels too. Then apply the following filter: at each time-step, check whether the current causal state and the previous state were in the same sub-machine. If they were, output the label of that sub-machine. If they weren't, then the process has moved from one sub-machine to another; output the label of that transition.

If the sub-machines have been chosen appropriately, the process derived from this filter will be emergent, since knowing what sub-machine we are in will reduce statistical complexity without impairing predictive power, or at least not impair it more than is gained by simplification. In this case, we may call the sub-machines *emergent structures*. For instance, a loop in the ϵ-machine — a closed cycle of states — would generally make a good sub-machine, and a fine emergent structure. By extension, a (part of a) configuration generated by the states-and-transitions in a sub-machine is also an emergent structure. The domains and particles we saw when looking at spatial and spatio-temporal processes were all examples of ϵ-machine based filters and emergent structures.

[4]It's often tempting to imagine a family of processes where $\mathbf{E}$ and C_μ both tend to 0 in some limit, and to use L'Hopital's Rule to calculate the limiting value of e, but I haven't found a way to make that precise.

[5]Thanks to Scott Page for pointing out this connection.

11.2.3 Example: Thermodynamics and Statistical Mechanics

As I mentioned, many people (Crutchfield 1992; Sklar 1993; Holland 1998) claim that thermodynamic regularities are emergent phenomena, emerging out of microscopic statistical mechanics. Let's check whether this agrees with my definition, both as a sanity-check for the definition and an illustration of how it can be applied.

Consider everyone's favorite companion from introductory statistical mechanics, a box full of gas. To be more specific, consider a cubic centimeter of argon, which is conveniently spinless and monoatomic, at standard temperature and pressure. Using the well-known formula (Landau and Lifshitz 1980, sec. 43), the thermodynamic entropy is

$$S(N,T,V) \quad = \quad Nk_B(\log V/N + c_v \log k_B T + \zeta + c_v + 1) \tag{11.2}$$

where ζ is the "chemical constant" of the gas, given by the Sackur-Tetrode formula (Landau and Lifshitz 1980, sec. 45), $\zeta = \frac{3}{2}\log\frac{m}{2\pi\hbar^2}$ and of course $c_v = 3/2$. Argon has an atomic mass of just under 40. We are taking $P = 10^5\mathrm{Nm}^{-2}$, $T = 293\mathrm{K}$, $V = 10^{-6}\mathrm{m}^3$. Thus $N = 2.47 \cdot 10^{19}$ and

$$\begin{aligned} S(N,T,V) &= 6.3 \cdot 10^{-3}\mathrm{J/K} && (11.3)\\ &= 6.6 \cdot 10^{20} \text{ bits} , && (11.4) \end{aligned}$$

using the conversion factor $k_B \log 2 = 1$ bit. Now, at the micromechanical level (almost by definition) the dynamics of the gas are Markovian, so each microstate *is* a causal state. If we sample the gas at time intervals of (say) 10^{-9} seconds, we have a first-order Markov process. Then $\mathbf{E} = C_\mu - h_\mu$ (Feldman and Crutchfield 1998a), so we need to know h_μ to calculate the efficiency of prediction. As it happens, Gaspard (1998, ch. 0) estimates the entropy rate of one cubic centimeter of argon at standard temperature and pressure to be around $3.3 \cdot 10^{29}$ bits per second. The efficiency of prediction is thus about 0.5, taking a time-step of one nanosecond. If we use a much larger time-step, the predictive efficiency of the system is essentially zero, which reflects the fact that the gas is very rapidly mixing.

Now consider looking at the macroscopic variables; it will be convenient to only consider extensive ones, so let's use total energy, particle number and volume, rather than the traditional number, pressure and volume. (Recall that $E = Nc_v k_B T$.) Their mean values are, of course, $\overline{E} = .16$ joules, $\overline{N} = 2.5 \cdot 10^{19}$ and $\overline{V} = 10^{-6}\mathrm{m}^3$. All of them fluctuate with a Gaussian distribution, but let's consider just fluctuations in E. Define $a = E - \overline{E}$. By the Einstein fluctuation formula (Keizer 1987, ch. 2), the variance is

$$\sigma^2 \quad = \quad -k_B C^{-1} \tag{11.5}$$

where $C = \partial^2 S/\partial a^2$. Explicitly evaluating that, $C = -Nc_v k_B/\overline{E}^2 = -2.3 \cdot 10^{-2}\mathrm{J}^{-1}\mathrm{K}^{-1}$, and so $\sigma^2 = 6.1 \cdot 10^{-22}\mathrm{J}^2$.

Assume we are sensitive to measurements at absurdly small level of $\Delta E = 10^{-15}$ joules. Then the entropy of the macrovariable energy is

$$\begin{aligned} H[E] &= \frac{1}{2}\log_2 2\pi e \frac{\sigma^2}{\Delta E^2} && (11.6)\\ &\approx 33.28 \text{ bits} . && (11.7) \end{aligned}$$

(If we set ΔE to a much larger value, there *isn't* any noticeable uncertainty in the macrovariable!)

What of the dynamics? Suppose that the gas stays in the linear regime. Then deviations from equilibrium values are followed by (on average) exponential return to equilibrium, plus noise. The dynamics of the macrovariables, too, are Markovian. The relevant stochastic differential equation is (Keizer 1987, p. 68):

$$\frac{da}{dt} \quad = \quad LCa + f , \tag{11.8}$$

where L, the *phenomenological coefficient*, governs the mean rate of decay of fluctuations, and f is white noise, i.e., $\overline{f(t)} = 0$ and $\overline{f(t+\tau)f(t)} = 2k_B L\delta(\tau)$. Ignoring (as we did above) fluctuations and coupling in the other extensive variables, we get (Balian 1991, sec. 14.2.3) $L = \lambda T^2$, where λ is the heat conductivity. For argon at STP, $\lambda \approx 1.017 \cdot 10^{-9}$ watts per kelvin, so $L \approx 8.6 \cdot 10^{-5}$ watt-kelvins. If we solve Eq. 11.8, we find that the conditional distribution at time t is Gaussian, with a conditional variance given by (Keizer 1987, Eq. 1.8.12)

$$\sigma^2(t) = \sigma_0^2 \left(1 - e^{2LCt}\right) . \tag{11.9}$$

If we take our time-step to be one millisecond, $\sigma^2(10^{-3}\mathrm{s}) \approx 2.0 \cdot 10^{-9}\sigma_0^2$. The entropy of the conditional distribution, coarse-grained at the same level as before, is 4.4 bits, and this is the entropy rate per time-step (i.e. $h_\mu = 4.4 \cdot 10^3$ bits/second). So the efficiency of prediction is 0.87. If we used the same time-step of 10^{-9}s as before, the efficiency is indistinguishable from 1. Hence thermodynamics emerges from the statistical mechanics, and does so very strikingly, since almost all of the information needed at the statistical-mechanical level is simply irrelevant thermodynamically.

11.3 Self-Organization Defined

Recall from Section 4.1.1 that the theory of causal states and ϵ-machines requires only conditional stationarity, not strict stationarity. When the process we are dealing with is non-stationary, the distribution of its causal states changes over time, and so the statistical complexity is a function of time, and we ought to write it $C_\mu(t)$ (cf. Crutchfield 1992). Under what conditions will $C_\mu(t)$ be an increasing function of time?

Here is an example to serve as an intuition pump. Prepare an ensemble of copies of a process so that all the copies start in the same causal state. $C_\mu(t)$ is then $\log 1 = 0$. Informally, there is only one thing the system can do, so it is simple. Suppose, however, that some transitions lead from this initial state to other causal states, specifically to a chain of causal states of period p, and that these states are very unlikely to lead back to the original state. Then $C_\mu(t)$ will increase over time from 0 to $\sim \log_2 p$. That is to say, when a system spontaneously goes from uniform to periodic behavior (which is one of the canonical examples of self-organization), its statistical complexity increases.

What I want to propose, therefore, is that *an increase in statistical complexity is a necessary condition for self-organization.* While the fundamental causal architecture remains unchanged, the degree of organization — measured by the amount of information needed to place the process in a state within the architecture — is variable. (Cf. the "knowledge diffusion" of Crutchfield (1992).) In every case I can think of, where people are pretty well agreed that self-organization happens, it's also pretty manifest that the statistical complexity increases.

If we compare this criterion for self-organization with the definition of emergence in chapter 11.2, we see that self-organization increases complexity, while emergence, generally speaking, reduces it, or requires us to use it more effectively for prediction. At first glance, then, self-organization and emergence are incompatible, but this is too hasty. Self-organization is something a process does over time, like being stationary, or having a growing variance. Emergence is, primarily, a relation between two processes, one of which is derived from the other, like "has a smaller entropy rate than". By extension, a process has the property of emergence if any of its derived processes is emergent (comparable to "is a function of a Markov chain"). There is nothing contradictory in saying that a process is becoming more structurally complex, while at the same time saying that there is another description of the process which is always simpler than the raw data.

We can now make sense of the way so many authors have linked self-organization and emergence. When something self-organizes, it becomes more statistically complex, i.e., optimal prediction requires more information. A cognitively-limited observer (such as a human scientist) is therefore motivated to look for a new way of describing the process which has a higher predictive efficiency. That is, the desire to describe things simply makes us *look for* emergent behavior in self-organizing systems. (Imagine describing an excitable medium, not by saying where the spiral waves are centered and how their spirals curve, but by giving the complete field of molecular concentrations at each point.) Emergence without self-organization is definitely possible — for example, we've seen that thermodynamics emerges from statistical mechanics in a stationary (and so definitely non-self-organizing) system. I presume there can be self-organizing, non-emergent processes, though it *might* be that some constraint on possible ϵ-machines rules that out. Assuming, however, that self-organization does not imply emergence, then it is conceivable that there are processes which organize themselves into conditions so complex that no human being can grasp them. They would be so organized, in other words, that they would look very like noise. (Cf. Crutchfield and Feldman 2001a; Crutchfield and Feldman 2001b; Lem 1968/1983.) Emergence may be a pre-condition of *detectable* self-organization.

There is an obstacle blocking the way to simply defining self-organization as increasing statistical complexity. This is that a rise in C_μ does not distinguish self-organization from *getting organized by something else*[6]. We want, in Grassberger's (1986) phrase, "self-generated complexity," not any other sort. This leads me to the following definition.

Definition 51 (Self-Organization) *If a time series (resp. spatial process) is dynamically autonomous, then it has* self-organized *between time t and time $t+T$ if and only if $C_\mu(t) < C_\mu(t+T)$ (resp., $\overline{{C_\mu}^{loc}}(t) < \overline{{C_\mu}^{loc}}(t+T)$).*

It would seem safe enough to apply this definition to non-feedback transducers if the complexity of the input process is zero, and similarly to non-autonomous spatial systems. It is not clear, however, how much an input with positive C_μ can contribute to increasing the organization of a transducer or a spatial process.

Second, it would be nice to *test* the formalization, by applying it to a large number of cases where we have clear intuitions, even proofs (Hanson and Crutchfield 1997) and seeing that it agrees with our intuition, before accepting it. The largest class of examples which combine intuitive consensus about self-organization, a guaranteed absence of outside organizers, and mathematical tractability are cellular automata.

What I hope to do in future work, therefore, is the following. I will assemble a large collection of two-dimensional CA rules, where a consensus exists as to whether or not they are self-organizing. Then, for each CA rule, I'll produce a large sample of its evolution from different random initial conditions, using the CAM8, a parallel computer specialized for running cellular automata[7]. This will give me enough data for the automatic reconstruction of each CA's ϵ-machine, and the estimation of $\overline{{C_\mu}^{loc}}$ as a function of time. Finally, I'll be able to see whether the rules which people think are self-organizing have increasing statistical complexity or not. It'll be particularly nice to be able to look at families of rules sharing a common form, and differ only by parameters, since some of them (e.g., the cyclic cellular automata) self-organize, but others don't, and the $\overline{{C_\mu}^{loc}}$ test ought to pick that up.

11.4 What Remains to Be Accomplished, or Things That Are Not Yet Theorems

11.4.1 Non-Stationarity

As I mentioned in Chapter 4, we do not need to assume we are dealing with stationary processes, merely with ones that are "conditionally stationary," i.e., the distribution of futures, conditional on histories, must be independent of when the history comes to an end. The conditionally-stationary processes form a comfortably large and roomy class, but they're not everything, and it would be nice if we could write down computational mechanics in a way which didn't invoke any sort of stationarity assumption.

The obvious thing to do, in the case of time series, is to say that $\overleftarrow{s}_{t_1}$ and $\overleftarrow{s}_{t_2}$ are causally equivalent when $\mathrm{P}(\overrightarrow{S}_{t_1} \in F | \overleftarrow{S}_{t_1} = \overleftarrow{s}_{t_1}) = \mathrm{P}(\overrightarrow{S}_{t_2} \in F | \overleftarrow{S}_{t_2} = \overleftarrow{s}_{t_2})$. If the process is conditionally stationary, this reduces to the normal notion of causal state. These states ought to be optimal minimal predictors, by the usual arguments, and I suspect they'll have deterministic transitions, though that's harder to see. What the ϵ-machine would look like, I really have no idea.

[6] I first learned of this point from Mitchell Porter.

[7] For details on the CAM8, see `http://www.im.lcs.mit.edu/cam8/`. For an even more detailed description of an earlier machine in the series, the CAM6, see Toffoli and Margolus (1987).

11.4.2 The Permutation City Problem

I haven't even bothered to state as an assumption that the order of observations in time and space is a given. However, if we're eliminating *a priori* assumptions, that one is questionable too. It may well be that if we re-ordered our data in some fashion, it would become easier to predict — in which case, why not do it? Why not form the causal states and the ϵ-machine on the re-organized data which are most efficiently predicted? Let's call this the Permutation City Problem, after the novel by Greg Egan (1994), which employs a similar conceit[8]. This feels very silly, but what, exactly, is wrong with it?

The basic flaw seems to be that re-arranging the data shouldn't be free; it takes a certain amount of information to specify the new arrangement, and the re-ordered predictor should be penalized by this amount. (Cf. the "recoding equivalence" of Crutchfield (1990).) If we have n data-points, specifying a permutation of them requires $\log n!$ bits, so predictive ability has to increase by $\frac{\log n!}{n}$ bits per symbol, or approximately $\log n$ bits per symbol as n gets large. That the predictive advantage of the rearranged series should increase at least logarithmically with n, for arbitrarily large n, is more than a bit implausible. Moreover, we really ought to perform the *same* rearrangement for every series from the same ensemble, if we want to capture anything about the *process*, as opposed to a particular realization. On the other hand, if we re-arrange the data at random, without performing any preliminary computations, then, almost by definition, we are simply randomizing the data stream, and destroying any predictable patterns it may contain.

There is no rigorous version of this argument. However, in the 1930s von Mises (1928/1981) and Reichenbach (Russell 1948) defined a "random collective" as, roughly, an infinite population whose every sub-population has the same distribution as the whole distribution. While this definition does not quite work, subsequent research has shown that it is adequate if we restrict ourselves to sub-samples which can be specified algorithmically (Salmon 1984). This suggests that it may be possible to give a rigorous answer to the Permutation City Problem, if we agree that only effectively-specifiable permutations are allowed[9].

11.4.3 Continuity

Throughout this book, I have assumed that space, time, and observables are all discrete; this is in keeping with all previous work on computational mechanics that I know of. It is fairly easy to formally extend the definitions of causal states to continuous variables. For instance, for time series with continuous time and values, we might say that two histories are causally equivalent when they give us the same conditional distribution over future trajectories.[10] There are three difficulties in the way of such a development.

First, it is not clear when the necessary conditional probability measures will mathematically exist. The regularity of conditional probabilities is quite easy for discrete processes; not so for continuous ones. It becomes a problem of functional analysis, so the mathematical foundations, if we aim at keeping to even the present standard of rigor, will get much more complicated. Still, we might invoke the physicist's license to ignore foundations, on the grounds that if it works, the mathematicians will find a way of making it right.

Second, much of the information theory I've used this development becomes inapplicable. Entropy and conditional entropy are defined for continuous variables, but they are coordinate-dependent — entropy is different if distances are measured in meters or in inches. This is distressing and unphysical. But mutual information is independent of coordinates, and so are statistical sufficiency and conditional independence, so we might be able to recover most of the results by leaning on them. (The role of C_μ, for instance, might

[8]To be precise, the novel's premise is as follows. The basic constituents of reality are an infinity of events. Every logical possible relation or set of relations which generates a spatio-temporal and causal ordering over some of those events leads to, or rather *is*, a universe containing just those events in that order. All possible universes co-exist outside time (since time is internal to universes), and all are equally real. For more on the generation of spatio-temporal order from relations among events, see Russell (1927, chs. 28–31). There it is proved that a countable infinity of (extended) events can generate a continuum of point-instants.

[9]The Permutation City Problem is due originally to Jim Crutchfield (1990), who also put forth the core of the answer above. The problem's most forceful current advocate is undoubtedly Murray Gell-Mann, who I hasten to add is not responsible for the name. A similar problem was considered earlier by Jorma Rissanen (1989, ch. 6), from whom I took the $\log n!$ idea.

[10]Having just come in to possession of a copy of Knight (1992), I suspect the resulting theory would look rather like his, but I'm not sure. A detailed comparison between his theory and computational mechanics should be made.

be taken by $I(\overleftarrow{S};\mathcal{S})$.)

Third, reconstruction from data becomes a *lot* harder. Even with the best analog instrumentation, we will never have an exact record of a continuous time series over a certain interval, which is what we would want. Even if we could get it, it would be hard (to say the least) to get repetitions of such a series, so that we could empirically estimate the necessary conditional probability densities. So it would seem that continuous computational mechanics could never be applied. But this is too hasty: any statistical analysis of continuous data faces the same problem, which isn't any *worse* for computational mechanics than for other methods. We might even join forces with them, by, say, using a nonparametric technique to estimate the conditional probabilities from sample data (Bosq 1998), or try fitting data to various basis functions (Crutchfield and McNamara 1987). This would impose prior restrictions on the function ϵ, which is something we want to avoid as much as possible, but, again, continuous computational mechanics certainly can't be worse in this regard than the existing techniques.

The other out would be to make a virtue of our limitations and explore the computational mechanics, not of continuous physical processes, but of continuous *models* of processes. For instance, the above definition of causal states can be applied to the Wiener process, $W(t)$: since, for any $T > 0$, $W(t+T) - W(t)$ is independent of all previous increments of the process, it is evident that each distinct value of $W(t)$, each distinct point in physical space, is a distinct causal state. This is a trivial example, but more interesting processes would yield to the same kind of analysis, with potentially interesting results, since very little is known about continuous, stochastic computation.

11.5 What Is to Be Done, or, Neat Things to Hack

I want to close by sketching out some areas in which computational mechanics could be, and is being, applied. One of the advantages of an abstract theory is that, because it is free of substantive assumptions, it can be applied to many problems which differ, perhaps radically, in their substance. This is by no means the only reason to want a general, abstract theory, but it may be a relief to descend from the empyrean to the muck of the lab-bench.

The ideas described in this section were developed in the Computation, Dynamics and Inference group at SFI, under the leadership of Jim Crutchfield. The terminology, in particular, is due to Jim. They represent active areas of research, and in some cases of collaboration. When I have a particular debt to someone outside the group, I've indicated it with a footnote.

11.5.1 The Real World

There are lots of data-sets crying out to be fed through ϵ-machine reconstruction algorithms. Mostly these are things where it's either very hard to come up with a decent model from first principles, or there's a real need to understand the intrinsic computation going on, or both. Most of the rest of this section will be about applications where exploratory work has been done in the Computation, Dynamics and Inference group at SFI.

A word first, though, about cases where it's just hard to come up with a good model. There is a large area in statistics, going by such names as "non-parametric regression," that tries to address the problem of finding predictive relationships between variables, without the benefit of a pre-set functional form for the relationship (Vapnik 1979/1982; Ripley 1996). Neural networks, in some of their avatars, are nonparametric regression functions (Zapranis and Refenes 1999). Maybe the most elegant theory of nonparametric regression is that employing the piecewise-polynomial functions called "splines" (Wahba 1990). Computer scientists study related techniques, typically in a less rigorous, more pragmatic way, as "data mining" (Weiss and Indurkhya 1998). Generally speaking, nonparametric regression methods employ a class of regression functions which are "universal approximators" — any well-behaved function can be approximated to arbitrary accuracy by some member of the class. (This is easily shown for neural networks, for instance.)

The difficulty comes when you try to cash in on this promise. When using a neural network, for instance, you must fix the architecture — so many nodes, in so many layers, and so on — and then train the network, given the data available. You then check its performance on new data, and decide whether or not it is adequate; if not, there is nothing for it but to pick a new architecture and try again. Moreover, to avoid over-fitting the training data, you have to start with small, simple, dumb networks, only going to more complex architectures when it is clear that no simple one can do the job (Vapnik 2000); this is called "capacity control". But why should we expect that a simple *relationship* should always be well-represented by a simple *neural network*? Maybe we should be using something else, like a spline, or a radial basis function. (Just because a series expansion converges doesn't mean that there isn't another expansion that converges faster; cf. Crutchfield (1992) on the distinction between complete and appropriate function bases.) Indeed computational learning theory has examples of problems that are easily learned using one class of representation but intractable with another (Kearns and Vazirani 1994). And what is true, in these respects, of neural networks is true of splines and all other conventional nonparametric methods.

It is not true, however, of ϵ-machines and ϵ-transducers. Since computational mechanics actually builds its models from data, architecture is not *guessed* but *inferred.* With the appropriate inference algorithm, the *simplest* possible architecture is inferred, eliminating the need for explicit capacity control. In other words, in almost any application domain where nonparametric or data-mining methods are used, computational mechanics is at least a contender.

11.5.1.1 Turbulence

I am *not* going to even hint that computational mechanics will solve the problem of turbulence (Frisch 1995). But it is often important to have a good model of, say, the velocity fluctuations at a point in a turbulent flow (e.g., for climate models). This has inspired a couple of attempts to infer causal states and ϵ-machines from turbulent flow data (Palmer, Fairall and Brewer 2000; Nicholas Watkins, personal communication, 2000). These efforts should be revisited, using the new reconstruction algorithm developed here. It would be very interesting to make an attack on how the statistical complexity and causal architecture of turbulence varies with Reynolds number (and, possibly, other control parameters). We might, for instance, settle the question of whether the transition to turbulence is self-organizing, with which we began.

11.5.1.2 Physical Pattern Formation

There are now a huge number of situations where experimentalists can reliably produce self-organized patterns of specific types. Excellent image-sequence data are available from many of them, thanks to digital cameras. An obvious but worthwhile project would be to take such a data set (from the Belousov-Zhabotinskii reagent, say) and feed it through a spatial ϵ-machine reconstruction algorithm. The output — the ϵ-machine — *should* include representations of all the acknowledged emergent structures (in the BZ case, spiral waves and organizing centers). If it does not, something is seriously wrong with the computational mechanics approach, simply because we *know* what's going on, macroscopically anyway, in these pattern-formers. Once experimentalists get comfortable with this sort of analysis, it will be natural for them to do it on *new* pattern-formers they encounter or devise, including ones from outside the lab.

11.5.1.3 Biosequences

About the second application of computational mechanics people suggest, on learning of it for the first time, is "DNA"[11]. Simply taking genome data and running it through an inference algorithm would be of relatively little interest, though it might turn up something. More promising would be to take ensembles of sequences which are *known* to have different functional properties (e.g., coding or non-coding, or belonging to different regulatory complexes), build their ϵ-machines, and see how those differ.[12] These could even be used to

[11] I'm so sick of the first application people suggest I won't even name it.

[12] We don't even have to do this for genes in the strict sense. For instance, single-stranded RNA folds up on itself, owing to interactions between bases along the strand, much as proteins do. While predicting the shape into which proteins will fold is very difficult, the RNA folding problem is fairly easily solved, at least for the ground-state conformation of the secondary structure. It's pretty simple to get large databases of RNA sequences and their folds. It would then be easy to construct the ϵ-machine for all the RNA sequences which fold into the same configuration. (Thanks to Walter Fontana for suggesting this

identify the family to which newly-sequenced genes belong; hidden Markov models are already used for this purpose, but those HMMs are constructed by the usual *ad hoc* methods, and could certainly be improved. An ϵ-transducer, built from the same data-set, would provide classifications directly.

11.5.1.4 Neural Coding

Neurons convey information to one another by means of brief, intense, highly-stereotyped electrical impulses known as *action potentials* or *spikes*.[13] Presumably, the pattern of spikes one neuron receives from another — the *spike train* — conveys information about what the upstream neuron has computed about the world, itself based on the spike trains it received. Ultimately, spike trains encode information about the world, or at least about the organism's sensory organs. The neural coding problem (Rieke, Warland, de Ruyter van Steveninck and Bialek 1997) is, simply, How is that information encoded? Given a spike train, how ought it be decoded? If we regard the neuron as a transducer, this amounts to attempting to model its input-output relation. Remarkable progress has been made recently by applying techniques from information theory (Dimitrov and Miller 2001) and by calculating the first Wiener kernel (Dayan and Abbott 2001), i.e., by attempting linear decoding.

Clearly, we should calculate ϵ-transducers and so nonlinear decoding "to all orders". The transducer states would tell us what features of input spike trains a given neuron is actually sensitive to, for instance, and so what kinds of computations it is able to perform. The full ϵ-transducer would allow us to calculate what ensemble of inputs will maximize the information content of the neuron's output, and see whether, as many speculate (Rieke, Warland, de Ruyter van Steveninck and Bialek 1997), and seems reasonable on evolutionary grounds, the distribution of natural stimuli is close to that which maximizes output-information.

Neurons do not work in isolation; in particular, it's pretty well established that "population codes" are a key part of neural representation and computation (Abbott and Sejnowski 1998). In these cases, the actions of individual neurons are comparatively insignificant, information being encoded in the pattern of activity across the population. There is no in-principle reason why we could not construct a single ϵ-transducer for the entire population and use it to figure out the population code, just as we could for an individual neuron. In fact, by extending the results of the section on feedback above, we could in some sense compose the population's ϵ-transducer from those of the individual neurons.

11.5.1.5 Signal Transduction, Gene Regulation, and Metabolic Networks

Signal transduction is the process by which cells detect and respond to environmental conditions, such as the concentrations of different sorts of chemicals, pressure, heat, light, electrical fields, and so forth. It is carried on by an intricate array of specialized and general-purpose signaling molecules, ranging from large protein complexes to calcium ions. Signal transduction is intimately related to gene regulation, the turning on or off of the expression of the various genes in the cell's genome, or more generally the control of the rate at which different genes are expressed. Gene regulation, in turn, is part of the control of metabolism, which is also connected directly to signal transduction.[14]

Huge volumes of data are now becoming available about all three processes, largely because of new experimental devices, such as "gene chips", which record the expression levels of thousands of genes simultaneously.

application.)

[13] They also communicate by chemical means, but let's pretend otherwise for now.

[14] The literature on all these biological processes, taken separately, is vast, and by some estimates doubles every twelve months. Gonick and Wheelis (1991) has a characteristically engaging discussion of the fundamentals of gene regulation. For philosophical views of these topics, see Monod (1970/1971) and Goodenough (1998). Loewenstein (1999), while written by a very distinguished experimenter, is full of misconceptions about information theory and nonlinear dynamics.

Hancock (1997) is intended as an introduction to signal transduction for biology students; it is straightforward, but presumes a high capacity for memorizing molecular names. Ptashne (1992) describes one of the very first instances of gene regulation to be understood in full detail, but mercifully stuffs the experimental details into appendices. Krauss (1999) was authoritative when it was published, and so should not be *absurdly* out of date when you read this. Milligan (1999) and Carraway and Carraway (2000) have practical details on experimental systems and approaches.

Quantitative treatments of these topics are rare. Fell (1997) may be the best point of entry for physicists or mathematicians. I have not had a chance to read Bower and Bolouri (2001).

These data sets cry out for statistical modeling, but very little is known about the kinds of relationships we should expect to find in the data, meaning that traditional statistical methods, attempting to estimate pre-defined parameters, are simply not applicable. This has lead those doing bioinformatics (Baldi and Brunak 1998) to develop non-parametric and data-mining methods.

The role of computational mechanics here would be, again, to provide a method for discovering patterns in the data which does not require prior assumptions about what those patterns are like, yet has proven optimality properties, and will find any patterns in the data which have any predictive power. The ϵ-transducer estimated from biological data would be an abstract model of the input-output characteristics of the signaling or regulatory network that provided the data, including its computational abilities. The information-processing ability of a single cell is often considerable, even if it is *not* a nerve cell (Holcombe and Paton 1998), and it would be very good to understand it, particularly since it's so important in keeping us alive.

The ϵ-transducer would also serve to constrain more conventional models of the functional and chemical-kinetic architecture of the network, things of the "this kinase phosphorylates that enzyme" type: the conventional models would have to reproduce the behavior of the ϵ-transducer, would have to provide (in the logical sense) models for it. But the constraint could also go the other way: given that we know a certain functional pathway exists, it would be nice if our reconstruction algorithm could use that knowledge to narrow its search. I have no idea of how to implement such constraints, but it would make for an important addition to the theory.[15]

11.5.1.6 Agents

An agent, etymologically, is something which acts; in the lapidary formulation of Stuart Kauffman, a "thing which does things to things". From the point of view of computational mechanics, an agent is simply a transducer. The input series represents the agent's environment; the output, the actions the agent takes. Putting things this way does *not* imply that the agent is limited to simple stimulus-response behaviors; that would imply a memoryless transducer. Instead the agent can do arbitrarily complicated internal information processing, all of it represented by the internal states and connections of the ϵ-transducer[16]. If an agent's actions influence the part of its environment to which it is sensitive (generally the case), then the feedback states represent the effects of its actions, its ability to make differences to its environment. The problem confronting an adaptive agent, or an agent designer, isn't so much selecting good actions, as selecting actions which produce desirable causal states.

Saying that the agent has "*an* environment" does not mean that it will not, sometimes, be desirable to explicitly represent the various parts of that environment, including, potentially, observable attributes of other agents. Reconstructing the ϵ-transducers from data for a population of interacting agents would allow us to infer the network of interactions among them, as well as the intrinsic computations that take place within each agent in its dealings with others. We might even be able to adapt the techniques of spatial computational mechanics (Chapter 10) to characterize the *global* information-processing capabilities of the population of agents — their collective cognition (Shalizi 1998a) and other distributed adaptations (Crutchfield, personal communication), and do so in impeccably materialist, individualist terms.

A simple example[17] may make these abstractions a bit clearer. Consider an ant. At any given time, it is performing one of a number of behaviors, which are readily observed and categorized. In the course of its activities, it moves about a varying physical environment, and comes into contact with other ants, performing other behaviors. From time to time, the ant switches behaviors. Take the state of the ant's immediate physical environment, and the outward behavior of the ant it is currently dealing with (if any), as the input. The output is the manifest behavior of the ant. By treating it as a transducer, we see how the

[15] I am grateful to Aryaman Shalizi for suggesting this application, and educating me about signal transduction.

[16] Since a transducer is a channel with memory, an adaptive agent is a *learning channel* — a pun for which Jim Crutchfield is solely responsible. Actually, ϵ-transducers very easily include the "operator models" of psychological learning theory as special cases (Bush and Mosteller 1955; Sternberg 1963; Holland 1990), but they can handle other modes of learning too, such as those of Holland, Holyoak, Nisbett and Thagard (1986).

[17] Suggested by Michael Lachmann. Cf. Delgado and Solé (1997).

ant's past history, physical environment, *and dealings with other ants* control its task-switching. We could also build transducers for all the other ants in the colony (perhaps by treating things like caste as fixed inputs), and ultimately compose them into the global ϵ-machine for the ant colony.

11.5.2 The Dynamics of Learning

Computational mechanics sets limits on how well processes can be predicted, and shows how, at least in principle, those limits can be attained. ϵ-machines are what any prediction method would build, if only they could. But any learning problem which is formal and definite enough that we can say whether or not it's been successfully solved is also a prediction problem, or at least equivalent to one (Thornton 2000). So, in a sense, ϵ-machines are also what every *learning* method wants to build. Computational mechanics thus has some important things to say about how well learning can succeed in different environments, and what optimal learning looks like (Shalizi and Crutchfield 2000c).

Conversely, when we try to reconstruct an ϵ-machine from actual data, we are engaging in a kind of learning, or at least our code is. If we want to learn well, i.e., do reconstruction well, we need to take into account results from learning theory about when and how learning is possible. I have already gestured at some results of this sort (for instance, claiming that constricting the space of possible models speeds convergence on the best one), but the literature has quantitative and powerful results. Unfortunately, most of them assume both a fixed mode of representation (a fixed model class) *and* IID data. Developing a quantitative learning theory for ϵ-machines, therefore, will mean extending statistical learning theory to dependent data. The ultimate goal would be a theory of learning in a changing environment, where the learner is itself a dynamical system — to understand the *dynamics of learning*, in Crutchfield's phrase.

Animals prove that this kind of learning is possible, and set a lower bound on how well it can be achieved: anything a sea slug, a lorikeet, or a tenured professor can do, a learning algorithm can do. What is not clear is that any of them, even the most highly adapted of them[18], learns as well as possible, i.e. that any of them attains the upper bound on learning ability, if there is one. To answer that question, we need theory, especially the kind of optimality theory computational mechanics is able to provide.

11.5.3 Phenomenological Engines

"Phenomenology", for physicists, is the study and modeling of phenomena, without much if any attempt to get at underlying mechanisms[19]. An immense amount of what people do in applied science, engineering, and related technical fields is basically phenomenology. They need to make day-to-day predictions, but either don't know the underlying mechanisms, or those mechanisms are too cumbersome to use for practical problems. Empirical regularities must take their place. Sometimes entire fields are devoted to teasing such regularities out of data; econophysics, for instance, consists of little more than attempts to get the phenomenology of financial time series right (Mantegna and Stanley 2000).

More respectably, phenomenology is often a crucial preliminary to understanding mechanisms, since an accurate knowledge of the phenomena and their relations constraints mechanical models; the classic case is the relationship between Mendelian and molecular genetics. The former is quite abstract, merely saying that there are causal factors, called genes, which influence the observable traits of organisms and are passed, in a certain manner, from parents to offspring. This is enough to have very important consequences, for instance, most of evolutionary genetics (Gillespie 1998), but it's quite mechanism-free; it is even compatible with the assumption that genetic influences are mediated by immaterial souls. Molecular genetics provides all the grubby mechanical details missing from Mendelism and is in many cases much more accurate into the bargain; but we were only led to it because it at least approximately fulfilled Mendelian expectations.[20]

[18] The sea slug.

[19] "Phenomenology" in philosophy also disdains mechanisms, but for entirely different, and far less creditable, reasons (Husserl 1913/1931; Kolakowski 1975; Gellner 1974).

[20] The relationship between the abstract, structural theory and the mechanical one is somewhat like that between an axiom system and one of its models in logic (Manzano 1990/1999), but not quite, because the abstract theory may only approximate the more realistic one.

I touched on this briefly when considering empirical applications above.

In computational mechanics, we have an automatic method for doing phenomenology. An ϵ-machine reconstruction algorithm takes in data and gives back a representation of causal patterns, suitable for use in prediction or intervention, "untouched by human hands". Such an algorithm is a *phenomenological engine* or *phenomenologimat*[21]. There is no in-principle reason why they could not become fast, reliable, standard pieces of software, with potentially amusing and even important consequences. They would spell the end of on-line gambling and human weathermen; but also stock-market quants, biomedical statisticians, many sorts of engineer, and routine medical diagnosticians[22]. Even data-analysts at high-energy physics experiments will find it hard to justify their existence — once a phenomenologimat gets written in Fortran.

[21]Thanks to Jon Fetter for these names.

[22]It has been known for a long time that, in many areas, human clinical judgment is significantly less accurate than the results of simple linear decision rules (Dawes, Faust and Meehl 1989). Phenomenologimats could invade domains where linear rules do not apply, but nonlinear ones do.

Appendix A

Mathematical Review

A.1 Equivalence Relations and Partitions

The following definitions and properties are well-known, and may be found in almost any book on abstract algebra or set theory.

Definition 52 (Equivalence Relation) *An* equivalence relation $\sim$ *on a set A is a relation on A that is reflexive, symmetric and transitive:*

$$\text{Reflexive}: \quad \forall a \in A \qquad a \sim a \tag{A.1}$$
$$\text{Symmetric}: \quad \forall a, b \in A \qquad (a \sim b) \Leftrightarrow (b \sim a) \tag{A.2}$$
$$\text{Transitive}: \quad \forall a, b, c \in A \qquad (a \sim b) \wedge (b \sim c) \Rightarrow (a \sim c) \tag{A.3}$$

Definition 53 (Equivalence Class) *An* equivalence class e *in A is a maximal sub-set of mutually equivalent elements: for all $a \in e$, $a \sim b$ iff $b \in e$. The equivalence class containing a is sometimes written $[a]$. The collection of all equivalence classes induced by $\sim$ in A is written $A/\sim$.*

Definition 54 (Partition) *A partition P of a set A is a class $P_0, P_1, \ldots$ of mutually exclusive and jointly exhaustive subsets of A:*

$$\text{Mutually exclusive}: \quad \forall P_i, P_j \in P \qquad P_i \cap P_j = \emptyset \tag{A.4}$$
$$\text{Jointly exhaustive}: \quad \forall a \in A,\ \exists P_i \in P \quad a \in P_i \tag{A.5}$$

The members of P are called the cells *of the partition. If there is only one cell, the partition is* trivial. *If each element of A has its own cell, the partition is the* identity partition.

Definition 55 (Refinement) *One partition P* refines *another partition, Q, if each cell of P is a subset of a cell of Q:*

$$\forall p \in P \quad \exists q \in Q \text{ s.t.} \quad p \subseteq q \tag{A.6}$$

P is finer *than Q; it is a* refinement *of Q; Q is* coarser *than P.*

Proposition 3 (Equivalence Relations and Partitions) *For any equivalence relation $\sim$ on A, the collection of equivalence classes $A/\sim$ forms a partition of A. Conversely, every partition of A corresponds to an equivalence relation.*

A.2 Information Theory

Information theory appeared in essentially its modern form with Shannon (1948), though there had been predecessors in both communications (Hartley 1928) and statistics, notably Fisher (see (Kullback 1968) for an exposition of these notions), and similar ideas were developed by Wiener and von Neumann, more or less independently of Shannon (Wiener 1961). Shannon and Weaver (1963) contains the classic papers; Pierce (1961) is a decent popular treatment.

Appendix A.2.4 lists a number of useful information-theoretic formulæ, which get called upon in our proofs. Throughout, our notation and style of proof follow those in (Cover and Thomas 1991), the definitive modern reference.

A.2.1 Entropy Defined

Given a random variable X taking values in a countable set $\mathcal{A}$, the entropy of X is

$$H[X] \equiv -\sum_{x \in \mathcal{A}} \mathrm{P}(X = x) \log_2 \mathrm{P}(X = x) \ , \tag{A.7}$$

taking $0 \log 0 = 0$. Notice that $H[X]$ is the expectation value of $-\log_2 \mathrm{P}(X = x)$ and is measured in *bits* of information. Caveats of the form "when the sum converges to a finite value" are implicit in all statements about the entropies of infinite countable sets $\mathcal{A}$.

Shannon interpreted $H[X]$ as the *uncertainty in* X. (Those leery of any subjective component in notions like "uncertainty" may read "effective variability" in its place.) He showed, for example, that $H[X]$ is the mean number of yes-or-no questions needed to pick out the value of X on repeated trials, if the questions are chosen to minimize this average (Shannon 1948).

A.2.2 Joint and Conditional Entropies

We define the joint entropy $H[X, Y]$ of two variables X (taking values in $\mathcal{A}$) and Y (taking values in $\mathcal{B}$) in the obvious way,

$$H[X, Y] \equiv -\sum_{(x,y) \in \mathcal{A} \times \mathcal{B}} \mathrm{P}(X = x, Y = y) \log_2 \mathrm{P}(X = x, Y = y) \ . \tag{A.8}$$

We define the conditional entropy $H[X|Y]$ of one random variable X with respect to another Y from their joint entropy:

$$H[X|Y] \equiv H[X, Y] - H[Y] \ . \tag{A.9}$$

This also follows naturally from the definition of conditional probability, since $\mathrm{P}(X = x|Y = y) \equiv \mathrm{P}(X = x, Y = y)/\mathrm{P}(Y = y)$. $H[X|Y]$ measures the mean uncertainty remaining in X once we know Y.

A.2.3 Mutual Information

The *mutual information* $I[X; Y]$ between two variables is

$$I[X; Y] \equiv H[X] - H[X|Y] \ . \tag{A.10}$$

This is the average reduction in uncertainty about X produced by fixing Y. It is non-negative, like all entropies here, and symmetric in the two variables.

The conditional mutual information $I[X; Y|Z]$ is

$$I[X; Y|Z] \equiv H[X|Z] - H[X|Y, Z] \ . \tag{A.11}$$

It is also non-negative and symmetric in X and Y. It can be larger or smaller than the unconditional mutual information.

A.2.4 Information-Theoretic Formulæ

The following formulæ prove useful in the development. They are relatively intuitive, given our interpretation, and they can all be proved with little more than straight algebra; see Cover and Thomas (1991, ch. 2). Below, f and g are functions.

$$
\begin{aligned}
H[X,Y] &= H[X] + H[Y|X] && \text{(A.12)}\\
H[X,Y] &\geq H[X] && \text{(A.13)}\\
H[X,Y] &\leq H[X] + H[Y] && \text{(A.14)}\\
H[X|Y] &\leq H[X] && \text{(A.15)}\\
H[X|Y] = H[X] &\text{ iff } X \text{is independent of } Y && \text{(A.16)}\\
H[X,Y|Z] &= H[X|Z] + H[Y|X,Z] && \text{(A.17)}\\
H[X,Y|Z] &\geq H[X|Z] && \text{(A.18)}\\
H[X] - H[X|Y] &= H[Y] - H[Y|X] && \text{(A.19)}\\
I[X;Y] &\leq H[X] && \text{(A.20)}\\
I[X;Y] = H[X] &\text{ iff } H[X|Y] = 0 && \text{(A.21)}\\
H[f(X)] &\leq H[X] && \text{(A.22)}\\
H[X|Y] = 0 &\text{ iff } X = f(Y) && \text{(A.23)}\\
H[f(X)|Y] &\leq H[X|Y] && \text{(A.24)}\\
H[X|f(Y)] &\geq H[X|Y] && \text{(A.25)}\\
I[f(X);g(Y)] &\leq I[X;Y] && \text{(A.26)}\\
I[f(X);g(Y)|Z] &\leq I[X;Y|Z] && \text{(A.27)}
\end{aligned}
$$

Eqs. A.12 and A.17 are called the *chain rules* for entropies. Strictly speaking, the right hand side of Eq. A.23 should read "for each y, $\mathrm{P}(X = x|Y = y) > 0$ for one and only one x".

A.3 Statistical Independence and Conditional Independence

Definition 56 (Statistical Independence) *Two random variables X and Y are* statistically independent *iff their joint probability distribution factors:*

$$\mathrm{P}(X = x, Y = y) = \mathrm{P}(X = x)\mathrm{P}(Y = y) \qquad \text{(A.28)}$$

or, equivalently, conditioning the one on the other makes no difference:

$$
\begin{aligned}
\mathrm{P}(X = x|Y = y) &= \mathrm{P}(X = x) && \text{(A.29)}\\
\mathrm{P}(Y = y|X = x) &= \mathrm{P}(Y = y) && \text{(A.30)}
\end{aligned}
$$

The classic treatment of statistical independence is Kac (1959).

Proposition 4 (Statistical Independence and Mutual Information) *(Cover and Thomas 1991, p. 27) X and Y are independent iff $I[X;Y] = 0$.*

Vitally important for our purposes is the derivative notion of conditional independence.

Definition 57 (Conditional Independence) *Two random variables X, Y are* conditionally independent *given a third, Z, (or "independent given Z") if and only if*

$$\mathrm{P}(X|Y,Z) = \mathrm{P}(X|Z) \qquad \text{(A.31)}$$

or, equivalently,

$$\mathrm{P}(X,Y|Z) = \mathrm{P}(X|Z)\mathrm{P}(Y|Z) \tag{A.32}$$

When this relation obtains, we write $X \perp\!\!\!\perp Y|Z$.

Proposition 5 (Conditional Independence and Mutual Information) *(Cover and Thomas 1991, p. 27) $X \perp\!\!\!\perp Y|Z$ iff $I(X;Y|Z) = 0$.*

A.3.1 Properties of Conditional Independence

We list only those we need. Monographs on graphical models (Pearl 2000; Spirtes, Glymour and Scheines 2001) contain more extensive lists.

$$(A \perp\!\!\!\perp B|CD) \wedge (A \perp\!\!\!\perp D|CB) \Rightarrow (A \perp\!\!\!\perp BD|C) \tag{A.33}$$
$$(A \perp\!\!\!\perp BC|D) \Rightarrow (A \perp\!\!\!\perp B|CD) \tag{A.34}$$
$$(A \perp\!\!\!\perp B|C) \wedge (A \perp\!\!\!\perp D|CB) \Rightarrow (A \perp\!\!\!\perp BD|C) \tag{A.35}$$

Life would be much easier if

$$(A \perp\!\!\!\perp B) \Rightarrow (A \perp\!\!\!\perp B|C) \tag{A.36}$$
$$(A \perp\!\!\!\perp B|C) \Rightarrow (A \perp\!\!\!\perp B|CD) \tag{A.37}$$

but sadly, neither of these implications holds in general; *adding* a conditional variable can make A and B dependent again!

The following property, while not included in most lists of conditional independence properties, is of some use to us:

$$(A \perp\!\!\!\perp B|C) \Rightarrow (A \perp\!\!\!\perp f(B)|C) \wedge (A \perp\!\!\!\perp B, f(B)|C) \tag{A.38}$$

for any measurable, nonrandom function f. It follows directly from the combination of Eq. A.27 and Proposition 5.

There is an important connection between conditional independence and statistical sufficiency; see Appendix A.5 below.

A.4 Automata Theory

Definition 58 (Formal Language) *A* formal language *$\mathcal{L}$ over the finite alphabet Σ is a subset of Σ^* — the set of all possible words, or strings, made up of symbols from Σ.*

Definition 59 (Determinism) *An automaton is* deterministic *or* has deterministic transitions *if, given its current state and its next input, there is only one possible next state for it.*

This definition often causes confusion, since many stochastic automata (i.e., ones with probabilistic transitions) are deterministic in this sense. But it is too thoroughly entrenched in computer science to be changed.

Definition 60 (Deterministic Finite Automaton) *A* deterministic finite automaton *(DFA) M is defined as a 5-tuple:*

$$M = \{Q, \Sigma, \delta, q_0, F\} , \tag{A.39}$$

where Q is a finite set of states, *Σ is an* alphabet, *$q_0 \in Q$ is the* initial *state, $F \subseteq Q$ is a set of* final *states, and $\delta : Q \times \Sigma \to Q$ is a* transition function*: $\delta(q,a) = q'$, where $q, q' \in Q$ and $a \in \Sigma$.*

A DFA can be used to read, or scan, words $w = w_1 \dots w_L$ over the alphabet Σ. Starting in the initial state q_0, the DFA reads the first symbol w_1 of the word w. It then makes a transition to another state $q' = \delta(q_0, w_1)$. The DFA then reads the next symbol w_2 and makes a transition to $q'' = \delta(q', w_2)$, and so on until all symbols in w have been read or until an undefined transition is encountered. If, after reading w, the DFA ends in a final state $q \in F$, M *accepts* w; otherwise M *rejects* it.

Definition 61 (Regular Language) *A* regular *language $\mathcal{L}$ is a formal language for which there exists a DFA that accepts all words in $\mathcal{L}$ and rejects all words not in $\mathcal{L}$.*

Regular languages are the simplest class of formal languages in a hierarchy (the Chomsky hierarchy) of language classes of increasing complexity (Lewis and Papadimitriou 1998).

There are generally many DFAs that recognize the same regular language $\mathcal{L}$, but there is a unique minimal DFA for $\mathcal{L}$, which we write $M(\mathcal{L})$. (For a nice proof of this proposition, see (Lewis and Papadimitriou 1998).) Similarly, for every DFA M there is a corresponding regular language $\mathcal{L}(M)$ consisting of all and only the words that are accepted by M.

Definition 62 (Regular Process Language) *A regular language is a* regular process language *if every subword of a word in $\mathcal{L}$ is also a word in $\mathcal{L}$.*

Definition 63 (Process Graph) *A DFA is a process graph if its every state is both an initial and an accepting state.*

The DFAs corresponding to regular process languages are process graphs, and vice versa (Hanson and Crutchfield 1992).

Definition 64 (Finite State Transducer) *A* finite-state transducer *(FST) is a finite automaton with two kinds of symbol associated with each transition: inputs and outputs. An FST R is defined by a 7-tuple:*

$$R = \{Q, \Sigma_{in}, \Sigma_{out}, \delta, \lambda, q_0, F\} \ , \tag{A.40}$$

where Q, δ, q_0, and F are as in a DFA, Σ_{in} is the input alphabet*, Σ_{out} is the* output alphabet*, and $\lambda : Q \times \Sigma_{in} \to \Sigma_{out}$ is the* observation function*: $\lambda(q, a) = b$ where $q \in Q$, $a \in \Sigma_{in}$, and $b \in \Sigma_{out}$. An FST effectively implements a mapping f_R from one language over Σ_{in} to another language over Σ_{out}. In other words, it reads a word $w \in \Sigma_{in}^*$ and transforms it to another word $w' \in \Sigma_{out}^*$ by mapping each symbol $w_i \in \Sigma_{in}$ to a symbol $w'_i \in \Sigma_{out}$ such that $w'_i = \lambda(q, w_i)$, where $q \in Q$ is the current state of R when reading w_i.*

In formal language theory, languages and automata play the role of sets and transducers the role of functions.

A.5 Sufficient Statistics

Definition 65 (A Statistic) *Let X be a random variable taking values from $\mathbf{X}$. Then a* statistic *T on (or "over") X is any measurable, non-random function of X, i.e., $T = f(X)$. To each statistic T there corresponds a partition $\mathbf{T}$ of $\mathbf{X}$.*

Remark 1. It is common to define the "same" statistic over any number of samples $X_1, X_2, \dots X_n$ taken in the same space, such as a stochastic process. For simplicity, what follows always writes X as a single variable, but this should be kept in mind.

Definition 66 (Predictive Sufficiency) *A statistic T over a random variable X is a sufficient statistic for predicting another random variable Y iff and only if $\mathrm{P}(Y|T = f(x)) = \mathrm{P}(Y|X = x)$, $\forall x$. If T is sufficient, then we also say that its associated partition $\mathbf{T}$ of $\mathbf{X}$ is sufficient.*

Remark 1. Sufficiency is important to prediction because it can be shown that, for any prediction method which uses a non-sufficient statistic can be bettered by one which does. (A more precise statement can be found in Appendix D below.)

Remark 2. Predictive sufficiency is related to, but not identical with, the idea of *parametric sufficiency.* That, roughly, is when a statistic contains all available information about the parameters of an unknown distribution. That is, if the probability distribution is parameterized by θ, written P_θ, a statistic T is parametrically sufficient if and only if $\mathrm{P}_\theta(X|T=t) = \mathrm{P}_{\theta'}(X|T=t)$ for all θ, θ'. If θ itself can be regarded as a random variable (as in de-noising, Bayesian statistics, etc.), then parametric and predictive sufficiency are identical.

Lemma 37 (Sufficiency and Conditional Independence) *Consider two random variables X and Y, and a statistic T on X. Then $X \perp\!\!\!\perp Y|T$ if and only if T is sufficient for predicting Y from X.*

Proof. "Only if": By conditional independence (Eq. A.31), $\mathrm{P}(Y|X,T) = \mathrm{P}(Y|T)$. But since $T = f(X)$, by Lemma 39 $\mathrm{P}(Y|X,T) = \mathrm{P}(Y|X)$. (Informally, T is a "coarser" variable than X, so conditioning on T has no effect once we've conditioned on X.) So $\mathrm{P}(Y|T) = \mathrm{P}(Y|X)$, which means T is sufficient. "If": We start with $\mathrm{P}(Y|T) = \mathrm{P}(Y|X)$. As before, since $T = f(X)$, $\mathrm{P}(Y|X) = \mathrm{P}(Y|X,T)$. Hence $\mathrm{P}(Y|T) = \mathrm{P}(Y|X,T)$, so (Eq. A.31), $X \perp\!\!\!\perp Y|T$. QED.

Proposition 6 (Predictive Sufficiency and Mutual Information) *(Cover and Thomas 1991, p. 37; Kullback 1968, sec. 2.4–2.5) T is a sufficient statistic over X for predicting Y if and only if $I(Y;T) = I(Y;X)$.*

Definition 67 (Minimal Sufficiency) *A statistic T is a* minimal sufficient statistic *for predicting Y from X if and only if it is predictively sufficient, and it is a function of every other sufficient statistic.*

Remark. If $\mathbf{T}$ is the partition corresponding to the minimal sufficient statistic T, then every other sufficient partition $\mathbf{Q}$ must be a refinement of $\mathbf{T}$. Turned around, no partition coarser than $\mathbf{T}$ is sufficient.

Appendix B

Mathematical Annex to Chapter 4

B.1 Time Reversal

We can imagine forming *reverse causal states* for futures, based on their conditional distribution of histories, i.e., assigning two futures to the same state if and only if they have the same conditional distribution for histories. While both the reverse states and the ordinary, forward ones render the past and the future of the process conditionally independent, there is no other general, systematic relationship between the two of them. The past and future morphs can be very different, and while both sets of causal states render past and future conditionally independent, one is a function of the past. In order to determine the reverse causal state from the forward state, we must be able to determine the forward state from the future of the process; to get the forward state from the reverse state, we must be able to determine the history uniquely from the reverse state. If the forward and reverse states can both be inferred from each other, so that there is a kind of time-reversal symmetry in the causal states, then there must be a $1-1$ correspondence between futures and histories.

In general, $\overleftarrow{C}_\mu \neq \overrightarrow{C}_\mu$ (Crutchfield 1992), whereas the entropy rates (Crutchfield and Shalizi 1999) and excess entropies must be equal. And so on.

B.2 ϵ-Machines are Monoids

A *semi-group* is a set of elements closed under an associative binary operator, but without a guarantee that every, or indeed any, element has an inverse (Ljapin 1963). A *monoid* is a semi-group with an identity element. Thus, semi-groups and monoids are generalizations of groups. Just as the algebraic structure of a group is generally interpreted as a symmetry, we propose to interpret the algebraic structure of a semi-group as a *generalized* symmetry. The distinction between monoids and other semi-groups becomes important here: only semi-groups with an identity element — i.e., monoids — can contain subsets that are groups and so represent conventional symmetries.

We claim that the transformations that concatenate strings of symbols from $\mathcal{A}$ onto other such strings form a semi-group G, the generators of which are the transformations that concatenate the elements of $\mathcal{A}$. The identity element is to be provided by concatenating the null symbol λ. The concatenation of string t onto the string s is forbidden if and only if strings of the form st have probability zero in a process. All such concatenations are to be realized by a single semi-group element denoted $\emptyset$. Since if $\mathrm{P}(st) = 0$, then $\mathrm{P}(stu) = \mathrm{P}(ust) = 0$ for any string u, we require that $\emptyset g = g\emptyset = \emptyset$ for all $g \in G$. Can we provide a representation of this semi-group?

Recall that, from our definition of the labeled transition probabilities, $T^{(\lambda)}_{ij} = \delta_{ij}$. Thus, $\mathbf{T}^{(\lambda)}$ is an identity element. This suggests using the labeled transition matrices to form a matrix representation of the semi-group. Accordingly, first define $U^{(s)}_{ij}$ by setting $U^{(s)}_{ij} = 0$ when $T^{(s)}_{ij} = 0$ and $U^{(s)}_{ij} = 1$ otherwise, to remove probabilities. Then define the set of matrices $\mathbf{U} = \{\mathbf{T}^{(\lambda)}\} \bigcup \{\mathbf{U}^{(\mathbf{s})}\ , \mathbf{s} \in \mathcal{A}\}$. Finally, define G as the set of all matrices generated from the set $\mathbf{U}$ by recursive multiplication. That is, an element g of G is

$$g^{(ab\ldots cd)} = \mathbf{U}^{(d)}\mathbf{U}^{(c)}\ldots\mathbf{U}^{(b)}\mathbf{U}^{(a)} \ , \tag{B.1}$$

where $a, b, \ldots c, d \in \mathcal{A}$. Clearly, G constitutes a semi-group under matrix multiplication. Moreover, $g^{(a\ldots bc)} = \mathbf{0}$ (the all-zero matrix) if and only if, having emitted the symbols $a \ldots b$ in order, we must arrive in a state from which it is impossible to emit the symbol c. That is, the zero-matrix $\mathbf{0}$ is generated if and only if the concatenation of c onto $a \ldots b$ is forbidden. The element $\emptyset$ is thus the all-zero matrix $\mathbf{0}$, which clearly satisfies the necessary constraints. This completes the proof of Proposition 9.

We call the matrix representation — Eq. B.1 taken over all words in $\mathcal{A}^k$ — of G the *semi-group machine* of the ϵ-machine $\{\boldsymbol{\mathcal{S}}, \mathbf{T}\}$ (Young 1991).

B.3 Measure-Theoretic Treatment of Causal States

In Section 4.2, where we define causal states, ϵ-machines, and their basic properties, we use a great many conditional probabilities. However, there are times when the events on which we condition — particular histories, or particular effective states — have probability zero. Then classical formulæ for conditional probability do not apply, and a more careful and technical treatment, going back to the measure-theoretic basis of probability, is called for. That's what I do here, showing that all the concepts we saw in Section 4.2 — the causal states, their morphs, and so forth — are well-defined measure-theoretically. The proofs in that section are equally valid whether we interpret the conditional probabilities they invoke classically or measure-theoretically. (The measure-theoretic interpretation raises a few technicalities, which we have flagged with footnotes to those proofs.) And we show here that our methods of proof in subsequent sections are not affected by this change in interpretation.

In what follows, I draw on Billingsley (1965, Billingsley (1979), Doob (1953), Gray (1990), Loéve (1955), and Rao (1993). I assume that the reader is familiar with measure-theoretic probability, at least in some basic way. The notation broadly follows that of Billingsley. A slightly different approach to these issues, and more than slightly different terminology and notation, may be found in chapter 2 of Upper (1997).

B.3.1 Abstract Definition of Conditional Probability

Definition 68 (Conditional Probability) *Consider a probability space $(\Omega, \mathcal{F}, P)$ and a σ-subalgebra $\mathcal{G} \subset \mathcal{F}$. The* conditional probability *of an event $A \in \mathcal{F}$, given the family of events $\mathcal{G}$, is a real-valued random function $P_{A||\mathcal{G}}(\omega)$, with the following properties:*

1. *$P_{A||\mathcal{G}}(\omega)$ is measurable with respect to $\mathcal{G}$; and*
2. *for any $G \in \mathcal{G}$,*

$$\int_G P_{A||\mathcal{G}}(\omega)dP \quad = \quad P(A \cap G) \tag{B.2}$$

The latter condition generalizes the classical formula that $P(A \cap G) = \sum_{g \in G} P(A|g)P(g)$.

Proposition 7 *There always exists a function $P_{A||\mathcal{G}}(\omega)$ satisfying the just-given conditions. Moreover, if f and g are two functions which both satisfy the above requirements, $f(\omega) = g(\omega)$ for P-almost-all ω.*

Proof: The existence of such random variables is vouchsafed to us by the Radon-Nikodym theorem; $P_{A||\mathcal{G}}(\omega)$ is the Radon-Nikodym derivative of $P(A \cap G)$, which is a measure over $\mathcal{G}$, with respect to P. (The latter is also restricted to the σ-subalgebra $\mathcal{G}$.) The Radon-Nikodym theorem also tells us that any two functions which satisfy the two conditions above agree for P-almost-all points ω. Any such function is called a *version* of the conditional probability. (See any of the standard references cited above for further details.)

If $\mathcal{G} = \mu(X)$, the σ-algebra generated by the random variable X, then we may write $P_{A||X=x}(\omega)$ or $P_{A||X}(\omega)$ in place of $P_{A||\mathcal{G}}(\omega)$.

It is not always the case that, if we let A vary, while holding ω fixed, we get a proper probability measure. Indeed, there are pathological examples where there are no conditional probability measures, though there are conditional probability functions. A conditional probability function which is a measure for all ω is said to be *regular*. If a regular conditional probability uses as its conditioning σ-algebra that generated by a random variable X, we write $P(\cdot|X = x)$, as usual.

B.3.1.1 Conditional Expectation

As well as conditional probabilities, we shall need conditional expectations. Their definition is completely analogous to Definition 68. The expectation of the random variable X conditional on the σ-subalgebra $\mathcal{G}$, denoted $\mathbf{E}\left\{X||\mathcal{G}\right\}$ is an integrable, $\mathcal{G}$-measurable random variable such that $\int_G \mathbf{E}\left\{X||\mathcal{G}\right\} dP = \int_G X dP$ for all $G \in \mathcal{G}$. Conditional probabilities are, of course, the conditional expectations of indicator functions. There is another important relationship between conditional probability and conditional expectation, which we give in the form of another proposition.

Proposition 8 (Coarsening Conditional Probability) *(Billingsley 1979; Doob 1953; Loéve 1955; Rao 1993) Consider any two σ-subalgebras $\mathcal{G}$ and $\mathcal{H}$, with $\mathcal{G} \subset \mathcal{H}$. Then*

$$\mathrm{P}_{A||\mathcal{G}}(\omega) \quad = \quad \mathbf{E}\left\{\mathrm{P}_{A||\mathcal{H}}||\mathcal{G}\right\}(\omega) \text{ almost surely (a.s.)}, \tag{B.3}$$

where we have been explicit about the conditional expectation's dependence on ω.

B.3.1.2 Conditional Independence

Let $\mathcal{G}$ be the conditioning σ-subalgebra, and let $\mathcal{A}$ and $\mathcal{B}$ be two other σ-subalgebras. Then $\mathcal{A}$ and $\mathcal{B}$ are *conditionally independent*, given $\mathcal{G}$, just when, for any pair of events A, B, $A \in \mathcal{A}$ and $B \in \mathcal{B}$, $P_{AB||\mathcal{G}}(\omega) = P_{A||\mathcal{G}}(\omega) P_{B||\mathcal{G}}(\omega)$ a.s.

Take any two σ-algebras over the same set, $\mathcal{A}$ and $\mathcal{B}$; their product, $\mathcal{AB}$, is the σ-algebra generated by the sets of the form $a \cap b$, where $a \in \mathcal{A}$ and $b \in \mathcal{B}$.

Proposition 9 *(Rao 1993, sec. 2.5) $\mathcal{A}$ and $\mathcal{B}$ are conditionally independent given $\mathcal{G}$ iff, for all $B \in \mathcal{B}$, $P_{B||\mathcal{A}G}(\omega) = P_{B||\mathcal{G}}(\omega)$ a.e., where $\mathcal{A}G$ is defined as above. This is also true if $\mathcal{A}$ and $\mathcal{B}$ are interchanged.*

Remark. Assuming regularity of conditional probability, this is equivalent to saying that the random variables Y and Z are independent given X if and only if

$$P(Z \in A|X = x, Y = y) \quad = \quad P(Z \in A|X = x) \tag{B.4}$$

Proposition 10 *(Loéve 1955, p. 351) Assuming regularity of conditional probability, for any three random variables*

$$\begin{aligned} & P(Z \in A, Y = y|X = x) \\ & \quad = \quad P(Z \in A|Y = y, X = x) P(Y = y|X = x) \end{aligned} \tag{B.5}$$

Lemma 38 *Let $\mathcal{A} = \mu(X)$, and $\mathcal{B} = \mu(f(X))$, for a measurable, nonrandom function f. Then $\mathcal{A}B = \mu(X, f(X)) = \mathcal{A} = \mu(X)$.*

Proof. Since f is measurable, every element of $\mathcal{B}$ is an element of $\mathcal{A}$, though not necessarily the reverse. Since $\mathcal{A}$ is a σ-algebra, it is closed under intersection. Therefore $\mathcal{A}B \subseteq \mathcal{A}$. But for every $a \in \mathcal{A}$, we can find a $b \in \mathcal{B}$ such that $a \subseteq b$, and $a \cap b = a$. Thus $\mathcal{A} \subseteq \mathcal{A}B$. Hence $\mathcal{A} = \mathcal{A}B$. QED.

Lemma 39 *Let f be a measurable, nonrandom function of the random variable X. Then*

$$P_{A||X,f(X)}(\omega) \quad = \quad P_{A||X}(\omega) \text{ a.e. }, \tag{B.6}$$

Proof. By Lemma 38, the conditioning σ-algebras on the left and right hand sides are the same. QED.

B.3.2 Restatements and Proofs of the Main Results

We begin by restating the definition of causal equivalence, and so of causal states, in terms adapted to abstract conditional probabilities. We then go through the results of Section 4.2 in order and, where necessary, give alternate proofs of them. (Where new proofs are not needed, we say so.)

B.3.2.1 Definition of Causal States

For us, Ω is the space of two-sided infinite strings over $\mathcal{A}$; $\mathcal{F}$ is the σ-algebra generated by cylinders over such strings; and the probability measure P is simply P (Definition 9).

What we want to do is condition on histories; so we make our conditioning σ-subalgebra $\mu(\overleftarrow{S})$, following the usual convention that $\mu(X)$ is the σ-algebra induced by the random variable X. This contains all finite-length histories, and even all semi-infinite histories, as events. Similarly, designate the σ-subalgebra for futures by $\mu(\overrightarrow{S})$. We want there to be a function $\mathrm{P}_{F||\mu(\overleftarrow{S})}(\omega)$, at least when $F \in \mu(\overrightarrow{S})$; and we want this to be a probability measure over $\mu(\overrightarrow{S})$, for fixed ω.

As we have seen (Proposition 7), the conditional probability function exists. Moreover, it is regular, since $\mu(\overleftarrow{S})$ is a subalgebra of the σ-algebra of cylinder sets, and S_t always takes its values from a fixed, finite set (Doob 1953; Rao 1993).

Thus, we *do* have a random variable $\mathrm{P}_{F||\overleftarrow{S}=\overleftarrow{s}}(\omega)$, which is the probability of the set $F \in \mu(\overrightarrow{S})$, given that $\overleftarrow{S}=\overleftarrow{s}$. We now define causal equivalence thus: $\overleftarrow{s} \sim_\epsilon \overleftarrow{s}'$ iff, for P-almost all pairs ω, ω', if $\omega \in \overleftarrow{s}$ and $\omega' \in \overleftarrow{s}'$, then $\mathrm{P}_{F||\overleftarrow{S}=\overleftarrow{s}}(\omega) = \mathrm{P}_{F||\overleftarrow{S}=\overleftarrow{s}'}(\omega')$, for all $F \in \mu(\overrightarrow{S})$. (It is clear that this is an equivalence relation — in particular, that it is transitive.)

It may be comforting to point out (following Upper (1997, sec. 2.5)) that the functions $\mathrm{P}_{F||\mu(\overleftarrow{S}^L)}(\omega)$, i.e., the probabilities of the fixed future event F conditional on longer and longer histories, almost always converge on $\mathrm{P}_{F||\mu(\overleftarrow{S})}(\omega)$. This is because of the martingale convergence theorem of Doob (Doob 1953, Theorem VII.4.3). For each L, $\mu(\overleftarrow{S}^L) \subset \mu(\overleftarrow{S}^{L+1})$ and the smallest σ-algebra containing them all is $\mu(\overleftarrow{S})$. Thus, for any random variable X with $\mathbf{E}\left\{|X|\right\} < \infty$, $\lim_{L\to\infty} \mathbf{E}\left\{X||\mu(\overleftarrow{S}^L)\right\} = \mathbf{E}\left\{X||\mu(\overleftarrow{S})\right\}$ almost surely. Applied to the indicator function 1_F of the future event F, this gives the desired convergence.

Note that if we want only causal equivalence for a finite future, matters are even simpler. Since for finite L every event in $\mu(\overrightarrow{S}^L)$ consists of the union of a finite number of disjoint elementary events (i.e., of a finite number of length-L futures), it suffices if the conditional probability assignments agree for the individual futures. If they agree for every finite L, then we have the alternate definition (Eq. 4.11) of causal states.

B.3.2.2 Measurability of ϵ

At several points, we need ϵ to be a measurable function, i.e., we need $\mu(\mathcal{S}) \subseteq \mu(\overleftarrow{S})$. This is certainly the case for processes that can be represented as Markov chains, stochastic deterministic finite automata, or conventional hidden Markov models generally. The strongest general result yet obtained is that ϵ is, so to speak, *nearly measurable*.

Proposition 11 *(Upper 1997, Prop. 2.5.3) For each causal state $\mathcal{S}_i$, the set $\epsilon^{-1}(\mathcal{S}_i)$ of histories mapping to $\mathcal{S}_i$ is either measurable or the intersection of a measurable set and a set of full measure.*

Thus, each $\epsilon^{-1}(\mathcal{S}_i)$ differs from a measurable set in $\mu(\overleftarrow{S})$ by at most a subset of a set of measure zero. This is close enough to complete measurability for our purposes, and we will speak of ϵ *as though* it were always measurable. Finding necessary and sufficient conditions on the process for ϵ to be measurable is an interesting problem.

B.3.2.3 The Morph

We wish to show that the morph of a causal state is well-defined, i.e., that the distribution of futures conditional on the entire history is the same as the distribution conditional on the causal state. Start with the fact that, since $\mathcal{S} = \epsilon(\overleftarrow{S})$, and ϵ is nearly measurable, $\mu(\mathcal{S}) \subseteq \mu(\overleftarrow{S})$. This lets us use Proposition 8, and see that $\mathrm{P}_{F||\mathcal{S}=\mathcal{S}_i}(\omega)$ is the expectation of $\mathrm{P}_{F||\overleftarrow{S}=\overleftarrow{s}}(\omega)$ over those $\omega \in \mathcal{S}_i$. But, by the construction of causal states, $\mathrm{P}_{F||\overleftarrow{S}=\overleftarrow{s}}(\omega)$ has the same value for P-almost-all ω. Thus $\mathrm{P}(F|\mathcal{S} = \mathcal{S}_i) = \mathrm{P}(F|\overleftarrow{S}=\overleftarrow{s})$ for (almost every) $\overleftarrow{s} \in \mathcal{S}_i$. (We can always find versions of the conditional probabilities which eliminate the "almost-all" and the "almost every" above.) So, since this works for arbitrary future events F, it works in general, and we may say that the distribution of futures is the same whether we condition on the past or on the causal state.

B.3.2.4 Existence of the Conditional Entropy of Futures

As we have seen, $\mathrm{P}_{\overrightarrow{S}^L||\overleftarrow{S}}(\omega)$ is a probability measure over a finite set, so (Gray 1990, sec. 5.5) we define the entropy of length-L futures conditional on a particular history $\overleftarrow{s}$ as

$$\begin{aligned} & H[\overrightarrow{S}^L|\overleftarrow{S}=\overleftarrow{s}] \\ \equiv & -\sum_{\{s^L\}} \mathrm{P}(\overrightarrow{S}^L = s^L|\overleftarrow{S}=\overleftarrow{s}) \log_2 \mathrm{P}(\overrightarrow{S}^L = s^L|\overleftarrow{S}=\overleftarrow{s}) \ , \end{aligned} \tag{B.7}$$

with the understanding that we omit futures of conditional probability zero from the sum. This is measurable, since $\mathrm{P}(\overrightarrow{S}^L = s^L|\overleftarrow{S}=\overleftarrow{s})$ is $\mu(\overleftarrow{S})$-measurable for each s^L. Now set

$$H[\overrightarrow{S}^L|\overleftarrow{S}] \equiv \int H[\overrightarrow{S}^L|\overleftarrow{S}=\overleftarrow{s}]d\mathrm{P}_{\overleftarrow{S}} \ , \tag{B.8}$$

where $\mathrm{P}_{\overleftarrow{S}}$ is the restriction of P to $\mu(\overleftarrow{S})$. (Measurability tells us that the integral exists.)

The procedure for $H[\overrightarrow{S}^L|\mathcal{R}]$ is similar, but if anything even less problematic.

Note that we do not need to re-do the derivations of Sections 4.3 and 4.4, since those simply exploit standard inequalities of information theory, which certainly apply to the conditional entropies we have just defined. (Cf. (Billingsley 1965; Gray 1990).)

B.3.2.5 The Labeled Transition Probabilities

Recall that we defined the labeled transition probability $T_{ij}^{(s)}$ as the probability of the joint event $\mathcal{S}' = \mathcal{S}_j$ and $\overrightarrow{S}^1 = s$, conditional on $\mathcal{S} = \mathcal{S}_i$. Clearly (Proposition 7), the existence of such conditional probabilities is not at issue, nor, as we have seen, is their regularity. We can thus leave Definition 14 alone.

B.4 Alternate Proof of the Refinement Lemma (Lemma 12)

The proof of Lemma 12 carries through verbally, but we do not wish to leave loop-holes. Unfortunately, this means introducing two new bits of mathematics.

First of all, we need the largest classes that are strictly homogeneous (Definition 6) with respect to $\overrightarrow{S}^L$ for fixed L; these are, so to speak, truncations of the causal states. Accordingly, we will talk about $\mathcal{S}^L$ and σ^L, which are analogous to $\mathcal{S}$ and σ. We will also need to define the function $\phi^L_{\sigma\rho} \equiv \mathrm{P}(\mathcal{S}^L = \sigma^L|\mathcal{R} = \rho)$.

Putting these together, for every L we have

$$H[\overrightarrow{S}^L|\mathcal{R} = \rho] = H[\sum_{\{\sigma^L\}} \phi^L_{\sigma\rho}\mathrm{P}(\overrightarrow{S}^L|\mathcal{S}^L = \sigma^L)] \tag{B.9}$$

$$\geq \sum_{\{\sigma^L\}} \phi^L_{\sigma\rho} H[\overrightarrow{S}^L|\mathcal{S}^L = \sigma^L] \; . \tag{B.10}$$

Thus,

$$H[\overrightarrow{S}^L \quad | \quad \mathcal{R}] = \sum_{\{\rho\}} \mathrm{P}(\mathcal{R} = \rho) H[\overrightarrow{S}^L|\mathcal{R} = \rho] \tag{B.11}$$

$$\geq \sum_{\{\rho\}} \mathrm{P}(\mathcal{R} = \rho) \sum_{\{\sigma^L\}} \phi^L_{\sigma\rho} H[\overrightarrow{S}^L|\mathcal{S}^L = \sigma^L] \tag{B.12}$$

$$= \sum_{\{\sigma^L,\rho\}} \mathrm{P}(\mathcal{R} = \rho)\phi^L_{\sigma\rho} H[\overrightarrow{S}^L|\mathcal{S}^L = \sigma^L] \tag{B.13}$$

$$= \sum_{\{\sigma^L,\rho\}} \mathrm{P}(\mathcal{S}^L = \sigma^L, \mathcal{R} = \rho) H[\overrightarrow{S}^L|\mathcal{S}^L = \sigma^L] \tag{B.14}$$

$$= \sum_{\{\sigma^L\}} \mathrm{P}(\mathcal{S}^L = \sigma^L) H[\overrightarrow{S}\ |\mathcal{S}^L = \sigma^L] \tag{B.15}$$

$$= H[\overrightarrow{S}^L|\mathcal{S}^L] \; . \tag{B.16}$$

That is to say,

$$H[\overrightarrow{S}^L|\mathcal{R}] \quad \geq \quad H[\overrightarrow{S}^L|\mathcal{S}^L] \; , \tag{B.17}$$

with equality if and only if every $\phi^L_{\sigma\rho}$ is either 0 or 1. Thus, if $H[\overrightarrow{S}^L|\widehat{\mathcal{R}}] = H[\overrightarrow{S}\ |\mathcal{S}^L]$, every $\widehat{\rho}$ is entirely contained within some σ^L; except for possible subsets of measure 0. But if this is true for every L — which, in the case of a prescient rival $\widehat{\mathcal{R}}$, it is — then every $\widehat{\rho}$ is at least weakly homogeneous (Definition 7) with respect to all $\overrightarrow{S}^L$. Thus, by Lemma 7, all its members, except for that same subset of measure 0, belong to the same causal state. QED.

B.5 Finite Entropy for the Semi-Infinite Future

While cases where $H[\overrightarrow{S}]$ is finite — more exactly, where $\lim_{L\to\infty} H[\overrightarrow{S}^L]$ exists and is finite — may be uninteresting for information-theorists, they are of great interest to physicists, since they correspond, among other things, to periodic and limit-cycle behaviors. There are, however, only two substantial differences between what is true of the infinite-entropy processes considered in the main body of the development and the finite-entropy case.

First, we can simply replace statements of the form "for all L, $H[\overrightarrow{S}^L] \ldots$ " with $H[\overrightarrow{S}]$. For example, the optimal prediction theorem (Theorem 5) for finite-entropy processes becomes for all $\boldsymbol{\mathcal{R}}$, $H[\overrightarrow{S}|\mathcal{R}] \geq H[\overrightarrow{S}|\mathcal{S}]$. The details of the proofs are, however, entirely analogous.

Second, we can prove a substantially stronger version of the Control Theorem (Theorem 11).

Theorem 25 (The Finite-Control Theorem) *For all prescient rivals* $\widehat{\boldsymbol{\mathcal{R}}}$,

$$H[\overrightarrow{S}] - H[\overrightarrow{S}|\widehat{\mathcal{R}}] \leq C_\mu \ . \tag{B.18}$$

Proof. By a direct application of Eq. A.20 and the definition of mutual information, Eq. A.10, we have that

$$H[\overrightarrow{S}] - H[\overrightarrow{S}|\mathcal{S}] \leq H[\mathcal{S}] \ . \tag{B.19}$$

But, by the definition of prescient rivals (Definition 17), $H[\overrightarrow{S}|\mathcal{S}] = H[\overrightarrow{S}|\widehat{\mathcal{R}}]$, and, by definition, $C_\mu = H[\mathcal{S}]$. Substituting equals for equals gives us the theorem. QED.

Appendix C

Proof of Lemma 20, Chapter 9

Recall the statement of Lemma 20

If a domain Λ^i has a periodic phase, then the domain is periodic, and the spatial periodicities $S(\Lambda^i_j)$ of all its phases $\Lambda^i_j, j = 0, \ldots, p-1$, are equal.

Proof. The proof consists of two parts. First, and most importantly, it is proved that the spatial periodicities of the temporal phases of a periodic domain Λ^i cannot increase and that the periodicity of one phase implies the periodicity of all its successors. Then it follows straightforwardly that the spatial periodicities have to be equal for all temporal phases and that they all must be periodic.

Our proof employs the update transducer T_ϕ, which is simply the FST which scans across a lattice configuration and outputs the effect of applying the CA update rule ϕ to it. For reasons of space, we refrain from giving full details on this operator — see rather (Hanson 1993). Here we need the following results. If ϕ is a binary, radius-r CA, the update transducer has 2^{2r} states, representing the 2^{2r} distinct contexts (words of previously read symbols) in which T_ϕ scans new sites, and we customarily label the states by these context words. The effect of applying the CA ϕ to a set of lattice configuration represented by the DFA M is a new machine, given by $T_\phi M$ — the "direct product" of the machines M and T_ϕ. Once again, for reasons of space, we will not explain how this direct product works in the general case. We are interested merely in the special case where $M = \Lambda^i_j$, the j^{th}, periodic phase of a domain, with spatial period n. The next phase of the domain, Λ^i_{j+1}, is the composed automaton $T_\phi M$, *once the latter has been minimized.* Before the latter step $T_\phi M$ consists of n "copies" of the FST T_ϕ, one for each of Λ^i_j's n states. There are no transitions within a copy. Transitions from copy k to copy k' occur only if $k' = k + 1 \pmod{n}$. In total, there are $n2^{2r}$ states in the direct composition.

$T_\phi M$ is finite and deterministic, but far from minimal. We are interested in its minimal equivalent machine, since that is what we have defined as the representative of the next phase of the domain. The key to our proof is an unproblematic part of the minimization, namely, removing states that have no predecessors (i.e., no incoming transitions) and so are never reached. (Recall that, by hypothesis, we are examining successive phases of a domain, all represented by strongly connected process graphs.) It can be shown, using the techniques in Hanson (1993), that if the transition from state k in Λ^i_j to state $k+1$ occurs on a 0 (respectively, on a 1), then in the composed machine, the transitions from copy k of T_ϕ only go to those states in copy $k+1$ whose context string ends in a 0 (respectively, in a 1). Since states in copy $k+1$ can be reached only from states in copy k, it follows that half of the states in each copy cannot be reached at all, and so they can be eliminated without loss.

Now, this procedure of eliminating states without direct predecessors in turn leaves some states in copy $k+2$ without predecessors. So we can re-apply the procedure, and once again, it will remove half of the remaining states. This is because applying it twice is the same as removing those states in copy $k+2$ for which the last two symbols in the context word differ from the symbols connecting state k to state $k+1$ and state $k+1$ to state $k+2$ in the original domain machine Λ^i_j.

What this procedure does is exploit the fact that, in a domain, every state is encountered only in a unique update-scanning context; we are eliminating combinations of domain-state and update-transducer-state that simply cannot be reached. Observe that we can apply this procedure exactly $2r$ times, since that suffices to establish the complete scanning context, and each time we do so, we eliminate half the remaining states. We are left then with $n2^{2r}/2^{2r} = n$ states after this process of successive halvings. Further observe that, since each state k of the original domain machine Λ^i_j occurs in *some* scanning context, we will never eliminate *all* the states in copy k. Since each of the n copies has at least one state left in it, and there are only n states remaining after the halvings are done, it follows that each copy contains exactly one state, which has one incoming transition, from the previous copy, and one outgoing transition, to the next copy. The result of eliminating unreachable states, therefore, is a machine of n states which is not just deterministic but (as we have defined the term) periodic. Note, however, that this is not *necessarily* the minimal machine, since we have not gone through a complete minimization procedure, merely the easy part of one. Λ^i_{j+1} thus might have fewer than n states, but certainly no more.

To sum up: We have established that, if Λ^i_j is a periodic domain phase, then Λ^i_{j+1} is also periodic and $S(\Lambda^i_{j+1}) \leq S(\Lambda^i_j)$. Thus, for any t, $S(\mathbf{\Phi}^t\Lambda^i_j) \leq S(\Lambda^i_j)$. But $\mathbf{\Phi}^t\Lambda^i_j) = \Lambda^i_{(j+t)\,\mathrm{mod}\,p}$ and if $t = p$, we have $\Lambda^i_{(j+t)\,\mathrm{mod}\,p} = \Lambda^i_{(j+p)\,\mathrm{mod}\,p} = \Lambda^i_j$. Putting these together we have

$$S(\Lambda^i_{j+1}) \leq S(\Lambda^i_j) \Rightarrow S(\Lambda^i_{j+1}) = S(\Lambda^i_j) \ , \tag{C.1}$$

for $j = 0, 1, \ldots, p-1$. This implies that the spatial period is the same, namely n, for all phases of the domain. And this proves the proposition when the CA alphabet is binary.

The reader may easily check that a completely parallel argument holds if the CA alphabet is not binary but $m-$ary, substituting m for 2 and $(m-1)/m$ for $1/2$ in the appropriate places. QED.

Appendix D

Prescience, Sufficiency, and Optimal Prediction

D.1 More General Notions of Predictive Power

In the preceding argument, we measured the predictive power of a class of effective states by how much they reduced the entropy of the outputs. Thinking of entropy as effective variability or uncertainty, this is not an unreasonable measure of ability to predict, but in many applications it is customary and/or sensible to use other measures, and, in any case, it would be naturally to be a little leery of the causal states if they were optimal *only* as measured by conditional entropy. It is for this reason that we have paid such attention to the statistical concept of sufficiency, since it lets us establish the predictive optimality of the causal states in a much more general sense.

Before we can do that, we need to introduce some concepts from statistical decision theory (Blackwell and Girshick 1954; Luce and Raiffa 1957; Lehmann and Casella 1998).

D.2 Decision Problems and Optimal Strategies

Definition 69 (A Decision Problem) *A decision problem consists of the pair $\Omega, \mathbf{A}$, where Ω is a random variable (ranging over Ω) and $\mathbf{A}$ is some set. Ω is called the* sample *or* state of nature*, and is supposed to represent data, observations, experimental results, etc. An $a \in \mathbf{A}$ is called an* action*, and the elements of $\mathbf{A}$ are supposed to represent different possible responses to the information about the world represented in Ω.*

Definition 70 (A Pure Strategy) *A* pure strategy *is a function specifying a unique action for each state of nature, $f : \Omega \mapsto \mathbf{A}$. If $f(\omega_1) = f(\omega_2)$ whenever ω_1 and ω_2 are in the same cell of a partition $\mathcal{Z}$, we say that the f* depends on *the corresponding random variable Z.*

Definition 71 (A Randomized Strategy) *A* randomized strategy *ϕ is a random function from states of nature to actions. We write the probability of taking action a given sample ω, under strategy ϕ, as $\mathrm{P}_\phi(A = a|\Omega = \omega)$. If $\mathrm{P}_\phi(A = a|\Omega = \omega_1) = \mathrm{P}_\phi(A = a|\Omega = \omega_2)$, for all a, whenever ω_1 and ω_2 are in the same cell of the partition $\mathcal{Z}$, we say that ϕ depends on the corresponding random variable Z.*

Given a set of randomized strategies, we can construct a set of pure strategies such that each randomized strategy picks a pure strategy at random. Hence the name "pure strategy".

Definition 72 (Utility of a Strategy) *The* utility *of a strategy ϕ, is a functional from ϕ's conditional distribution of actions to the non-negative real numbers, parameterized by the state of nature: $\mathcal{L}(\phi, \omega)$. It is often by not necessarily written in terms of a* loss function *defined for each action, $L : \mathbf{A} \times \Omega \mapsto \mathbf{R}^+$.*

Remark 1. Perhaps the two most common utility functionals are mean loss and maximum loss. *Remark 2.* Some authors prefer to make large values of utility preferable to small ones; no matter of principle is involved.

Definition 73 (Dominance among Strategies) *Strategy ϕ* dominates *ψ when $\mathcal{L}(\phi, \omega) \leq \mathcal{L}(\psi, \omega)$, for all $\omega \in \Omega$.*

Definition 74 (Optimal Strategies) *If a strategy ϕ dominates all other strategies ψ, then ϕ is called an* optimal strategy.

Remark. If the utility functional is the mean loss, then the optimal strategy is said to be a *Bayes* strategy, or a *Bayes optimal* strategy, or simply to maximize expected utility. If the utility functional is the maximum loss, then the optimal strategy is said to be *minimax*.

Definition 75 (Behaviorally Equivalent Strategies) *Two randomized strategies ϕ_1, ϕ_2 are behaviorally equivalent if and only if they lead to the same distribution of actions conditional on the state of nature, i.e., iff $\mathrm{P}_{\phi_1}(A = a|\Omega = \omega) = \mathrm{P}_{\phi_2}(A = a|\Omega = \omega)$ for all a, ω.*

Definition 76 (Behaviorally Equivalent Strategy Sets) *Two sets Φ_1, Φ_2 is randomized strategies are behaviorally equivalent iff each strategy in Φ_1 is behaviorally equivalent to at least one strategy in Φ_2, and vice versa.*

Remark. This is the same as the definition of "equally informative" strategies and statistics in Blackwell and Girshick (1954, Def. 8.3.1). We avoid the use of the word "informative" here, since we do not want to have to explain the relationship between this concept and those of information theory.

Proposition 12 (Strategies Based on Sufficient Statistics) *Given a set Φ of randomized strategies which are functions of the state of nature, and a sufficient statistic Z on Ω, there is a behaviorally equivalent set of randomized strategies Ψ, where each $\psi \in \Psi$ depends only on Z. Conversely, if Z is a statistic, and for any arbitrary set of randomized strategies Φ depending on Ω it is possible construct a set Ψ of randomized strategies depending only on Z which is behaviorally equivalent to Φ, then Z is a sufficient statistic.*

This is proved by Blackwell and Girshick (1954) as their Theorem 8.3.2. Their proof is constructive. (See their p. 218 for the converse part of the theorem.) See also Lehmann and Casella (1998, ch. 1).

Remark. The gist of the lemma is that, whatever behavior you might want to get from strategies which are sensitive to the whole of the state of nature, or to some arbitrary partition over it, you can get the same behavior using strategies which are sensitive only to a sufficient statistic.

D.3 Prediction Problems and Optimal Prediction

Definition 77 (A Prediction Problem) *Let* **A** *be the set of future behaviors of which the system is capable. Then a strategy is a (possibly random) mapping from present data to future events, i.e., a prediction method. Let Ω be the space of possible observations-to-date, and suppose that there is an optimal prediction method, ϕ_{opt} which is a (possibly random) function of Ω. Then we say that the decision problem is a* prediction problem.

Remark. The essential parts of the definition are that (1) the "state of nature" is a record of past observations — in the case of memoryless transduction, the current input to the transducer — and (2) there *is* an optimal predictor based on that data.

Lemma 40 (General Predictive Optimality of Prescient States) *Let ϕ_{opt} be the optimal predictor for a given prediction problem, and let $\widehat{\boldsymbol{\mathcal{R}}}$ be a class of prescient states for that process. Then there is a behaviorally equivalent, and so equally optimal, predictor, $\psi_{\widehat{\boldsymbol{\mathcal{R}}}}$, which depends only on $\widehat{\mathcal{R}}$.*

Proof. We simply apply Proposition 12 with $\Phi = \{\phi_{\mathrm{opt}}\}$.

Lemma 41 (Generally Predictively Optimal States Are Prescient) *If, for any prediction problem on $\mathcal{P}$, one can construct an optimal predictor which depends only on $\mathcal{R}$, then $\boldsymbol{\mathcal{R}}$ is prescient.*

Proof. We simply consider the prediction problem implicit in the preceding development, the optimal (minimal-conditional-entropy) solutions to which all, by construction, involve prescient states. By hypothesis, we can make an optimal predictor, in this sense, using $\boldsymbol{\mathcal{R}}$, so $\boldsymbol{\mathcal{R}}$ must be prescient.

Theorem 26 ***(General Predictive Optimality and Minimality of Causal States)*** *The causal states are generally predictively optimal, and if $\boldsymbol{\mathcal{R}}$ is generally predictively optimal, then it is a refinement almost everywhere of $\boldsymbol{\mathcal{S}}$.*

Proof. The first part of the theorem, the predictive optimality of the causal states, is a direct application of Lemma 40, since the causal states are prescient. Second, we know from Lemma 41 that generally-predictively-optimal states are prescient, and from the Refinement Lemma (12) that prescient states are refinements a.e. of the causal states. Or, put differently, the generally-predictive states are sufficient statistics, and the causal states are the minimal sufficient statistics, so the second part of the theorem follows from Lemma 3 (in its various avatars) as well.

Bibliography

Abbott, Laurence F. and Terrence J. Sejnowski (eds.) (1998). *Neural Codes and Distributed Representations: Foundations of Neural Computation*, Cambridge, Massachusetts. MIT Press.

Ablowitz, Mark J., Martin D. Kruskal and J. F. Ladik (1979). "Solitary Wave Collisons." *SIAM Journal on Applied Mathematics*, **36**: 428–437.

Agre, Philip E. (1997). *Computation and Human Experience*. Learning in Doing: Social, Cognitive, and Computational Perspectives. Cambridge, England: Cambridge University Press.

Aizawa, Y., I. Nishikawa and Kunihiko Kaneko (1991). "Soliton Turbulence in Cellular Automata." In (Gutowitz 1991), pp. 307–327. Also published as *Physica D* **45** (1990), nos. 1–3.

Akaike, Hirotugu (1998). *Selected Papers of Hirotugu Akaike*. Springer Series in Statistics: Perspectives in Statistics. Berlin: Springer-Verlag. Edited by Emanuel Parzen, Kunio Tanabe and Genshiro Kitagawa.

al Ghazali, Abu Hamid Muhammad ibn Muhammad at-Tusi (1100/1997). *The Incoherence of the Philosophers = Tahafut al-falasifah: A Parallel English-Arabic Text*. Islamic translation series. Provo, Utah: Brigham Young University Press. Translated by Michael E. Marmura.

Algoet, Paul (1992). "Universal Schemes for Prediction, Gambling and Portfolio Selection." *The Annals of Probability*, **20**: 901–941. See also an important Correction, *The Annals of Probability*, **23** (1995): 474–478.

Alhazred, Abdul (c. 750). *Kitab al-Azif*. al-Iraq. Latin translation by Olaus Wormius, as *Necronomicon*, printed Spain, 1623; as cited in (Lovecraft 1929, p. 132).

Anas, Alex, Richard Arnott and Kenneth A. Small (1998). "Urban Spatial Structures." *Journal of Economic Literature*, **36**: 1426–1464.

Anderson, Philip W. and Daniel L. Stein (1987). "Broken Symmetry, Emergent Properties, Disspiative Structures, Life: Are They Related?" In *Self-Organizing Systems: The Emergence of Order* (F. Eugene Yates, ed.), pp. 445–457. New York: Plenum Press.

Andre, David, Forrest H. Bennett, III and John R. Koza (1997). "Evolution of Intricate Long-Distance Communication Signals in Cellular Automata Using Genetic Programming." In *Artificial Life V* (Christopher G. Langton and Katsunori Shimohara, eds.), pp. 513–520. Cambridge, Massachusetts: MIT Press.

Andrienko, Yu. A., N. V. Brilliantov and J. Kurths (2000). "Complexity of Two-Dimensional Patterns." *European Physical Journal B*, **15**: 539–546.

Anonymous (T'ang Dynasty). "Kuan Yin Tzu." Written in China during the T'ang dynasty. Partial translation in Joseph Needham, Science and Civilisation in China, vol. II (Cambridge University Press, 1956), p. 73.

Apostolico, Alberto and Gill Bejerano (2000). "Optimal Amnesic Probabilistic Automata, or, How to Learn and Classify Proteins in Linear Time and Space." In *RECOMB 2000: Proceedings of the 4th Annual Conference on Research in Computational Biology*, pp. 25–32. New York: Association for Computing Machinery Press.

Arnold, Dirk (1996). "Information-theoretic Analysis of Phase Transitions." *Complex Systems*, **10**: 143–155.

Arrow, Kenneth (1974). *The Limits of Organization*. Fels Lectures on Public Policy Analysis. New York: W. W. Norton.

Arthur, Wallace (1990). *The Green Machine: Ecology and the Balance of Nature*. Oxford: Basil Blackwell.

Ashby, W. Ross (1956). *An Introduction to Cybernetics*. London: Chapman and Hall.

— (1960). *Design for a Brain: The Origins of Adaptive Behavior*. London: Chapman and Hall, 2nd edn. First edition, 1956.

— (1962). "Principles of the Self-Organizing System." In (Von Foerester and Zopf Jr 1962), pp. 255–278.

Auyang, Sunny Y. (1998). *Foundations of Complex-System Theories: In Economics, Evolutionary Biology, and Statistical Physics*. Cambridge, England: Cambridge University Press.

Badii, Remo and Antonio Politi (1997). *Complexity: Hierarchical Structures and Scaling in Physics*, vol. 6 of *Cambridge Nonlinear Science Series*. Cambridge: Cambridge University Press.

Bak, Per (1996). *How Nature Works: The Science of Self-Organized Criticality*. New York: Copernicus.

Bak, Per, C. Tang and Kurt Weisenfield (1987). "Self-Organized Criticality: An explanation of 1/f noise." *Physical Review Letters*, **59**: 381–384.

Baldi, Pierre and Søren Brunak (1998). *Bioinformatics: The Machine Learning Approach*. Adaptive Computation and Machine Learning. Cambridge, Massachusetts: MIT Press.

Balian, Roger (1991). *From Microphysics to Macrophysics: Methods and Applications of Statistical Physics*. Berlin: Springer-Verlag.

Ball, Philip (1999). *The Self-Made Tapestry: Pattern Formation in Nature*. Oxford: Oxford University Press.

Banks, Stephen P. (1990). *Signal Processing, Image Processing, and Pattern Recognition*. New York: Prentice Hall.

Barabási, Albert-László and H. Eugene Stanley (1995). *Fractal Concepts in Surface Growth*. Cambridge, England: Cambridge University Press.

Bateson, Gregory (1979). *Mind and Nature: A Necessary Unity*. New York: E. P. Dutton.

Bennett, Charles H. (1985). "Dissipation, Information, Computational Complexity and the Definition of Organization." In (Pines 1985), pp. 215–234.

— (1986). "On the Nature and Origin of Complexity in Discrete, Homogeneous Locally-Interacting Systems." *Foundations of Physics*, **16**: 585–592.

— (1990). "How to Define Complexity in Physics, and Why." In (Zurek 1990), pp. 137–148.

Berlekamp, Elwyn R., John H. Conway and Richard K. Guy (1982). *Winning Ways for your Mathematical Plays*. New York: Academic Press.

Bialek, William and Naftali Tishby (1999). "Predictive Information." E-print, arxiv.org, cond-mat/9902341.

Bickel, P. J. and Y. Ritov (1995). *Inference in Hidden Markov Models I: Local Asymptotic Normality in the Stationary Case*. Tech. Rep. 383, Statistics Department, University of California-Berkeley. URL `http://www.stat.berkeley.edu/tech-reports/383.abstract`.

Billingsley, Patrick (1961). *Statistical Inference for Markov Processes*, vol. 2 of *Statistical Research Monographs*. Chicago: University of Chicago Press.

— (1965). *Ergodic Theory and Information*. Tracts on Probablity and Mathematical Statistics. New York: Wiley.

— (1979). *Probability and Measure*. Wiley Series in Probability and Mathematical Statistics. New York: Wiley.

Binder, P.-M. and Nicolás Perry (2000). "Comment II on 'Simple Measure for Complexity'." *Physical Review E*, **62**: 2998–2999.

Blackwell, David and M. A. Girshick (1954). *Theory of Games and Statistical Decisions*. New York: Wiley. Reprinted New York: Dover Books, 1979.

Blackwell, David and Lambert Koopmans (1957). "On the Identifiability Problem for Functions of Finite Markov Chains." *Annals of Mathematical Statistics*, **28**: 1011–1015.

Blum, Lenore, Michael Shub and Steven Smale (1989). "On a theory of computation and complexity over the real numbers: NP-completeness, recursive functions and universal machines." *Bulletin of the American Mathematical Society*, **21**: 1–46.

Boccara, Nino (1993). "Transformations of One-Dimensional Cellular Automaton Rules by Translation-invariant Local Surjective Mappings." *Physica D*, **68**: 416–426.

Boccara, Nino, J. Nasser and M. Roger (1991). "Particle-like Structures and Their Interactions in Spatio-Temporal Patterns Generated by One-Dimensional Deterministic Cellular Automaton Rules." *Physical Review A*, **44**: 866 – 875.

Boccara, Nino and M. Roger (1991). "Block Transformations of One-Dimensional Deterministic Cellular Automata." *Journal of Physics A*, **24**: 1849 – 1865.

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

Bohren, Craig F. and B. A. Albrecht (1998). *Atmospheric Thermodynamics*. New York: Oxford University Press.

Booth, Taylor L. (1967). *Sequential Machines and Automata Theory*. New York: Wiley.

Borges, Jorge Luis (1964). *Other Inquisitions, 1937–1952*. Austin: University of Texas Press. Trans. Ruth L. C. Simms.

Bosq, Denis (1998). *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction*, vol. 110 of *Lecture Notes in Statistics*. Berlin: Springer-Verlag, 2nd edn.

Bower, James M. and Hamid Bolouri (eds.) (2001). *Computational Modeling of Genetic and Biochemical Networks*, Computational Molecular Biology, Cambridge, Massachusetts. MIT Press.

Bucy, Richard S. (1994). *Lectures on Discrete Time Filtering*. Signal Processing and Digital Filtering. Berlin: Springer-Verlag.

Bühlmann, Peter and Abraham J. Wyner (1999). "Variable Length Markov Chains." *Annals of Statistics*, **27**: 480–513. URL `http://www.stat.berkeley.edu/tech-reports/479.abstract1`.

Bunge, Mario (1959). *Causality: The Place of the Causal Princple in Modern Science*. Cambridge, Massachusetts: Harvard University Press. Reprinted as *Causality and Modern Science*, NY: Dover Books, 1979.

Burks, Arthur W. (ed.) (1970). *Essays on Cellular Automata*, Urbana. University of Illinois Press.

Burton, David M. (1976). *Elementary Number Theory*. Boston: Allyn and Bacon.

Bush, Robert R. and Frederick Mosteller (1955). *Stochastic Models for Learning*. New York: Wiley.

Bussemaker, Harmen J., Hao Li and Eric D. Siggia (2000). "Building a Dictionary for Genomes: Identification of Presumptive Regulatory Sites by Statistical Analysis." *Proceedings of the National Academy of Sciences USA*, **97**: 10,096–10,100. URL `http://www.pnas.org/cgi/doi/10/1073/pnas.180265397`.

Caines, Peter E. (1988). *Linear Stochastic Systems*. New York: Wiley.

Calvino, Italo (1965/1968). *Cosmicomics*. New York: Harcourt Brace Jovanovich. Translated by William Weaver from *Le cosmicomiche* (Turin: Giulio Einaudi).

Carraway, Kermit L. and Coralie A. Carothers Carraway (eds.) (2000). *Cytoskeleton: Signalling and Cell Regulation*. The Practical Approach Series. Oxford: Oxford University Press, 2nd edn.

Carroll, John and Darrell Long (1989). *Theory of Finite Automata: With an Introduction to Formal Languages*. Englewood Cliffs, New Jersey: Prentice-Hall.

Chaikin, Paul M. and T. C. Lubensky (1995). *Principles of Condensed Matter Physics*. Cambridge, England: Cambridge University Press.

Chaitin, Gregory (1966). "On the Length of Programs for Computing Finite Binary Sequences." *Journal of the Association for Computing Machinery*, **13**: 547–569.

Charniak, Eugene (1993). *Statistical Language Learning*. Language, Speech and Communication. Cambridge, Massachusetts: MIT Press.

Chomsky, Noam (1956). "Three Models for the Description of Language." *IRE Transactions on Information Theory*, **2**: 113.

— (1957). *Syntactic Structures*, vol. 4 of *Janua linguarum, series minor*. The Hauge: Mouton.

Chopard, Bastien and Michel Droz (1998). *Cellular Automata Modeling of Physical Systems*, vol. 6 of *Collection Aléa Saclay*. Cambridge, England: Cambridge University Press.

Chorin, Alexander J. (1994). *Vorticity and Turbulence*, vol. 103 of *Applied Mathematical Sciences*. Berlin: Springer Verlag.

Clarke, Keitch C., Leonard Gaydos and Stacy Hoppen (1996). "A Self-modifying Cellular Automaton Model of Historical Urbanization in the San Francisco Bay Area." *Environment and Planning B*, **24**: 247–261.

Collings, Peter J. (1990). *Liquid Crystals: Nature's Delicate Phase of Matter*. Princeton, New Jersey: Princeton University Press.

Conant, Roger (1974). "W. Ross Ashby (1903–1972)." *International Journal of General Systems*, **1**: 4–7.

Cover, Thomas M. and Joy A. Thomas (1991). *Elements of Information Theory*. New York: Wiley.

Cramér, Harald (1945). *Mathematical Methods of Statistics*. Uppsala: Almqvist and Wiksells. Republished by Princeton University Press, 1946, as vol. 9 in the Princeton Mathematics Series, and as a paperback, in the Princeton Landmarks in Mathematics and Physics series, 1999.

Cressie, Noel A. C. (1993). *Statistics for Spatial Data*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. New York: Wiley, revised edn.

Cross, Mark C. and Pierre Hohenberg (1993). "Pattern Formation Out of Equilibrium." *Reviews of Modern Physics*, **65**: 851–1112.

Crutchfield, James P. (1990). "Information and Its Metric." In *Nonlinear Structures in Physical Systems—Pattern Formation, Chaos and Waves* (L. Lam and H. C. Morris, eds.), p. 119. New York: Springer-Verlag.

— (1992). "Semantics and Thermodynamics." In *Nonlinear Modeling and Forecasting* (Martin Casdagli and Stephen Eubank, eds.), vol. 12 of *Santa Fe Institute Studies in the Sciences of Complexity*, pp. 317–359. Reading, Massachusetts: Addison-Wesley.

— (1994a). "The Calculi of Emergence: Computation, Dynamics, and Induction." *Physica D*, **75**: 11–54.

— (1994b). "Is Anything Ever New? Considering Emergence." In *Complexity: Metaphors, Models, and Reality* (G. Cowan and D. Pines and D. Melzner, eds.), vol. 19 of *Santa Fe Institute Studies in the Sciences of Complexity*, pp. 479–497. Reading, Massachusetts: Addison-Wesley.

Crutchfield, James P. and Christopher Douglas (1999). "Imagined Complexity: Learning a Random Process." Manuscript in preparation.

Crutchfield, James P. and David P. Feldman (1997). "Statistical Complexity of Simple One-Dimensional Spin Systems." *Physical Review E*, **55**: 1239R–1243R.

— (2001a). "Regularities Unseen, Randomness Observed: Levels of Entropy Convergence." *Physical Review E*, **submitted**. E-print, arxiv.org, cond-mat/0102181.

— (2001b). "Synchronizing to the Environment: Information Theoretic Constraints on Agent Learning." *Chaos*, **submitted**. E-print, arxiv.org, nlin.AO/0103038.

Crutchfield, James P., David P. Feldman and Cosma Rohilla Shalizi (2000a). "Comment I on 'Simple Measure for Complexity'." *Physical Review E*, **62**: 2996–2997. E-print, arxiv.org, chao-dyn/9907001.

Crutchfield, James P. and James E. Hanson (1993a). "Attractor Vicinity Decay for a Cellular Automaton." *Chaos*, **3**: 215–224.

— (1993b). "Turbulent Pattern Bases for Cellular Automata." *Physica D*, **69**: 279–301.

Crutchfield, James P., Wim Hordijk and Melanie Mitchell (2000b). "Computational Performance of Evolved Cellular Automata: Parts I and II." Manuscript in preparation.

Crutchfield, James P. and Bruce S. McNamara (1987). "Equations of Motion from a Data Series." *Complex Systems*, **1**: 417–452.

Crutchfield, James P. and Melanie Mitchell (1995). "The Evolution of Emergent Computation." *Proceedings of the National Academy of Sciences*, **92**: 10,742–10,746.

Crutchfield, James P. and Norman H. Packard (1983). "Symbolic dynamics of noisy chaos." *Physica D*, **7**: 201–223.

Crutchfield, James P. and Cosma Rohilla Shalizi (1999). "Thermodynamic Depth of Causal States: Objective Complexity via Minimal Representations." *Physical Review E*, **59**: 275–283. E-print, arxiv.org, cond-mat/9808147.

— (2001). "Intrinsic Computation versus Engineered Computation: The Computational Mechanics of Rule 110." Manuscript in preparation.

Crutchfield, James P. and Karl Young (1989). "Inferring Statistical Complexity." *Physical Review Letters*, **63**: 105–108.

— (1990). "Computation at the Onset of Chaos." In (Zurek 1990), pp. 223–269.

Das, Rajarshi (1996). *The Evolution of Emergent Computation in Cellular Automata*. Ph.D. thesis, Colorado State University.

Das, Rajarshi, Melanie Mitchell and James P. Crutchfield (1994). "A Genetic Algorithm Discovers Particle Computation in Cellular Automata." In *Proceedings of the Conference on Parallel Problem Solving in Nature — PPSN III* (Y. Davidor and H.-P. Schwefel and R. Manner, eds.), Lecture Notes in Computer Science, pp. 344–353. Berlin: Springer-Verlag.

Dawes, Robyn M., David Faust and Paul E. Meehl (1989). "Clinical Versus Actuarial Judgment." *Science*, **243**: 1668–1674.

Dayan, Peter and Laurence F. Abbott (2001). *Theoretical Neuroscience*. Cambridge, Massachusetts: MIT Press. URL `http://play.ccs.brandeis.edu/abbott/book/`.

de Gennes, Pierre-Gilles and Jacques Prost (1993). *The Physics of Liquid Crystals*, vol. 83 of *International Series of Monographys on Physics*. Oxford: Clarendon Press, 2nd edn.

Debreu, Gerard (1959). *Theory of Value: An Axiomatic Analysis of Economic Equilibrium*, vol. 17 of *Cowles Foundation Monographs*. New York: Wiley.

Delgado, Jordi and Ricard V. Solé (1997). "Collective-induced Computation." *Physical Review E*, **55**: 2338–2344.

Dennett, Daniel C. (1991). "Real Patterns." *Journal of Philosophy*, **88**: 27–51. Reprinted in (Dennett 1997).

— (1997). *Brainchildren: Essays on Designing Minds*. Representation and Mind. Cambridge, Massachusetts: MIT Press.

Descartes, René (1637). *Discours de la Méthode pour bien conduire sa raison, et chercher la vérité dans les sciences*. Leiden. Translated, as *Discoures on the Method of rightly conducting one's reason and seeking the truth in the sciences*, in (Descartes 1985, vol. I, pp. 111–151).

— (1664). *Le Monde de M. Descartes ou le Traité de la Lumière*. Paris. Written 1629–33 but published posthumously; partially translated, as *The World or Treatise on Light*, in (Descartes 1985, vol. I, pp. 81–98).

— (1985). *The Philosophical Writings of Descartes*. Cambridge, England: Cambridge University Press. 2 vols. Translated by John Cottingham, Robert Stoothoff and Duglad Murdoch.

Dimitrov, Alexander G. and John P. Miller (2001). "Neural Coding and Decoding: Communication Channels and Quantization." *Network: Computation in Nervous Systems*, **submitted**. URL `http://cns.montana.edu/~alex/publications/coding.pdf`.

Doob, Joseph L. (1953). *Stochastic Processes*. Wiley Publications in Statistics. New York: Wiley.

Dorfman, Jay Robert (1998). *An Introduction to Chaos in Nonequilibrium Statistical Mechanics*, vol. 14 of *Cambridge Lecture Notes in Physics*. Cambridge, England: Cambridge University Press.

D'Souza, Raissa M. and Norman H. Margolus (1999). "Thermodynamically Reversible Generalization of Diffusion Limited Aggregation." *Physical Review E*, **60**: 264–274. E-print, arxiv.org, cond-mat/9810258.

Edmonds, Bruce H. (1997). "Hypertext Bibliography of Measures of Complexity." URL `http://www.cpm.mmu.ac.uk/~bruce/combib/`.

Egan, Greg (1994). *Permutation City*. London: Millennium.

Eggertsson, Thráinn (1990). *Economic Behavior and Institutions*. Cambridge Surveys of Economic Literature. Cambridge, England: Cambridge University Press.

Eigen, Manfred and Peter Schuster (1979). *The Hypercycle: A Principle of Natural Self-Organization*. Berlin: Springer-Verlag.

Elliot, R. J., L. Aggoun and J. B. Moore (1995). *Hidden Markov Models: Estimation and Control*, vol. 29 of *Applications of Mathematics*. New York: Springer.

Ellis, Richard S. (1985). *Entropy, Large Deviations, and Statistical Mechanics*, vol. 271 of *Grundlehren der mathematischen Wissenschaften*. Berlin: Springer-Verlag.

— (1999). "The Theory of Large Deviations: from Boltzmann's 1877 Calculation to Equilibrium Macrostates in 2D Turbulence." *Physica D*, **133**: 106–136.

Eloranta, Kari (1993). "Partially Permutive Cellular Automata." *Nonlinearity*, **6**: 1009–1023.

— (1994). "The Dynamics of Defect Ensembles in One-Dimensional Cellular Automata." *Journal of Statistical Physics*, **76**: 1377–1398.

Eloranta, Kari and E. Nummelin (1992). "The Kink of Cellular Automaton Rule 18 Performs a Random Walk." *Journal of Statistical Physics*, **69**: 1131–1136.

Elster, Jon (1989a). *The Cement of Society: A Study of Social Order*. Studies in Rationality and Social Change. Cambridge, England: Cambridge University Press.

— (1989b). *Nuts and Bolts for the Social Sciences*. Cambridge, England: Cambridge University Press.

Eppstein, David (ongoing). "Gliders in Life-Like Cellular Automata." Interactive online database of two-dimensional cellular automata rules with gliders. URL `http://fano.ics.uci.edu/ca/`.

Epstein, Joshua M. (1999). "Agent-based Computational Models and Generative Social Science." *Complexity*, **4(5)**: 41–60.

Eriksson, Karl-Erik and Kristian Lindgren (1987). *Entropy and Correlations in Lattice Systems*. Tech. rep., Physical Resource Theory Group, Chalmers University, Göteborg, Sweden.

Eskin, Eleazar, William Noble Grundy and Yoram Singer (2000). "Protein Family Classification Using Spare Markov Transducers." In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology* (Philip Bourne and Michael Gribskov and Russ Altman and Nancy Jensen and Debra Hope and Thomas Lengauer and Julie Mitchell and Eric Scheeff and Chris Smith and Shawn Strande and Helge Weissig, eds.), pp. 134–145. Menlo Park, California: American Association for Artificial Intelligence Press.

Evans, William, Sridhar Rajagopalan and Umesh Vazirani (1993). "Choosing a Reliable Hypothesis." In *Proceedings of the 6th Workshop on Computational Learning Theory*, pp. 269–276. New York: ACM Press.

Farley, B. G. and W. A. Clark (1954). "Simulation of Self-Organizing Systems by Digital Computer." *Transactions of the Institute of Radio Engineers*, **Professional Group on Information Theory (PGIT) 4**: 76–84.

Farmer, J. Doyne, Alan Lapedes, Norman Packard and Burton Wendroff (eds.) (1986). *Evolution, Games, and Learning: Models for Adaptation in Machines and Nature: Proceedings of the Fifth Annual International Conference of the Center for Nonlinear Studies, Los Alamos, New Mexico 87545, USA, May 20-24, 1985*, Amsterdam. North-Holland. Also published as *Physica D* **22**.

Fedorova, Antonina N. and Michael G. Zeitlin (2000). "Localized Coherent Structures and Patterns Formation in Collective Models of Beam Motion." In *Proceedings of the Capri ICFA Workshop, October, 2000*. E-print, arxiv.org, physics/0101007.

Feixas, Miquel, Esteve del Acebo, Philippe Bekaert and Mateu Sbert (1999). "An Informathion Theory Framework for the Analysis of Scene Complexity." *Europgraphics*, **18**: C–95–C–106.

Feldman, David P. (1998). *Computational Mechanics of Classical Spin Systems*. Ph.D. thesis, University of California, Davis. URL `http://hornacek.coa.edu/dave/Thesis/thesis.html`.

Feldman, David P. and James P. Crutchfield (1998a). "Discovering Non-Critical Organization: Statistical Mechanical, Information Theoretic, and Computational Views of Patterns in Simple One-Dimensional Spin Systems." *Journal of Statistical Physics*, **submitted**. URL `http://www.santafe.edu/projects/CompMech/papers/DNCO.html`. Santa Fe Institute Working Paper 98-04-026.

— (1998b). "Measures of Statistical Complexity: Why?" *Physics Letters A*, **238**: 244–252.

Fell, David (1997). *Understanding the Control of Metabolism*, vol. 2 of *Frontiers in Metabolism*. London: Portland Press.

Feng, Jin and Thomas G. Kurtz (2000). "Large Deviations for Stochastic Processes." On-line manuscript. URL `http://www.math.wisc.edu/~kurtz/feng/ldp.htm`.

Fischer, K. H. and J. A. Hertz (1988). *Spin Glasses*. Cambridge Studies in Magnetism. Cambridge: Cambridge University Press.

Flake, Gary William (1998). *The Computational Beauty of Nature: Computer Explorations of Fractals, Chaos, Complex Systems and Adaptation*. Cambridge, Massachusetts: MIT Press.

Fontana, Walter and Leo W. Buss (1994). ""The Arrival of the Fitest": Towards a Theory of Biological Organization." *Bulletin of Mathematical Biology*, **56**: 1–64. URL `http://www.santafe.edu/~walter/publications.html`.

Forrest, Stehpanie (ed.) (1990). *Emergent Computation: Self-Organizing, Collective, and Cooperative Phenomena in Natural and Artificial Computing Networks: Proceedings of the Ninth Annual International Conference of the Center for Nonlinear Studies, Los Alamos, New Mexico 87545, USA, May 22–26, 1989*, Amsterdam. North-Holland. Also published as *Physica D* **42**.

Forster, Dieter (1975). *Hydrodynamic Fluctuations, Broken Symmetry, and Correlation Functions*, vol. 47 of *Frontiers in Physics*. Reading, Massachusetts: Benjamin Cummings.

Fox, Ronald F. (1988). *Energy and the Evolution of Life*. New York: W. H. Freeman.

Frisch, Uriel (1995). *Turbulence: The Legacy of A. N. Kolmogorov*. Cambridge, England: Cambridge University Press.

Garncarek, Z. and R. Piasecki (1999). "What Is a Physical Measure of Spatial Inhomogeneity Comparable to the Mathematical Approach?" *European Journal of Physics: Applied Physics*, **5**: 243–249.

Gaspard, Pierre (1998). *Chaos, Scattering and Statistical Mechanics*. Cambridge Nonlinear Science. Cambridge, England: Cambridge University Press.

Gel'fand, I. M. and A. M. Yaglom (1956). "Calculation of the Amount of Information About a Random Function Contained in Another Such Function." *Uspekhi Matematicheski Nauk*, **12**: 3–52. Trans. in American Mathematical Society Translations, 2nd series, 12 (1959): 199–246.

Gellner, Ernest (1974). *Legitimation of Belief*. Cambridge, England: Cambridge University Press.

Gillespie, John H. (1998). *Population Genetics: A Concise Guide*. Baltimore: Johns Hopkins University Press.

Glymour, Clark (1992). *Thinking Things Through: An Introduction to Philosophical Issues and Achievements*. Cambridge, Massachusetts: MIT Press.

Goldberg, Charles H. (1988). "Parity Filter Automata." *Complex Systems*, **2**: 91–141.

Goldsworthy, Andy (1990). *A Collaboration with Nature*. New York: Harry N. Abrams.

Gompper, Gerhard and Michael Schick (1994). *Self-assembling Amphiphilic Systems*, vol. 16 of *Phase Transitions and Critical Phenomena*. San Diego, California: Academic Press.

Gonçalves, W. M., R. D. Pinto, J. C. Sartorelli and M. J. de Oliveira (1998). "Inferring Statistical Complexity in the Dripping Faucet Experiment." *Physica A*, **257**: 385–389.

Gonick, Larry and Mark Wheelis (1991). *The Cartoon Guide to Genetics*. New York: HarperCollins, 2nd edn.

Goodenough, Ursula (1998). *The Sacred Depths of Nature*. Oxford: Oxford University Press.

Graham, Norma Van Surdam (1989). *Visual Pattern Analyzers*, vol. 16 of *Oxford Psychology Series*. Oxford: Oxford University Press.

Grassberger, Peter (1983). "New Mechanism for Deterministic Diffusion." *Physical Review A*, **28**: 3666–7.

— (1986). "Toward a Quantitative Theory of Self-Generated Complexity." *International Journal of Theoretical Physics*, **25**: 907–938.

Gray, Robert M. (1990). *Entropy and Information Theory*. New York: Springer-Verlag.

Grenander, Ulf (1996). *Elements of Pattern Theory*. Johns Hopkins Studies in the Mathematical Sciences. Baltimore, Maryland: Johns Hopkins University Press.

Grenander, Ulf, Y. Chow and D. M. Keenan (1991). *Hands: A Pattern Theoretic Study of Biological Shapes*, vol. 2 of *Research Notes in Neural Computing*. New York: Springer-Verlag.

Grenander, Ulf and K. Manbeck (1993). "A Stochastic Shape and Color Model for Defect Detection in Potatoes." *American Statistical Association*, **2**: 131–151.

Griffeath, David (1979). *Additive and Cancellative Interacting Particle Systems*, vol. 724 of *Lecture Notes in Mathematics*. Berlin: Springer-Verlag.

Griffeath, David and Cristopher Moore (eds.) (forthcoming). *New Constructions in Cellular Automata*, Oxford. Oxford University Press. Proceedings of the Constructive Cellular Automata Theory conference, Santa Fe Institute, Santa Fe, New Mexico, November 1998.

Gurzadyan, V. G. (1999). "Kolmogorov Complexity as a Descriptor of Cosmic Microwave Background Maps." *Europhysics Letters*, **46**: 114–117. E-print, arxiv.org, astro-phy/9902123.

Gutowitz, Howard (ed.) (1991). *Cellular Automata: Theory and Experiment*, Cambridge, Massachusetts. MIT Press. Also published as *Physica D* **45** (1990), nos. 1–3.

Guttorp, Peter (1995). *Stochastic Modeling of Scientific Data*. Stochastic Modeling. London: Chapman and Hall.

Guyon, Xavier (1995). *Random Fields on a Network: Modeling, Statistics, and Applications*. Springer Series in Statistics: Probability and Its Applications. Berlin: Springer-Verlag.

Haken, Herman (1977). *Synergetics: An introduction: Nonequilibrium Phase Transitions and Self-Organization in Physics, Chemistry, and Biology*. Berlin: Springer-Verlag.

Hancock, John T. (1997). *Cell Signaling*. Harlow, Essex: Longman.

Hanson, James E. (1993). *Computational Mechanics of Cellular Automata*. Ph.D. thesis, University of California, Berkeley.

Hanson, James E. and James P. Crutchfield (1992). "The Attractor-Basin Portrait of a Cellular Automaton." *Journal of Statistical Phyics*, **66**: 1415–1462.

— (1997). "Computational Mechanics of Cellular Automata: An Example." *Physica D*, **103**: 169–189.

Harrison, Lionel G. (1993). *Kinetic Theory of Living Pattern*. Cambridge, England: Cambridge University Press.

Hartley, R. V. L. (1928). "Transmission of Information." *Bell System Technical Journal*, pp. 535–563.

Hartmanis, Juris and R. E. Stearns (1966). *Algebraic Structure Theory of Sequential Machines*. Prentice-Hall International Series in Applied Mathematics. Englewood Cliffs, New Jersey: Prentice-Hall.

Hebb, D. O. (1949). *The Organization of Behavior: A Neuropsychological Theory*. New York: Wiley.

Heims, Steve J. (1991). *The Cybernetics Group: Constructing a Social Science for Postwar America, 1946–1953*. Cambridge, Massachusetts: MIT Press.

Hein, Piet (1966). *Grooks*. Cambridge, Massachusetts: MIT Press.

Hinton, Geoffrey and Terrence J. Sejnowski (eds.) (1999). *Unsupervised Learning: Foundations of Neural Computation*. Computational Neuroscience. Cambridge, Massachusetts: MIT Press.

Holcombe, Mike and Ray Paton (eds.) (1998). *Information Processing in Cells and Tissues*, New York. Plenum Press.

Holland, John H. (1992). *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. Cambridge, Massachusetts: MIT Press, 2nd edn. First edition, Ann Arbor, Michigan: University of Michigan Press, 1975.

— (1998). *Emergence: From Chaos to Order*. Reading, Massachusetts: Addison-Wesley.

Holland, John H., Keith J. Holyoak, Richard E. Nisbett and Paul R. Thagard (1986). *Induction: Processes of Inference, Learning, and Discovery*. Computational Models of Cognition and Perception. Cambridge, Massachusetts: MIT Press.

Holland, Paul W. (1990). “Stochastic Models for Learning.” In *A Statistical Model: Frederick Mosteller's Contributions to Statistics, Science, and Public Policy* (Stephen E. Fienberg and David C. Hoaglin and William H. Kruskal and Judith M. Tanur, eds.), Springer Series in Statistics: Perspectives in Statistics, pp. 194–198. Berlin: Springer-Verlag.

Hollander, Myles and Douglas A. Wolfe (1999). *Nonparametric Statistical Methods*. Wiley Series in Probability and Statistics: Applied Probability and Statistic. New York: Wiley, 2nd edn.

Hopcroft, John E. and Jeffrey D. Ullman (1979). *Introduction to Automata Theory, Languages, and Computation*. Reading: Addison-Wesley. 2nd edition of *Formal Languages and Their Relation to Automata*, 1969.

Hordijk, Wim (1999). *Dynamics, Emergent Computation, and Evolution in Cellular Automata*. Ph.D. thesis, University of New Mexico, Albuquerque, New Mexico. URL `http://www.santafe.edu/projects/evca/Papers/WH-Diss.html`.

Hordijk, Wim, Melanie Mitchell and James P. Crutchfield (1998). “Mechanisms of Emergent Computation in Cellular Automata.” In *Parallel Problem Solving in Nature—PPSN V* (A. E. Eiben and T. Bäck and M. Schoenaur and H.-P. Schwefel, eds.), Lecture Notes in Computer Science, pp. 613–622. Berlin: Springer-Verlag.

Hordijk, Wim, Cosma Rohilla Shalizi and James P. Crutchfield (2001). “An Upper Bound on the Products of Particle Interactions in Cellular Automata.” *Physica D*, **154**: 240–258. E-print, arxiv.org, nlin.CG/0008038.

Huberman, Bernardo A. (1985). “Computing with Attractors: From Self-Repairing Computers to Ultradiffusion, and the Application of Dynamical Systems to Human Behavior.” In (Pines 1985), pp. 167–174.

Hume, David (1739). *A Treatise of Human Nature: Being an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects*. London: John Noon. Reprint (Oxford: Clarendon Press, 1951) of original edition, with notes and analytical index.

Husserl, Edmund (1913/1931). *Ideas: General Introduction to Pure Phenomenology*. New York: Macmillan. Translated by W. R. Boyce Gibson, from *Ideen au einer reinen Phänomenologie und phänomenologischen Philosophie*.

Hutchins, Edwin (1995). *Cognition in the Wild*. Cambridge, Massachusetts: MIT Press.

ibn Rushd (Averroës), Abu al-Walid Muhammad ibn Ahmad ibn Muhammad (1180–1185/1954). *Tahafut al-Tahafut: The Incoherence of the Incoherence*. Cambridge, England: Gibbs Memorial Trust. Translated by Simon Van Den Bergh. Original version published in Cordova, al-Andalus.

Infeld, Eryk and George Rowlands (1990). *Nonlinear Waves, Solitions, and Chaos*. Cambridge, England: Cambridge University Press.

Ito, H., S.-I. Amari and K. Kobayashi (1992). “Identifiability of Hidden Markov Information Sources and Their Minimum Degrees of Freedom.” *IEEE Transactions on Information Theory*, **38**: 324–333.

Jaeger, Herbert (2000). “Observable Operator Models for Discrete Stochastic Time Series.” *Neural Computation*, **12**: 1371–1398. URL `ftp://ftp.gmd.de/GMD/ais/publications/1999/`.

Jakubowski, Mariusz H., Ken Steiglitz and Richard Squier (1997). “Information Transfer between Solitary Waves in the Saturable Schrödinger Equation.” *Physical Review E*, **56**: 7267–7272.

Jensen, Henrik J. (1998). *Self-Organized Criticality: Emergent Complex Behavior in Physical and Biological Systems*, vol. 10 of *Cambridge Lecture Notes in Physics*. Cambridge, England: Cambridge University Press.

Kac, Mark (1959). *Statistical Independence in Probability, Analysis and Number Theory*. New York: Wiley.

Kantz, Holger and Thomas Schreiber (1997). *Nonlinear Time Series Analysis*, vol. 7 of *Cambridge Nonlinear Science Series*. Cambridge, England: Cambridge University Press.

Kearns, Michael J. and Umesh V. Vazirani (1994). *An Introduction to Computational Learning Theory*. Cambridge, Massachusetts: MIT Press.

Keizer, Joel (1987). *Statistical Thermodynamics of Nonequilibrium Processes*. New York: Springer-Verlag.

Kelly, Kevin T. (1996). *The Logic of Reliable Inquiry*, vol. 2 of *Logic and Computation in Philosophy*. Oxford: Oxford University Press.

Kemeny, John G. and J. Laurie Snell (1976). *Finite Markov Chains*. New York: Springer-Verlag.

Kemeny, John G., J. Laurie Snell and Anthony W. Knapp (1976). *Denumerable Markov Chains*. New York: Springer-Verlag, 2nd edn.

Kim, Jaegwon (1998). *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. Representation and Mind. MIT Press.

Kipling, Rudyard (1940). *Complete Verse*. New York: Doubleday, definitive edn.

Klimontovich, Yuri Lvovich (1990/1991). *Turbulent Motion and the Structure of Chaos: A New Approach to the Statistical Theory of Open Systems*. Dordrecht: Kluwer Academic. Translated by Alexander Dobroslavsky from *Turbulentnoe dvizhenie i struktura khaosa: Novyi podkhod k statisticheskoi teorii otkrytykh sistem* (Moscow: Nauka).

Klinkner, Kristina L. and Cosma Rohilla Shalizi (2001). "Extensive State Estimation: Reconstructing Causal States by Splitting." Manuscript in preparation.

Knight, Frank B. (1992). *Foundations of the Prediction Process*, vol. 1 of *Oxford Studies in Probability*. Oxford: Clarendon Press.

Kogan, Barry S. (1985). *Averroes and the Metaphysics of Causation*. Albany, New York: State University of New York Press.

Kohonen, Tuevo (1984). *Self-Organizing Maps and Associative Memory*. Berlin: Springer-Verlag.

— (2001). *Self-Organizing Maps*. Berlin: Springer-Verlag, 3rd edn.

Kolakowski, Leszek (1975). *Husserl and the Search for Certitude*. Cassirer Lectures. New Haven, Connecticut: Yale University Press.

— (1978). *Main Currents of Marxism: Its Rise, Growth and Dissolution*. Oxford: Oxford University Press. Translated by P. S. Falla from *Glowne nurty marksizmu*, otherwise unpublished.

Kolmogorov, Andrei N. (1941). "Interpolation und Extrapolation von stationären Zufälligen Folgen." *Bull. Acad. Sci. U.S.S.R., Math.*, **3**: 3–14. In Russian with German summary.

— (1965). "Three Approaches to the Quantitative Definition of Information." *Problems of Information Transmission*, **1**: 1–7.

— (1983). "Combinatorial Foundations of Information Theory and the Calculus of Probabilities." *Russ. Math. Surveys*, **38**: 29.

Koppel, Moshe (1987). "Complexity, Depth, and Sophistication." *Complex Systems*, **1**: 1087–1091.

Koppel, Moshe and Henri Atlan (1991). "An Almost Machine-Independent Theory of Program-Length Complexity, Sophistication and Induction." *Information Sciences*, **56**: 23–44.

Krauss, Gerhard (1999). *Biochemistry of Signal Transduction and Regulation*. Wienheim, Germany: Wiley-VCH.

Krohn, Wolfgang, Günther Küppers and Helga Nowotny (eds.) (1990). *Selforganization: Portrait of a Scientific Revolution*, Dordrecht. Kluwer Academic.

Krugman, Paul R. (1996). *The Self-Organizing Economy*. Oxford: Blackwell.

Kullback, Solomon (1968). *Information Theory and Statistics*. New York: Dover Books, 2nd edn. First edition New York: Wiley, 1959.

Landau, L. D. and E. M. Lifshitz (1980). *Statistical Physics*. Oxford: Pergamon Press.

Lange, Oksar and Fred M. Taylor (1938). *On the Economic Theory of Socialism*. Minneapolis: University of Minnesota Press. Collects previously published articles; ed. Benjamin E. Lippincott. Reprinted New York: McGraw-Hill, 1964.

Lauritzen, S. L. (1996). *Graphical Models*. New York: Oxford University Press.

Lehmann, E. L. and George Casella (1998). *Theory of Point Estimation*. Springer Texts in Statistics. Berlin: Springer-Verlag, 2nd edn.

Lem, Stanislaw (1968/1983). *His Master's Voice*. San Diego: Harcourt Brace Jovanovich. Translated by Michael Kandel from *Glos pana*.

Lempel, Abraham and Jacob Ziv (1976). "On the Complexity of Finite Sequences." *IEEE Transactions in Information Theory*, **IT-22**: 75–81.

— (1986). "Compression of Two-Dimensional Data." *IEEE Transactions in Information Theory*, **IT-32**: 2–8.

Levin, Leonid A. (1974). "Laws of Information Conservation (Nongrowth) and Aspects of the Foundation of Probability Theory." *Problemy Peredachi Informatsii*, **10**: 30–35. Translation: Problems of Information Transmission 10 (1974) 206–210.

Levin, S. A., T. M. Powell and J. H. Steele (eds.) (1993). *Patch Dynamics*, vol. 96 of *Lecture Notes in Biomathematics*, Berlin. Springer-Verlag.

Lewis, Harry R. and Christos H. Papadimitriou (1998). *Elements of the Theory of Computation*. Upper Saddle River, New Jersey: Prentice-Hall, 2nd edn.

Li, Ming and Paul M. B. Vitanyi (1993). *An Introduction to Kolmogorov Complexity and its Applications*. New York: Springer-Verlag.

Li, Wentien (1991). "On the Relationship between Complexity and Entropy for Markov Chains and Regular Languages." *Complex Systems*, **5**: 381–399.

Liggett, Thomas M. (1985). *Interacting Particle Systems*. Berlin: Springer-Verlag.

Lim, Jae S. (1990). *Two-Dimensional Signal and Image Processing*. New York: Prentice Hall.

Lindgren, Kristian, Cristopher Moore and Mats Nordahl (1998). "Complexity of Two-Dimensional Patterns." *Journal of Statistical Physics*, **91**: 909–951.

Lindgren, Kristian and Mats G. Nordahl (1988). "Complexity Measures and Cellular Automata." *Complex Systems*, **2**: 409–440.

— (1990). "Universal Computation in a Simple One-Dimensional Cellular Automaton." *Complex Systems*, **4**: 299–318.

Lindley, David V. (1972). *Bayesian Statistics, a Review*. Philadelphia: Society for Industrial and Applied Mathematics.

Ljapin, E. S. (1963). *Semigroups*, vol. 3 of *Translations of Mathematical Monographs*. Providence, Rhode Island: American Mathematical Society.

Lloyd, Seth and Heinz Pagels (1988). "Complexity as Thermodynamic Depth." *Annals of Physics*, **188**: 186–213.

Loehlin, John C. (1992). *Latent Variable Models: An Introduction to Factor, Path, and Structural Analysis*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 2nd edn.

Loéve, Michel (1955). *Probability Theory*. New York: D. Van Nostrand Company, 1st edn.

Loewenstein, Werner R. (1999). *The Touchstone of Life: Molecular Information, Cell Communication, and the Foundations of Life*. Oxford: Oxford University Press.

Lopez-Ruiz, R., H. L. Mancini and X. Calbet (1995). "A Statistical Measure of Complexity." *Physics Letters A*, **209**: 321–326.

Lotka, Alfred J. (1924). *Elements of Physical Biology*. Baltimore, Maryland: Williams and Wilkins. Reprinted as *Elements of Mathematical Biology*, New York: Dover Books, 1956.

Lovecraft, H. P. (April 1929). "The Dunwich Horror." *Weird Tales*. As reprinted in (Lovecraft 1997, pp. 101–174).

— (1997). *The Annotated H. P. Lovecraft*. New York: Dell. Edited by S. T. Joshi.

Luce, R. Duncan and Howard Raiffa (1957). *Games and Decisions: Introduction and Critical Survey*. New York: Wiley.

Luria, Aleksandr R. (1973). *The Working Brain: An Introduction to Neuropsychology*. New York: Basic Books.

Lwoff, Andre (1962). *Biological Order*. The Karl Taylor Compton Lectures, MIT, 1960. Cambridge, Massachusetts: MIT Press.

Manneville, Paul (1990). *Dissipative Structures and Weak Turbulence*. Boston, Massachusetts: Academic Press.

Manneville, Paul, Nino Boccara, Gérard Y. Vichniac and R. Bidaux (eds.) (1990). *Cellular Automata and Modeling of Complex Systems: Proceedings of the Winter School, Les Houches, France, February 21–28, 1989*, vol. 46 of *Springer Proceedings in Physics*, Berlin. Springer-Verlag.

Manning, Christopher D. and Hinrich Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: MIT Press.

Manrubia, Susanna C., Damian H. Zanette and Ricard V. Sole (1999). "Transient Dynamics and Scaling Phenomena in Urban Growth." *Fractals*, **7**: 1–8.

Mantegna, Rosario N. and H. Eugene Stanley (2000). *An Introduction to Econophysics: Correlations and Complexity in Finance*. Cambridge, England: Cambridge University Press.

Manzano, María (1990/1999). *Model Theory*, vol. 37 of *Oxford Logic Guides*. Oxford: Clarendon Press. Translated by Ruy J. G. B. de Quieroz from *Teoría de Modelos* (Madrid: Alianza Editorial).

March, James G. and Herbert A. Simon (1993). *Organizations*. Oxford: Blackwell, 2nd edn. First ed. New York: Wiley, 1958.

Margolus, Norman (1999). "Crystalline Computation." In *Feynman and Computation: Exploring the Limits of Computers* (Anthony J. G. Hey, ed.), pp. 267–305. Reading, Massachusetts: Perseus Books. E-print, arxiv.org, comp-gas/9811002.

Martin-Löf, P. (1966). "The Definition of Random Sequences." *Information and Control*, **9**: 602–619.

Mayo, Deborah G. (1996). *Error and the Growth of Experimental Knowledge*. Science and Its Conceptual Foundations. Chicago: University of Chicago Press.

Mayo, Deborah G. and Aris Spanos (2000). "A Post-Data Interpretation of Neyman-Pearson Methods Based on a Conception of Severe Testing." Preprint, Virginia Tech.

Mayr, Otto (1986). *Authority, Liberty, and Automatic Machinery in Early Modern Europe*, vol. 8 of *Johns Hopkins Studies in the History of Technology, new series*. Baltimore, Maryland: Johns Hopkins University Press.

McIntosh, Harold V. (2000). "Rule 110 as it Relates to the Presence of Gliders." Electronic manuscript. URL `http://delta.cs.cinvestav.mx/~mcintosh/comun/RULE110W/RULE110.html`.

Medawar, Peter B. (1982). "Herbert Spencer and General Evolution." In *Pluto's Republic*, pp. 209–227. Oxford: Oxford University Press.

Miller, James Grier (1978). *Living Systems*. New York: McGraw-Hill.

Milligan, Graeme (ed.) (1999). *Signal Transduction*. The Practical Approach Series. Oxford University Press, 2nd edn.

Minsky, Marvin (1967). *Computation: Finite and Infinite Machines*. Englewood Cliffs, New Jersey: Prentice-Hall.

Mittenthal, Jay E. and Arthur B. Baskin (eds.) (1992). *The Principles of Organization in Organisms*, vol. 13 of *Santa Fe Institute Studies in the Sciences of Complexity*, Reading, Massachusetts. Addison-Wesley.

Monod, Jacques (1970/1971). *Chance and Necessity: An Essay on the Natural Philosophy of Modern Biology*. New York: Alfred A. Knopf. Translated by Austryn Wainhouse from *Le Hasard et la Necessite* (Paris: Editions du Seuil).

Moore, Cristopher (1990). "Unpredictability and Undecidability in Dynamical Systems." *Physical Review Letters*, **64**: 2354–2357.

— (1996). "Recursion Theory on the Reals and Continuous-Time Computation." *Theoretical Computer Science*, **162**: 23–44.

— (1998). "Dynamical Recognizers: Real-Time Language Recognition by Analog Computers." *Theoretical Computer Science*, **201**: 99–136.

Moore, Edward F. (1956). "Gedanken-Experiments on Sequential Machines." In *Automata Studies* (Claude E. Shannon and John McCarthy, eds.), vol. 34 of *Annals of Mathematics Studies*, pp. 129–153. Princeton, New Jersey: Princeton University Press.

— (1970). "Machine Models of Self-Reproduction." In (Burks 1970), pp. 187–203.

Needham, Joseph (1936). *Order and Life*. Terry Lectures, Yale University, 1935. New Haven, Conneticut: Yale University Press. Reprinted with a new foreword, Cambridge, Massachusetts: MIT Press, 1968.

— (1943a). "Evolution and Thermodynamics." In (Needham 1943c), pp. 207–232. First published 1941.

— (1943b). "Integrative Levels: A Revaluation of the Idea of Progress." In (Needham 1943c), pp. 233–272. Herbert Spencer Lecture at Oxford University, 1937.

— (1943c). *Time: The Refreshing River (Essays and Addresses, 1932–1942)*. New York: Macmillan.

Nehaniv, Chrystopher L. and John L. Rhodes (1997). "Krohn-Rhodes Theory, Hierarchies, and Evolution." In *Mathematical Hierarchies and Biology: DIMACS workshop, November 13–15, 1996* (Boris Mirkin and F. R. McMorris and Fred S. Roberts and Andrey Rzhetsky, eds.), vol. 37 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*. Providence, Rhode Island: American Mathematical Society.

Neyman, Jerzy (1950). *First Course in Probability and Statistics*. New York: Henry Holt.

Nicolis, Grégoire and Ilya Prigogine (1977). *Self-Organization in Nonequilibrium Systems: From Dissipative Structures to Order through Fluctuations*. New York: John Wiley.

Ockham, William of (1964). *Philosophical Writings: A Selection, Translated, with an Introduction, by Philotheus Boehner, O.F.M., Late Professor of Philosophy, The Franciscan Institute*. Indianapolis: Bobbs-Merrill. First pub. various European cities, early 1300s.

Orponen, Pekka (1997). "A Survey of Continuous-Time Computation Theory." In *Advances in Algorithms, Languages, and Complexity* (D.-Z. Du and K.-I Ko, eds.), pp. 209–224. Dordrecht: Kluwer Academic.

Ortoleva, Peter J. (1994). *Geochemical Self-Organization*, vol. 23 of *Oxford Monographs on Geology and Geophysics*. Oxford: Oxford University Press.

Packard, Norman H., James P. Crutchfield, J. Doyne Farmer and Robert S. Shaw (1980). "Geometry from a Time Series." *Physical Review Letters*, **45**: 712–716.

Pagels, Heinz R. (1988). *The Dreams of Reason: The Computer and the Rise of the Sciences of Complexity*. New York: Simon and Schuster.

Palmer, A. Jay, C. W. Fairall and W. A. Brewer (2000). "Complexity in the Atmosphere." *IEEE Transactions on Geoscience and Remote Sensing*, **38**: 2056–2063.

Palmer, Richard (1989). "Broken Ergodicity." In *Lectures in the Sciences of Complexity* (Daniel L. Stein, ed.), vol. 1 of *Santa Fe Institute Studies in the Sciences of Complexity (Lectures)*, pp. 275–300. Redwood City, California: Addison-Wesley.

Park, James K., Kenneth Steiglitz and William P. Thurston (1986). "Soliton-like Behavior in Automata." *Physica D*, **19**: 423–432.

Peacocke, Arthur R. (1983). *An Introduction to the Physical Chemistry of Biological Organization*. Oxford: Clarendon Press.

Pearl, Judea (2000). *Causality: Models, Reasoning, and Inference*. Cambridge, England: Cambridge University Press.

Perry, Nicolás and P.-M. Binder (1999). "Finite Statistical Complexity for Sofic Systems." *Physical Review E*, **60**: 459–463.

Peyrard, M. and M. D. Kruskal (1984). "Kink Dynamics in the Highly Discrete sine-Gordon System." *Physica D*, **14**: 88–102.

Piasecki, R. (2000). "Entropic Measure of Spatial Disorder for Systems of Finite-Sized Objects." *Physica A*, **277**: 157–173.

Pierce, J. R. (1961). *Symbols, Signals, and Noise*. Harper Brothers.

Pines, David (ed.) (1985). *Emerging Syntheses in Science: Proceedings of the Founding Workshops of the Santa Fe Institute*, vol. 1 of *Santa Fe Institute Studies in the Sciences of Complexity (Proceedings)*, Santa Fe, New Mexico. Santa Fe Institute.

Pollard, David (1984). *Convergence of Stochastic Processes*. Springer Series in Statistics. New York: Springer-Verlag.

Popper, Karl R. (1945). *The Open Society and Its Enemies*. London: Routledge.

— (1960). *The Poverty of Historicism*. London: Routledge, 2nd edn. First edition, 1957.

Poundstone, William (1984). *The Recursive Universe: Cosmic Complexity and the Limits of Scientific Knowledge*. New York: William Morrow.

Press, William H., Saul A. Teukolsky, William T. Vetterling and Brian P. Flannery (1992). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge, England: Cambridge University Press, 2nd edn. URL `http://lib-www.lanl.gov/numerical/`.

Prigogine, Ilya (1980). *From Being to Becoming: Time and Complexity in the Physical Sciences*. San Francisco: W. H. Freeman.

Prigogine, Ilya and Isabelle Stegners (1979/1984). *Order Out of Chaos: Man's New Dialogue with Nature*. Bantam New Age Books. New York: Bantam. Translation and emmendation of *La Nouvelle alliance* (Paris: Gallimard).

Ptashne, Mark (1992). *A Genetic Switch: Phage λ and Higher Organisms*. Cambridge, Massachusetts: The Cell Press and Blackwell Scientific, 2nd edn.

Quine, Willard Van Orman (1961). *From a Logical Point of View: Logico-Philosophical Essays*. Cambridge, Mass.: Harvard University Press, 2nd edn. First edition, 1953.

Rao, Malempati Madhusudana (1993). *Conditional Measures and Applications*, vol. 177 of *Monographs and Textbooks in Pure and Applied Mathematics*. New York: Marcel Dekker.

Rayner, J. C. W. and D. J. Best (1989). *Smooth Tests of Goodness of Fit*. Oxford: Oxford University Press.

Resnick, Mitchel (1994). *Turtles, Termites and Traffic Jams: Explorations in Massively Parallel Microworlds*. Complex Adaptive Systems. Cambridge, Massachusetts: MIT Press.

Rhodes, John (1971). *Applications of Automata Theory and Algebra via the Mathematical Theory of Complexity to Biology, Physics, Psychology, Philosophy, Games, and Codes*. Berkeley, California: University of California.

Rieke, Fred, David Warland, Rob de Ruyter van Steveninck and William Bialek (1997). *Spikes: Exploring the Neural Code*. Computational Neuroscience. Cambridge, Massachusetts: MIT Press.

Ripley, Brian D. (1981). *Spatial Statistics*. New York: Wiley.

— (1988). *Statistical Inference for Spatial Processes*. Cambridge, England: Cambridge University Press.

— (1996). *Pattern Recognition and Neural Networks*. Cambridge, England: Cambridge University Press.

Rissanen, Jorma (1978). "Modeling by Shortest Data Description." *Automatica*, **14**: 465–471.

— (1983). "A Universal Data Compression System." *IEEE Transactions in Information Theory*, **IT-29**: 656–664.

— (1989). *Stochastic Complexity in Statistical Inquiry*. Singapore: World Scientific.

Rothman, Daniel H. and Stéphane Zaleski (1997). *Lattice-Gas Cellular Automata: Simple Models of Complex Hydrodynamics*, vol. 5 of *Collection Aléa Saclay*. Cambridge, England: Cambridge University Press.

Ruelle, David (1989). *Statistical Mechanics: Rigorous Results*. Reading, Massachusetts: Addison Wesley.

— (1999). "Smooth Dynamics and New Theoretical Ideas in Nonequilibrium Statistical Mechanics." *Journal of Statistical Physics*, **95**: 393–468. E-print, arxiv.org, chao-dyn/9812032.

Russell, Bertrand (1920). *Introduction to Mathematical Philosophy*. The Muirhead Library of Philosophy. London: George Allen and Unwin, revised edn. First edition, 1919. Reprinted New York: Dover Books, 1993.

— (1927). *The Analysis of Matter*. International Library of Philosophy, Psychology and Scientific Method. London: K. Paul Trench, Trubner and Co. Reprinted New York: Dover Books, 1954.

— (1948). *Human Knowledge: Its Scope and Limits*. New York: Simon and Schuster.

Salmon, Wesley C. (1971). *Statistical Explanation and Statistical Relevance*. Pittsburgh: University of Pittsburgh Press. With contributions by Richard C. Jeffrey and James G. Greeno.

— (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Schelling, Thomas C. (1978). *Micromotives and Macrobehavior*. New York: W. W. Norton.

Schetzen, Martin (1989). *The Volterra and Wiener Theories of Nonlinear Systems*. Malabar, Florida: Robert E. Krieger Publishing Company, 2nd edn. Reprint, with additions, of the first edition, New York: John Wiley, 1980.

Schinazi, Rinaldo B. (1999). *Classical and Spatial Stochastic Processes*. Boston: Birkhäuser.

Schrecker, Ellen, John Schrecker and Chiang Jung-feng (1987). *Mrs. Chiang's Szechwan Cookbook*. New York: Harper and Row, 2nd edn. First edition 1976.

Schutz, Bernard F. (1980). *Geometrical Methods of Mathematical Physics*. Cambridge, England: Cambridge University Press.

Selfridge, Oliver G. (1959). "Pandemonium: A Paradigm for Learning." In *The Mechanisation of Thought Processes* (D. V. Blake and A. M. Uttley, eds.), vol. 10 of *National Physical Laboratory Symposia*, pp. 511–529. London: Her Majesty's Stationery Office.

Sethna, James P. (1991). "Order Parameters, Broken Symmetry, and Topology." In *1990 Lectures in Complex Systems* (Lynn Nadel and Daniel L. Stein, eds.), vol. 3 of *Santa Fe Institute Studies in the Sciences of Complexity (Lectures)*, pp. 243–265. Redwood City, California: Addison-Wesley.

Shalizi, Cosma Rohilla (1996a). "Is the Primordial Soup Done Yet? Quantifying Self-Organization, Especially in Cellular Automata." Talk given 30 April 1996 to the Madison Chaos and Complex Systems Seminar. URL `http://www.santafe.edu/~shalizi/Self-organization/soup-done/`.

— (1996b). "Review of (Krugman 1996)." *The Bactra Review*, **11**. URL `www.santafe.edu/~shalizi/reviews/self-organizing-economy/`.

— (1998a). "Naval Collective Intelligence (Review of (Hutchins 1995))." *The Bactra Review*, **54**. URL `www.santafe.edu/~shalizi/reviews/cognition-in-the-wild/`.

— (1998b). "*Homo economicus* on the Grand Tour, or, When Is a Lizard a Good Enough Dragon for Government Work? (Review of (Eggertsson 1990))." *The Bactra Review*, **44**. URL `www.santafe.edu/~shalizi/reviews/economic-behavior-and-institutions/`.

— (1999). "A Myopic (and Sometimes Blind) Eye on the Main Chance, or, the Origins of Custom (Review of (Young 1998))." *The Bactra Review*, **78**. URL `www.santafe.edu/~shalizi/reviews/young-strategy-and-structure/`.

— (2000). "Growth, Form, Function, and Crashes." *Santa Fe Institute Bulletin*, **15(2)**: 6–11.

Shalizi, Cosma Rohilla and James P. Crutchfield (2000a). "ϵ-Transducers: The Causal States of General Functional Relationships." MS. in preparation.

— (2000b). "Information Bottlenecks, Causal States, and Statistical Relevance Bases: How to Represent Relevant Information in Memoryless Transduction." *Physical Review E*, **submitted**. E-print, arxiv.org, nlin.AO/0006025.

— (2000c). "Pattern Discovery and Computational Mechanics." *Annals of Mathematics and Artificial Intelligence*, **submitted**. E-print, arxiv.org, cs.LG/0001027.

— (2001). "Computational Mechanics: Pattern and Prediction, Structure and Simplicity." *Journal of Statistical Physics*, **forthcoming**. E-print, arxiv.org, cond-mat/9907176.

Shalizi, Cosma Rohilla, Robert A. Haslinger and James P. Crutchfield (2001). "Spatiotemporal Emergent Structures from First Principles: Causal Architectures for Spatial Processes." Manuscript in preparation.

Shalizi, Cosma Rohilla and Cristopher Moore (2001). "What Is a Macrostate? Some Notes on the Intersection of Statistical and Computational Mechanics." Manuscript in preparation.

Shannon, Claude E. (1948). "A Mathematical Theory of Communication." *Bell System Technical Journal*, **27**: 379–423. Reprinted in (Shannon and Weaver 1963).

Shannon, Claude E. and Warren Weaver (1963). *The Mathematical Theory of Communication*. Urbana, Illinois: University of Illinois Press.

Shaw, Robert (1984). *The Dripping Faucet as a Model Chaotic System*. Santa Cruz, California: Aerial Press.

Shettleworth, Sara J. (1998). *Cognition, Evolution and Behavior*. Oxford: Oxford University Press.

Shiner, J. S., M. Davison and P. T. Landsberg (1999). "Simple Measure for Complexity." *Physical Review E*, **59**: 1459–1464.

Simon, Herbert A. (1991). "Organizations and Markets." *Journal of Economic Perspectives*, **5**: 25–44.

— (1996). *The Sciences of the Artificial*. Karl Taylor Compton Lectures, MIT, 1968. Cambridge, Massachusetts: MIT Press, 3rd edn. First edition 1969.

Singer, Yoram (1997). "Adaptive Mixtures of Probabilistic Transducers." *Neural Computation*, **9**: 1711–1733.

Sinha, Sudeshna and William L. Ditto (1998). "Dynamics Based Computation." *Physical Review Letters*, **81**: 2156–2159.

Sklar, Lawrence (1993). *Physics and Chance: Philosophical Issues in the Foundations of Statistical Mechanics*. Cambridge, England: Cambridge University Press.

Smith, Mark Andrew (1994). *Cellular Automata Methods in Mathematical Physics*. Tech. Rep. 615, Laboratory for Computer Science, MIT.

Solé, Ricard V. and Bartolo Luque (1999). "Statistical Measures of Complexity for Strongly Interacting Systems." E-print, arxiv.org, adap-org/9909002.

Solomonoff, Raymond J. (1964). "A Formal Theory of Inductive Inference." *Information and Control*, **7**: 1–22 and 224–254.

Spirtes, Peter, Clark Glymour and Richard Scheines (2001). *Causation, Prediction, and Search*. Adaptive Computation and Machine Learning. Cambridge, Massachusetts: MIT Press, 2nd edn.

Spivak, Michael (1965). *Calculus on Manifolds: A Modern Approach to Classical Theorems of Advanced Calculus*. Menlo Park, California: Benjamin Cummings.

Steiglitz, Kenneth, I. Kamal and A. Watson (1988). "Embedding Computation in One-Dimensional Automata by Phase Coding Solitons." *IEEE Transactions on Computers*, **37**: 138–144.

Stengel, Robert F. (1994). *Optimal Control and Estimation*. New York: Dover. As *Stochastic Optimal Control: Theory and Application*, New York: Wiley, 1986.

Sterling, Bruce (1985). *Schismatrix*. New York: Arbor House. Reprinted in (Sterling 1996).

— (1996). *Schismatrix Plus*. New York: Ace Books.

Sternberg, Saul (1963). "Stochastic Learning Theory." In *Handbook of Mathematical Psychology* (R. Duncan Luce and Robert R. Bush and Eugene Galanter, eds.), vol. 2, pp. 1–120. New York: Wiley.

Stiglitz, Joseph E. (1994). *Whither Socialism?*. The Wicksell Lectures, 1990. Cambridge, Massachusetts: MIT Press.

Takens, Floris (1981). "Detecting Strange Attractors in Fluid Turbulence." In *Symposium on Dynamical Systems and Turbulence* (D. A. Rand and L. S. Young, eds.), vol. 898 of *Lecture Notes in Mathematics*, p. 366. Berlin: Springer-Verlag.

— (1983). "Distinguishing Deterministic and Random Systems." In *Nonlinear Dynmaics and Turbulence* (G. I. Barenblatt and G. Iooss and D. D. Joseph, eds.), Interaction of Mechanics and Mathematics. Boston: Pitman Advanced Publishing Program.

Thornton, Chris (2000). *Truth from Trash: How Learning Makes Sense*. Complex Adaptive Systems. Cambridge, Massachusetts: MIT Press.

Tilman, David and Peter Kareiva (eds.) (1997). *Spatial Ecology: The Role of Space in Population Dynamics and Interspecific Interactions*, vol. 30 of *Monographs in Population Biology*, Princeton, New Jersey. Princeton University Press.

Timmer, J., H. Rust, W. Horbelt and H. U. Voss (2000). "Parametric, Nonparametric and Parametric Modelling of a Chaotic Circuit Time Series." *Physics Letters A*, **274**: 123–134.

Tishby, Naftali, Fernando C. Pereira and William Bialek (1999). "The Information Bottleneck Method." In *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing* (B. Hajek and R. S. Sreenivas, eds.), pp. 368–377. Urbana, Illinois: University of Illinois Press. E-print, arxiv.org, physics/0004057.

Toffoli, Tommaso (1984). "Cellular Automata as an Alternative to (Rather Than an Approximation of) Differential Equations in Modeling Physics." In *Cellular Automata* (J. Doyne Farmer and Tommasso Toffoli and Stephen Wolfram, eds.), pp. 117–127. Amsterdam: North-Holland. Also published as *Physica D* **10**, nos. 1–2.

Toffoli, Tommaso and Norman Margolus (1987). *Cellular Automata Machines: A New Environment for Modeling*. MIT Press Series in Scientific Computation. Cambridge, Massachusetts: MIT Press.

Tou, Julius T. and Rafael C. Gonzalez (1974). *Pattern Recognition Principles*. Reading, Mass: Addison-Wesley.

Touchette, Hugo and Seth Lloyd (1999). "Information-Theoretic Limits of Control." *Physical Review Letters*, **84**: 1156–1159.

Upper, Daniel R. (1997). *Theory and Algorithms for Hidden Markov Models and Generalized Hidden Markov Models*. Ph.D. thesis, University of California, Berkeley. URL `http://www.santafe.edu/projects/CompMech/`.

Valiant, Leslie G. (1984). "A Theory of the Learnable." *Communications of the Association for Computing Machinery*, **27**: 1134–1142.

van de Geer, Sara (2000). *Empirical Processes in M-Estimation*, vol. 4 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge, England: Cambridge University Press.

van der Heyden, Marcel J., Cees G. C. Diks, Bart P. T. Hoekstra and Jacob DeGoede (1998). "Testing the Order of Discrete Markov Chains Using Surrogate Data." *Physica D*, **117**: 299–313.

Van Dyke, Milton (1982). *An Album of Fluid Motion*. Stanford, California: Parabolic Press.

Vapnik, Vladimir N. (1979/1982). *Estimation of Dependences Based on Empirical Data*. Springer Series in Statistics. Berlin: Springer-Verlag. Translated by Samuel Kotz from *Vosstanovlyenie Zavicimosteĭ po Émpiricheckim Dannim* (Moscow: Nauka).

— (2000). *The Nature of Statistical Learning Theory*. Statistics for Engineering and Information Science. Berlin: Springer-Verlag, 2nd edn.

Vartanian, Aram (1953). *Diderot and Descartes: A Study of Scientific Naturalism in the Enlightenment*, vol. 6 of *History of Ideas*. Princeton, New Jersey: Princeton University Press.

Verdu, Sergio (1994). "The Development of Information Theory." In *From Statistical Physics to Statistical Inference and Back* (Peter Grassberger and Jean-Pierre Nadal, eds.), vol. 428 of *NATO Science Series C: Mathematical and Physical Sciences*, pp. 155–168. Dordrecht: Kluwer Academic.

Vicsek, Tamás (1989). *Fractal Growth Phenomena*. Singapore: World Scientific.

Victor, Jonathan D. and P. Johannesma (1986). "Maximum-Entropy Approximations of Stochastic Nonlinear Transductions: An Extension of the Wiener Theory." *Biological Cybernetics*, **54**: 289–300.

Vitányi, Paul M. B. and Ming Li (1999). "Minimum Description Length Induction, Bayesianism, and Kolmogorov Complexity." E-print, arxiv.org, cs.LG/9901014.

Von Foerester, Heinz and George W. Zopf Jr (eds.) (1962). *Principles of Self-Organization: Transactions of the University of Illinois Symposium on Self-Organization, Robert Allerton Park, 8 and 9 June, 1961*, New York. Information Systems Branch, U.S. Office of Naval Research, Pergamon Press.

von Mises, Richard (1928/1981). *Probability, Statistics and Truth*. New York: Dover. Translated by J. Neyman, D. Sholl, and E. Rabinowitsch, first published as *Wahrscheinlichkeit, Statistik, und Wahrheit* (Vienna: J. Springer).

von Neumann, John (1966). *Theory of Self-Reproducing Automata*. Urbana: University of Illinois Press. Edited and completed by Arthur W. Burks.

Wahba, Grace (1990). *Spline Models for Observational Data*. Philadelphia: Society for Industrial and Applied Mathematics.

Watts, Duncan J. (1999). *Small Worlds: The Dynamics of Networks between Order and Randomness*. Princeton Studies in Complexity. Princeton, New Jersey: Princeton University Press.

Weiss, Benjamin (1973). "Subshifts of Finite Type and Sofic Systems." *Monatshefte für Mathematik*, **77**: 462–474.

Weiss, Sholom M. and Nitin Indurkhya (1998). *Predictive Data Mining: A Practical Guide*. San Francisco: Morgan Kaufmann.

White, Roger and Guy Engelen (1993). "Cellular Automata and Fractal Urban Form: A Cellular Modelling Approach to the Evolution of Urban Land-use Patterns." *Environment and Planning A*, **25**: 1175–1199.

Whitehead, Alfred North and Bertrand Russell (1925–27). *Principia Mathematica*. Cambridge, England: Cambridge University Press, 2nd edn.

Wiener, Norbert (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications*. Cambridge, Massachusetts: The Technology Press of the Massachusetts Institute of Technology. "First published during the war [1942] as a classifed report to Section D_2, National Defense Research Council".

— (1954). *The Human Use of Human Beings: Cybernetics and Society*. Garden City, New York: Doubleday, 2nd edn. Republished London: Free Association Books, 1989; first ed. Boston: Houghton Mifflin, 1950.

— (1958). *Nonlinear Problems in Random Theory*. Cambridge, Massachusetts: The Technology Press of the Massachusetts Institute of Technology.

— (1961). *Cybernetics: Or, Control and Communication in the Animal and the Machine*. Cambridge, Massachusetts: MIT Press, 2nd edn. First edition New York: Wiley, 1948.

Williamson, Oliver E. (1975). *Markets and Hierarchies: Analysis and Antitrust Implications*. New York: Free Press.

Winfree, Arthur T. (1980). *The Geometry of Biological Time*. Berlin: Springer-Verlag.

— (1987). *When Time Breaks Down: The Three-Dimensional Dynamics of Electrochemical Waves and Cardiac Arrhythmias*. Princeton: Princeton University Press.

Witt, A., A. Neiman and J. Kurths (1997). "Characterizing the Dynamics of Stochastic Bistable Systems by Measures of Complexity." *Physical Review E*, **55**: 5050–5059.

Wolfram, Stephen (1983). "Statistical Mechanics of Cellular Automata." *Reviews of Modern Physics*, **55**: 601–644. Reprinted in (Wolfram 1994).

— (1984a). "Computation Theory of Cellular Automata." *Communications in Mathematical Physics*, **96**: 15–57. Reprinted in (Wolfram 1994).

— (1984b). "Universality and Complexity in Cellular Automata." *Physica D*, **10**: 1–35. Reprinted in (Wolfram 1994).

Wolfram, Stephen (ed.) (1986). *Theory and Applications of Cellular Automata*. Singapore: World Scientific.

Wolfram, Stephen (1994). *Cellular Automata and Complexity: Collected Papers*. Reading, Massachusetts: Addison-Wesley. URL `http://www.stephenwolfram.com/publications/books/ca-reprint/`.

Wuensche, Andrew and Mike Lesser (1992). *The Global Dynamics of Cellular Automata: An Atlas of Basin of Attraction Fields of One-Dimensional Cellular Automata*, vol. 1 of *Santa Fe Institute Studies in the Sciences of Complexity (Reference)*. Reading, Massachusetts: Addison-Wesley.

Yeomans, Julia M. (1992). *Statistical Mechanics of Phase Transitions*. Oxford: Clarendon Press.

Young, H. Peyton (1998). *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton: Princeton University Press.

Young, Karl (1991). *The Grammar and Statistical Mechanics of Complex Physical Systems*. Ph.D. thesis, University of California, Santa Cruz.

Young, Karl and James P. Crutchfield (1993). "Fluctuation Spectroscopy." *Chaos, Solitons, and Fractals*, **4**: 5–39.

Yovits, Marshall C. and Scott Cameron (eds.) (1960). *Self-Organizing Systems: Proceedings of an Interdisciplinary Conference, 5 and 6 May, 1959*, vol. 2 of *International Tracts in Computer Science and Technology and Their Application*, Oxford. Pergamon Press.

Yunes, J. B. (1994). "Seven-State Solutions to the Firing Squad Synchronization Problem." *Theoretical Computer Science*, **127**: 313–332.

Zallen, Richard (1983). *The Physics of Amorphous Solids*. New York: Wiley.

Zapranis, Achilleas and Apostolos-Paul Refenes (1999). *Principles of Neural Model Identification, Selection and Adequacy: With Applications to Financial Econometrics*. Perspectives in Neural Computing. London: Springer-Verlag.

Ziv, Jacob and Abraham Lempel (1977). "A Universal Algorithm for Sequential Data Compression." *IEEE Transactions in Information Theory*, **IT-23**: 337–343.

Zurek, Wojciech H. (ed.) (1990). *Complexity, Entropy, and the Physics of Information*, vol. 8 of *Santa Fe Institute Studies in the Sciences of Complexity*, Reading, Massachusetts. Addison-Wesley.

Acknowledgements

Thanks in the first place go to the professors whose decision to advise me in this ill-conceived venture testifies to an abundance of compassion rather than a surfeit of judgement: Jim Crutchfield (Santa Fe Institute), David Griffeath (University of Wisconsin-Madison Math) and Martin Olsson (UW-Madison Physics). Due to circumstances beyound our control, Jim's name cannot go on this dissertation as that of my adviser, though that is what he was. All of the research reported here was done under his supervision, and almost of all of it in collaboration with him, over the course of three and a half years at SFI. He has been a model adviser, patient, encouraging, absurdly tolerant of my views and foibles, and genuinely collaborative. I am also grateful to Profs. J. C. Sprott, Charles Goebel, Ian Dobson and M. Deric Bownds of UW-Madison and Dr. Norman Margolus of MIT and Boston University for encouragement, advice, training and aid.

The money for this came from a variety of Santa Fe Institute programs ("Adaptive Computation", "Computation, Dynamics and Inference", "Dynamics of Learning" and "Network Dynamics"), via grants and contracts from the Defense Advanced Research Projects Agency (F30602-00-2-0583), Intel Corporation, the National Science Foundation (PHY-9970157 and HS-970583), the Office of Naval Research (N00014-95-1-0875) and Sandia National Laboratory (AU-4978). Additional money came from an research assistantship under David Griffeath in the spring of 1997, and from numerous unwitting students taking introductory physics at Madison. Scallion cakes, dry-fried beef and other necessary sustance came from Schrecker, Schrecker and Jung-feng (1987). Review copies of many neat and useful books came from the MIT, Oxford University, University of Chicago and Yale University Presses. Aural stimuli were provided by Tori Amos, Autechre, the Clash, Max Roach, the Reptile Palace Orchestra, the Talking Heads, and the CD collections of Dave Albers, Alex Lancaster and Aryaman Shalizi. (This implies an obvious debt to my office-mates, gratefully acknowledged.) Solidarity came from AFT local #3220.

SFI has been an extraordinarily pleasurable, stimulating and productive place to work, and I really ought to thank the entire community for making it that way. Some people, however, have gone so far beyond what they needed to do to help me that it would be churlish not to thank them. Margaret Alexander, our librarian, cheerfully got me copies of every document that looked even remotely relevant, often from fragmentary citations, with astonishing speed and accuracy. Lolly Brown and Tim Carlson kept everything with a chip in it, no matter how weird, running so smoothly it was possible to forget what computers are normally like. Suzanne Dulle, Ginny Greinger, Ellen Goldberg, Tom Kepler and Ginger Richardson all provided much-needed moral and institutional support at crucial times. Ann Clough, Sally Nelson and Harry Owens dealt very professionally with the mess that passed for my finances.

It is a pleasure to acknowledge the value of conversation and correspondence about various aspects of this work with Derek Abbott, Phil Agre, Jeff Brantingham, Jim Brink, Geoff Canright, Kim-Sau Chung, Carole Crumley, Chris Douglas, Adam Elga, Kellie Evans, Doyne Farmer, Gary William Flake, Walter Fontana, Murray Gell-Mann, Janko Gravner, Jaehum Han, Han Jing, John Holland, Erica Jen, Tom Kepler, Nelly Kovalevskaya, Thorbjorn Knudsen, Kristina Lerman, Li Wei, Deborah Mayo, Melanie Mitchell, Mark Newman, Mats Nordahl, Jim Odell, Nienke Oomes, Scott Page, Ludo Paige, Jay Palmer, Alex Penn, Mitch Porter, Yuzuru Sato, Carl Simon, Spyros Skouras, Eric Smith, Ricard Solé, Peter Stadler, Dowman Varn, Ben Vigoda, Nick Watkins, Doug White, Henry Wright, Andy Wuensche, Danny Yee and Karl Young. Danny Yee, in particular, proofread chapters 1 and 2 of his own free will, and they much better for his efforts.

The work has been much improved by repeated presentation to (very patient) audiences at UW-Madison, the University of Michigan's Center for the Study of Complex Systems, and SFI. I particularly want to thank a physics undergrad at Madison, name unknown, who after hearing me give a talk on this work called it "utterly trivial": I had never hoped to be so clear.

For humoring my enthusiams, enduring my flakiness, long-suffering forebearance, arguments incisive and interminable, and general niceness, I am indebted to many friends: Dave Albers, Phil Apps, Gunther Eble, Sean Elicker, Dave Feldman, Jon "nd." Fetter, Cate Griebel, Rob Haslinger, Ishbel Hely, Maia Hely, Tim Hely, Wim Hordijk, Peter Hraber, Amihan Huesmann, Bill Jameson, Shareen Joshi, Michael "Mooki" Lachmann, Alex Lancaster, Michelle Martin, Cris Moore, Hateshput Moore, Chris O'Dell, Delia Pepper, Gillian Pepper, John Pepper, Olivia Posner, Susan Ptak, Ginger Richardson, Nigel Snoad, Barbara Tozier, "Uncle" Bill Tozier, Sander van der Leuw and Erik "I'll solve your model" van Nimwegen.

A special debt is owed to those who endured collaborations with me while I was working on this dissertation: Dave Albers, Jim Crutchfield, Dave Feldman, Rob Haslinger, Wim Hordijk, Kris Klinkner and Cris Moore.

To my family, I owe more than I can say.

Last and largest: this is for Kris.