Measures and Metrics of Information Processing in Complex Systems: A Rope of Sand

By

RYAN GREGORY JAMES
B.S. (California State University, Long Beach) 2007
M.S. (University of California, Davis) 2009

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Physics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

James P. Crutchfield (Chair)

John B. Rundle

Raissa M. D'Souza

Committee in Charge

2013

To Amanda: for all her love, patience, and understanding.

Abstract

How much information do natural systems store and process? In this work we attempt to answer this question in multiple ways. We first establish a mathematical framework where natural systems are represented by a canonical form of edge-labeled hidden Markov models called ϵ-machines. Then, utilizing this framework, a variety of measures are defined and algorithms for computing them from an ϵ-machine are described.

The first two measures defined are related to the length of time a system remembers. The first, the Markov order, is a well-known measure of the time one must observe a system in order to make accurate predictions. Despite its statistical nature, it is shown to be a topological property of the process's ϵ-machine. The second, the recently defined cryptic order, quantifies the ability to retrodict a system's internal dynamics. It is also shown to be a topological property of the ϵ-machine, and efficient algorithms for computing both quantities are given.

The second batch of metrics quantify information generation and storage in a system by partitioning the observations. By considering the role of both the past and the future behavior of a system, a semantic understanding of information generation emerges, labeling some information generation as *ephemeral*, having no lasting effects on the system, and the rest as *bound*, playing a role temporal structure. Following through with this decomposition, other quantities of less straight-forward interpretation are also defined. This is followed by a thorough discussion of these quantities and other derived quantities.

Lastly the decomposition of the entropy rate into ephemeral and bound components is applied to several standard chaotic systems through a duality between the entropy rate and the Lyapunov exponent. This exposes new structural behaviors hitherto unknown in these systems. These revolutions hint at a method for tuning natural or engineered systems so as to maximize the ability to harness their intrinsic computing abilities.

Acknowledgments

This dissertation and the work leading up to it would not have been possible without the help, both directly and indirectly, of so very many people.

I would first like to thank my advisor, Jim Crutchfield, who has provided me with so much support, allowing me to grow and explore. Without his guidance and inspiration I would not have been able to travel to so many place and meet so many people.

And my father, Kenneth James, who instilled in me a persistent curiosity and wonder about the world.

My mother, Catherine Janssen, who showed me how to enjoy the written word.

My grandfather, Larry Smith, who taught me mathematical rigor.

I would like to thank Chris Ellison for exploring this crazy world of complexity and information with me; and John Mahoney, who always questioned the meaning of what we wrote down.

To the rest of the Complexity Sciences Center: Sean, Spencer, Benny, Jörge, Dowman, Ben, Nick, Nix, Greg R., Alec, Korana, Jordan, Paul, Chris S., Greg W., Pooneh. The conversations I've had with all of you are invaluable.

To John Greek and Chuhee Kwon, who taught me how to think about physics and develop intuition.

And to my loving wife Amanda, without whom I would be lost.

Table of Contents

CHAPTER 1

Background

The premise of this work is that many natural systems can be viewed as *processes*: a mathematical abstraction of temporal behavior which makes use of, among other fields, symbolic dynamics, stochastic dynamical systems, information theory, and statistical mechanics. Given this common vantage a vast variety of systems can be directly compared despite their myriad of physical differences.

The underpinning philosophy implicit in this abstraction is that many aspects of natural systems can be understood in terms of *information*. In particular, the ways and amounts by which a system stores and processes this information. Utilizing information as the common dimension for each system enables the direct comparisons desired.

1.1 Processes

A *process* is a bi-infinite sequence of random variables:

$$\mathcal{P} = \ldots, X_{-2}, X_{-1}, X_0, X_1, X_2, \ldots \tag{1.1}$$

where each random variable X_t is drawn from an alphabet $\mathcal{A}$ which, while often finite, can be countable or uncountable. An arbitrary variable is chosen to ground the indexing: X_0.

A process is also a subshift endowed with a shift-invariant metric:

$$\begin{aligned} \mathcal{P} &= (\mathbf{X}, \mu) \\ &\mathbf{X} \subseteq \mathcal{A}^{\mathbb{Z}} \wedge \mathcal{P} = \sigma(\mathcal{P}) \end{aligned} \tag{1.2}$$

where $\mathbf{X}$ is a subshift, μ is a metric, and σ is the standard shift operator.

To ground notation, we denote a contiguous string of contiguous random variables starting at time t and extending a length ℓ by $X_{t:t+\ell}$. Note that the interval notation used is closed on the left and open on the right. This copies the interval notation of several popular programming languages including Python

and Ruby. A particular realization of that string is $x_{t:t+\ell}$. For simplicity, we denote the infinite string $X_{-\infty:t}$ as $X_{:t}$ and the infinite string $X_{t:\infty}$ as $X_{t:}$. This implies that an entire process can be denoted $X_{:}$. We call a process *stationary* if $X_{t:t+\ell} = X_{t':t'+\ell}$ for all t, t'.

There exist a variety of ways to construct a process from a natural system. Perhaps the simplest way to do so is to simply measure some property of the system at regular intervals. For example, one can measure the voltage in an axon every nanosecond. Signals can also be processed in many ways, most notably by discretizing them. Building off the prior example, rather than the process being the actual voltage reading every nanosecond, it can be a 0 if there was no spike within the past nanosecond, and a 1 if there was. Throughout this work, we will be working with discretized processes, where $\mathcal{A}$ is finite in size.

1.2 Information Theory

Continuing our trend of system agnostic tools, we find our analysis in the form of information theory. Information theory, developed by Claude Shannon in 1948, posits two properties of information which have had vast implications in areas from engineering to philosophy: first that information is quantitative and can therefore be measured, and second that it is divorced from semantics. Part of this work addresses the second property by providing meaning to some measures in the process framework. Furthermore, this second property is what can make information-theoretic analysis a difficult task. Interpreting exactly what a measure quantifies can be very nuanced and problematic.

We will now provide a terse introduction to some of the most common measures used in information theory. There are many other useful measures, some of which are will be defined and utilized in the remainder of this work. Others still are particularly niche and nuanced and require a significant amount of work to define and interpret. It is an ongoing goal to unify and address all these measures in a single document.

1.2.1 Entropy

The fundamental measure of the information contained in a random variable X is the *entropy*:

$$\mathrm{H}[X] = -\sum_{x \in X} p_x \log_2 p_x \tag{1.3}$$

where each x is an event from X and p_x is the probability of that event. It is useful to note that the entropy is maximized when each event in X is equally likely, and in such a case $\mathrm{H}[X] = \log_2 |X|$.

The entropy is considered the information content of a random variable due to Shannon's source coding theorem. This theorem states that a given random variable X — for example the words contained in this dissertation — can be losslessly compressed down to a total of $\mathrm{H}[X]$ bits. This is independent of compression scheme.

1.2.2 Conditional Entropy

It is sometimes of interest to ask how much information a particular random variable has given knowledge of a second variable. This quantity is the *conditional entropy*:

$$\begin{aligned} \mathrm{H}[X|Y] &= \sum_{y \in Y} p_y \, \mathrm{H}[X|Y=y] && (1.4) \\ &= \sum_{x \in X, y \in Y} p_{xy} \log_2 p_{x|y} && (1.5) \end{aligned}$$

where $\mathrm{H}[X|Y=y]$ is the entropy of the conditional distribution of X given that $Y=y$.

The conditional entropy is often interpreted as the amount of information in one random variable that is not in another. This leads to a method of partitioning information sources, which can be particularly useful when attempting to interpret how information is stored and processed.

1.2.3 Mutual Information

The *mutual information* is the canonical measure of shared information. It quantifies the amount of information a random variable X contains about a second random variable Y. This is also the amount of information that Y contains about X, and so this measure is symmetric. It is given by a variety of forms:

$$\begin{aligned} \mathrm{I}[X:Y] &= \mathrm{H}[X] - \mathrm{H}[X|Y] && (1.6) \\ &= \mathrm{H}[Y] - \mathrm{H}[Y|X] && (1.7) \\ &= \mathrm{H}[X,Y] - \mathrm{H}[X|Y] - \mathrm{H}[Y|X] && (1.8) \\ &= \sum_{x \in X, y \in Y} p_{xy} \log_2 \frac{p_{xy}}{p_x p_y} && (1.9) \end{aligned}$$

1.2.4 I-Diagrams

Taken together, these measures form a cohesive framework directly analogous to set theory. The entropy is similar to set cardinality. Conditional entropy is similar to set difference. Mutual information is similar

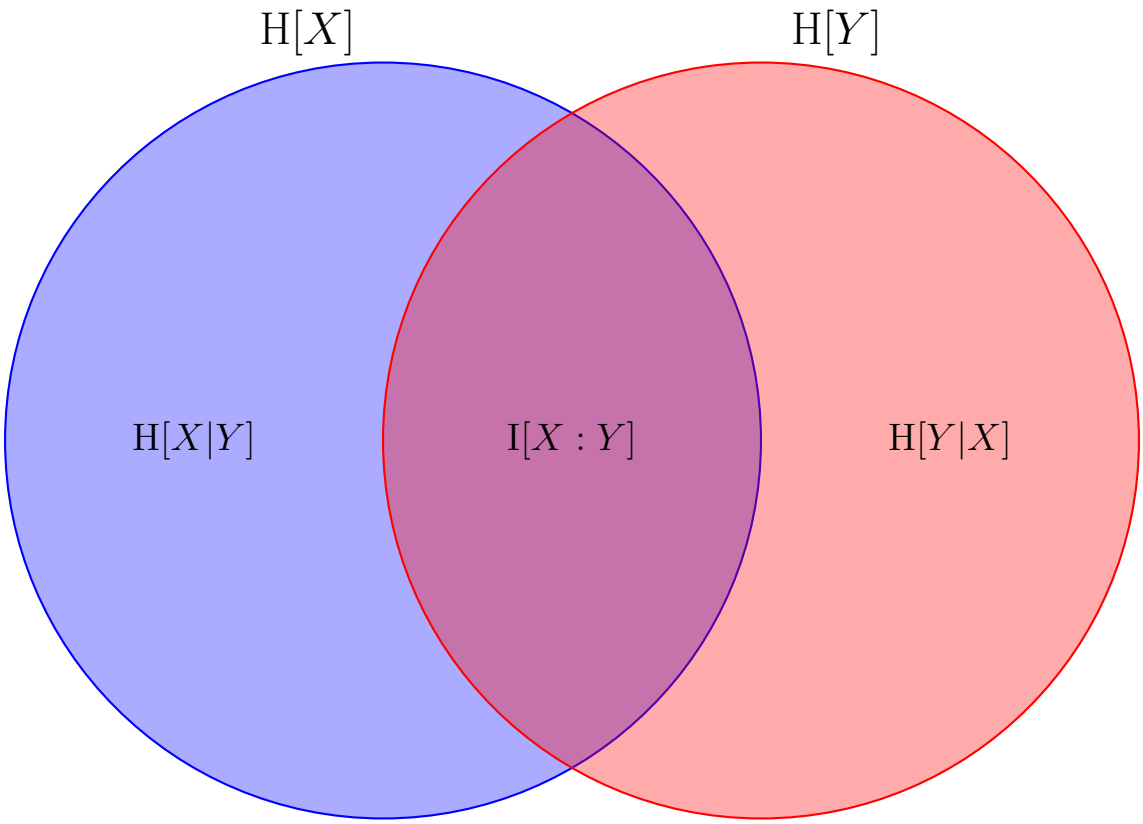


Figure 1.1: A generic I-diagram for two variables, X and Y. The entropy of each variable is represented by a circle — blue for X and red for Y. The information they share, $\mathrm{I}[X : Y]$, is the area where the two circles overlap. Information contained in the variable X but not in the variable Y, $\mathrm{H}[X|Y]$, is represented by the blue crescent on the left. Similarly, information in Y but not in X is the red crescent on the right representing $\mathrm{H}[Y|X]$. Jointly, the entire diagram represents the information $\mathrm{H}[X, Y]$, the joint entropy.

to set intersection. It is furthermore accurate to visualize these measures graphically similar to Venn diagrams. These are known as I-diagrams and an example can be seen in Fig. 1.1.

Although the diagram in Fig. 1.1 is fairly simple and straight forward, when faced with joint distributions of three, four and more variables the diagram can provided valuable insight as the the relationships and dynamics between variables. Within the context of this work, i-diagrams are most used in chapter 3.

1.3 ϵ-Machines

Perhaps the most outstanding issue with applying information theory to processes is that processes are infinite. The standard method for overcoming this is to encapsulate the process's behavior into an *automata*. These automata are a special type of probabilistic automata with a recurrent component that consists entirely of accepting states. This class of model is also known as a *hidden Markov model*.

There are an infinite number of hidden Markov models which accurately represent any given process. It is therefore useful to have a canonical model to associate with each process. Taking a cue from formal language theory, we consider the smallest unifilar[1] model and refer to this form as the *ϵ-machine*.

The ϵ-machine has many advantageous properties beyond being the smallest of its class. Its states, $\mathcal{S}$, correspond to the minimal sufficient statistic of the past about the future:

$$\mathrm{I}[\mathcal{S} : X_{0:}] = \mathrm{I}[X_{:0} : X_{0:}] \tag{1.10}$$

$$\forall \mathcal{R} \; s.t. \; \mathrm{I}[\mathcal{R} : X_{0:}] = \mathrm{I}[X_{:0} : X_{0:}] : \mathrm{H}[\mathcal{R}] \geq \mathrm{H}[\mathcal{S}] \tag{1.11}$$

It is also possible to compute many useful quantities in closed form from the ϵ-machine; including the entropy rate, the excess entropy, the statistical complexity, the Markov and cryptic orders, the bound and ephemeral information rates, and the synchronization information.

To develop some intuition for processes and their associated ϵ-machine, let us look at a few common examples. First consider the trivial independent, identically distributed (IID) process of flipping an unbiased coin with faces 0 and 1. The ϵ-machine for this process can be seen in Fig. 1.2a. There is a single state, A, and two transitions, $A \xrightarrow{\frac{1}{2}|0} A$ and $A \xrightarrow{\frac{1}{2}|1} A$. Each edge is labeled $p|s$ where p is the probability of following that edge given being in the originating state, and s is the symbol emitted by the system upon taking that transition. Each time a coin is flipped, it is equally likely to emit a 0 or a 1, and this is reflected in the single state of the ϵ-machine.

Our second example is that of a periodic process, infinitely repeating . . . 01010101 . . ., and its ϵ-machine is seen in Fig. 1.2b. In contrast to the unbiased coin example this process has no stochasticity: from each state there is only a single allowed transition and therefore that transition occurs with probability 1. This process, however, consists of two states. One state corresponds to the next symbol being a 0 and the other to the next symbol being a 1. The states thus keep track of the *phase* of the process.

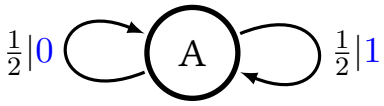
The third example is the *golden mean* process. This process has roots in one-dimensional chaotic maps and symbolic dynamics. It consists of all possible arrangements of 0s and 1s such that there are no consecutive 1s. Its ϵ-machine can be seen in Fig. 1.2c. The golden mean process is an example of a special class of process, a *Markov chain*. The easiest way to see this is that all edges leading into a state are labeled with the same unique symbol, here all edges entering state A are labeled with a 0 and all edges entering

[1] known as determinism in automata theory, we avoid that word due to the stochastic nature of our models and the confusion that may arise from the phrases such as "deterministic stochastic model".

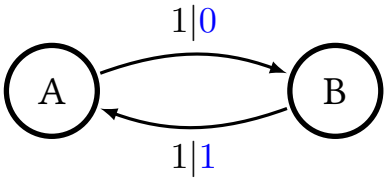
state B are labeled with a 1. Being a Markov chain means that the behavior of the golden mean process can be accurately predicted after observing just a single symbol. This process is also the first example where the stochasticity is structurally important — state A stochastically transitions to either itself or state B, and the subsequent behavior depends upon which of these two states it went to.

The next example is structurally very similar to that of the golden mean, and is known as the even process. The even process consists of even-length blocks of 1s punctuated by arbitrarily long blocks of 0s. Its ϵ-machine is represented in Fig. 1.2d. The even process is in many ways the opposite of the golden mean process. Most notably, the even process is *non-Markovian*, meaning that perfect prediction requires an arbitrarily long sequence of observations. Other differences between this and the golden mean process will be discussed in chapters 2, which discusses the Markov and cryptic orders, and 3, which discusses how they generate information.

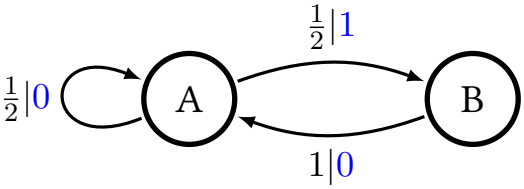
Finally we have the noisy random phase-slip (NRPS) process. Viewing its structure in Fig. 1.2e provides us with some intuition as to why it is named this. First, there is "noise" between states D and E — either a 0 or a 1 can be emitted and this does not affect future behavior. Second, the state visitation sequence is cyclic other that at state A, where the system can "slip" and state in that phase of the state visitation. As it turns out, the NRPS process is exceedingly generic and is therefore useful to illustrate a variety of information measures.



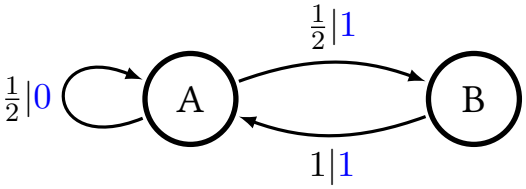
(a) Fair coin



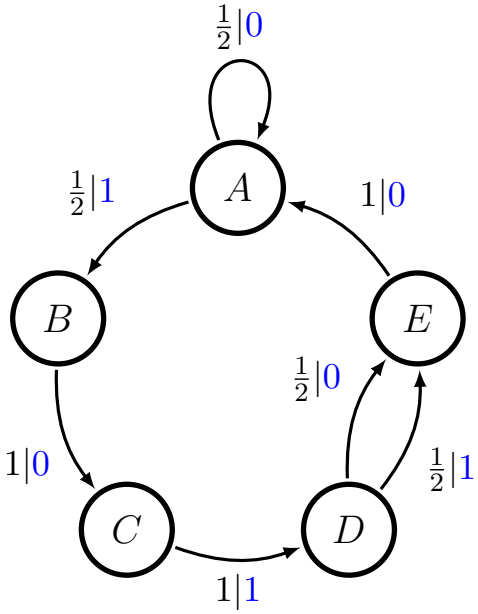
(b) Period 2



(c) Golden mean



(d) Even



(e) Noisy Random Phase-Slip

Figure 1.2: A variety of ϵ-machines. This sample displays a wide variety of behaviors and are representative of typical small ϵ-machines. The *states* of the system are represented by circles with an inscribed capital letter. The system transitions from one state to another with a particular probability p and emits a symbol s. Each edge notates this by being labeled $p|s$.

Chapter 2

Many Roads to Synchrony

2.1 Introduction

Stochastic processes are often described by the spatial and temporal length scales over which correlations exist. In physics, the range of correlations is a structural property, giving the range over which significant energetic coupling exists between the system's degrees of freedom. In time series analysis, knowing the length scale of correlations is key to successful forecasting. In biosequence analysis, the decay of correlations along DNA base pairs determines in some measure the difficulty faced by a replicating enzyme as it "decides" to begin copying a gene. The common element in these is that the correlation scale determines how quickly the observer (analyzer, forecaster, or enzyme) comes to *synchronize* to the process—that is, comes to know a relevant structure of the stochastic process.

We recently showed that there are a number of distinct, though related, length scales associated with synchronizing to presentations of stationary stochastic processes [Cru10]. Here, we show that these length scales are *topological*, depending only upon the underlying graph topology of a canonical model of the stochastic process. This reveals deep ties between the structure of the minimal sufficient statistic of a process with its observable synchronization. We will also touch upon another class of synchronization length scales introduced in Ref. [Jam11].

Specifically, we investigate measures of synchronization and their associated lengths scales for hidden Markov models (HMMs)—a particular class of processes that have an internal (hidden) Markovian dynamic that produces an observed data sequence. We focus on two such measures — the Markov order and the cryptic order — and show through a series of conceptual advances how they can be efficiently and accurately computed from the minimal sufficient statistic of the process, known as the ϵ-machine.

The development proceeds as follows. We begin by introducing the two primary measures of interest in Sec. 2.3 and demonstrate their calculation via naive methods in Sec. 2.4. Utilizing a surprising finding

in Sec. 2.5, we then show to how alleviate some of the issues with the naive approach in Sec. 2.6. We then close up the last of the issues by utilizing relevant data structures from the study of formal language theory in Sec. 2.7. These steps provide an algorithm for us. Building from this understanding of the first measure, we show how to compute the second one through similar means in Sec. 2.8. We then briefly touch upon other orders and their bounds in Sec. 2.9. Building off these computational advances, we survey the Markov and cryptic orders among ϵ-machinesin Sec. 2.10 and demonstrate that infinite correlation is a dominate property in the space of fixed-size processes and it is therefore generically hard to exactly synchronize [Tra11] to stationary processes. We next provide an example showing how these measure relate to one dimensional spin systems in Sec. 2.11. Finally, we conclude by discussing how these measure relate to other measure of interest and by suggesting applications where these algorithms will prove useful.

2.2 Background

We assume the reader has an introductory knowledge of information theory and finite-state machines, such as that found in the first few chapters of Ref. [Cov06] and Ref. [Hop01], respectively. Furthermore, we make use of ϵ-machines, a particular representation of a process that makes many properties directly and easily computable; see Ref. [Sha01]. A cursory understanding of symbolic dynamics, such as that found in the first few chapters of Ref. [Lin99] is useful for several of the results.

2.2.1 Processes

We denote subsequences in a time series as $X_{a:b}$, where $a < b$, to refer to the random variable sequence $X_a X_{a+1} X_{a+2} \cdots X_{b-1}$, which has length $b - a$. We will drop an index when it is infinite, for example the *past*, $X_{-\infty:0}$, will be denoted $X_{:0}$ and the *future*, $X_{0:\infty}$, will be denoted $X_{0:}$. We generally use w to refer to a *word*—a sequence of symbols drawn from an alphabet $\mathcal{A}$. We place two words, u and v, adjacent to each other to mean concatenation: $w = uv$. We define a *process* to be a joint probability distribution $\Pr(X_:)$ over $X_: = X_{:0} X_{0:}$.

A *presentation* of a given process is any state-based representation that generates the process. A process's ϵ*-machine* is its unique, minimal unifilar presentation [1]. The recurrent states of a process's ϵ-machine are known as the *causal states*, and are denoted $\mathcal{S}$. The causal states are the minimal sufficient statistic of $X_{:0}$ about $X_{0:}$. For a thorough treatment on presentations, please see Ref. [Cru10].

[1] Unifilar is known as "deterministic" in finite automata literature. Here, we avoid that term so that confusion does not arise due to the stochastic nature of the models being used.

2.3 Problem Statement

When confronted with a process, one of the most natural questions to ask is how much memory does it have. Is it like a coin or a die, with no memory? Does it alternate between two values, requiring that the process remember its phase? Does it express patterns that are arbitrarily long, requiring an equally arbitrarily long memory? This type of memory is quantified by the Markov order:

$$R \equiv \operatorname*{argmin}_{\ell} \{\Pr(X_0|X_{-\ell:0}) = \Pr(X_0|X_{0:})\} \tag{2.1}$$

To put it colloquially, how many prior observations must you remember to predict as well as if you remembered the infinite past? Markov chains have $R = 1$ by their very definition. Hidden Markov models, though their internal dynamics are Markovian ($R = 1$), their expressed, observed behavior can range from memoryless ($R = 0$) to arbitrary ($R = \infty$). A major topic of this paper will be to show how to compute a process's R efficiently and accurately given its ϵ-machine. In this vein it is prudent to recast Eq. 2.1 into a different form:

$$\begin{aligned}
&\Pr(X_0|X_{-R:0}) = \Pr(X_0|X_{:0}) \\
\implies & X_{:0} \sim_\epsilon X_{-R:0} \\
\implies & \mathrm{H}[\mathcal{S}_0|X_{-R:0}] = 0 \\
\implies & R = \operatorname*{argmin}_{\ell} \{\mathrm{H}[\mathcal{S}_0|X_{-\ell:0}] = 0\} \\
&\quad = \operatorname*{argmin}_{\ell} \{\mathrm{H}[\mathcal{S}_\ell|X_{0:\ell}] = 0\}
\end{aligned} \tag{2.2}$$

In essence this means that, because the past R observations predict just as well as the infinite past, the causal states are a function of length-R pasts.

The second primary length scale we will discuss is the *cryptic order*, k_χ [Mah09]. Its definition builds from Eq. 2.2:

$$k_\chi \equiv \operatorname*{argmin}_{\ell} \{\mathrm{H}[\mathcal{S}_\ell|X_{0:}] = 0\} \tag{2.3}$$

The difference between the two being that the cryptic order is conditioned on the *infinite* future, as opposed

to a finite one. This provides our interpretation of the cryptic order: it is the number of causal states which can not be *retrodicted*. That is, no matter how long we observe a process, the first k_χ internal states the process was in can not be inferred.

2.4 Naive Approach

To illustrate a direct method of observing the Markov and cryptic orders for a process, we appeal to yet another form of their definitions [CRU10]:

$$R = \underset{\ell}{\operatorname{argmin}} \left\{ \mathrm{H}[X_{0:\ell}] = \mathbf{E} + \ell h_\mu \right\} \tag{2.4}$$

$$k_\chi = \underset{\ell}{\operatorname{argmin}} \left\{ \mathrm{H}[X_{0:\ell}, \mathcal{S}_\ell] = \mathbf{E} + \ell h_\mu \right\} \tag{2.5}$$

where $\mathbf{E} = \mathrm{I}[X_{:0}, X_{0:}]$ is known as the *excess entropy* and $h_\mu = \mathrm{H}[X_0 | X_{:0}]$ is known as the *entropy rate*. The intuition for these is identical to those above: once we have reached the Markov (cryptic) order, we predict as accurately as possible. It is worth noting that these definitions only hold for *finitary* processes.

These definitions lead to a simple way of determining the Markov and cryptic orders of a system. To compute the Markov order, we calculate the entropy of longer and longer blocks of contiguous observations until the entropy grows linearly. We call this function the *block entropy curve*. The first length at which it is growth is linear is the Markov order. To compute the cryptic order, we perform a similar feat, but rather than calculating the entropy of blocks of observations alone, we calculate the entropy of those blocks as well as the causal states that could be induced by those observations. We call this function the *block-state entropy curve*. The length at which the block-state entropy curve is behaving asymptotically is the cryptic order. This approach can be seen in Fig. 2.1. The data for this plot comes from 4-state, 2-symbol machine #12810 (henceforth known simply as process #12810) shown in Fig. 2.2.

It is now important to point out the weaknesses of this approach, which are fourfold:

- Must know h_μ exactly
- Must know $\mathbf{E}$ exactly
- Must be able to differentiate *exactly on* the asymptote from *less than machine precision away from* the asymptote

- Must be able to "guess" when R or k_χ were infinite, else we'd be computing block entropies indefinitely

Two of these weaknesses are not too crippling. The entropy rate can be computed exactly from any unifilar model of the process, and so its calculation can be done fairly easily [Sha01]. Similarly, the excess entropy can be computed if the joint distribution over both a unifilar, gauge-free model of the process and a unifilar, gauge-free model of the reverse of the process is on hand [Ell09].

The later two weaknesses do not have such an easy solution. How are we to know if our calculation of the entropy at length ℓ is exactly equal to $\mathbf{E} + \ell h_\mu$, or if they are just so close that our finite precision calculations can not differentiate them? Further, if $\mathrm{H}[X_{0:\ell}]$ has not equaled $\mathbf{E}+\ell h_\mu$ after $\ell = 1,000,000$ we can not assume that it never will; perhaps the process is Markov order one billion. These are, in particular, the two weaknesses that we must overcome.

2.5 Topological

Our first step forward in solving this problem is to take a step back. Rather than considering the particular process generated by the machine in Fig. 2.2, let us consider the family of processes that exist when the probabilities are adjusted, but the structure remains the same. This family can be summarized by the machine in Fig. 2.3. If we then compute block and block-state entropy curves from a random ensemble of processes from this family, and plot the derivative of those curves and subtract out their asymptotic behavior, we arrive at Fig. 2.4.

As Figure 2.4 dramatically demonstrates, the Markov and cryptic orders are, surprisingly, independent of the transition probabilities in the structure. This means that any pattern relevant for prediction is encoded by the ϵ-machine's topology.

2.6 Synchronizing Words

On careful analysis of Eq. 2.2 it is actually not surprising that the Markov order is topological. A conditional entropy $\mathrm{H}[X|Y]$ is only zero if X is a deterministic function of Y. In this particular case, $\mathrm{H}[\mathcal{S}_R|X_{0:R}] = 0$ means that each word of length R determines a unique state of our model. That is, each word of length R is *synchronizing*. If one were to begin observing a process having no inkling as to what state it was in, after observing R symbols the exact state would be known.

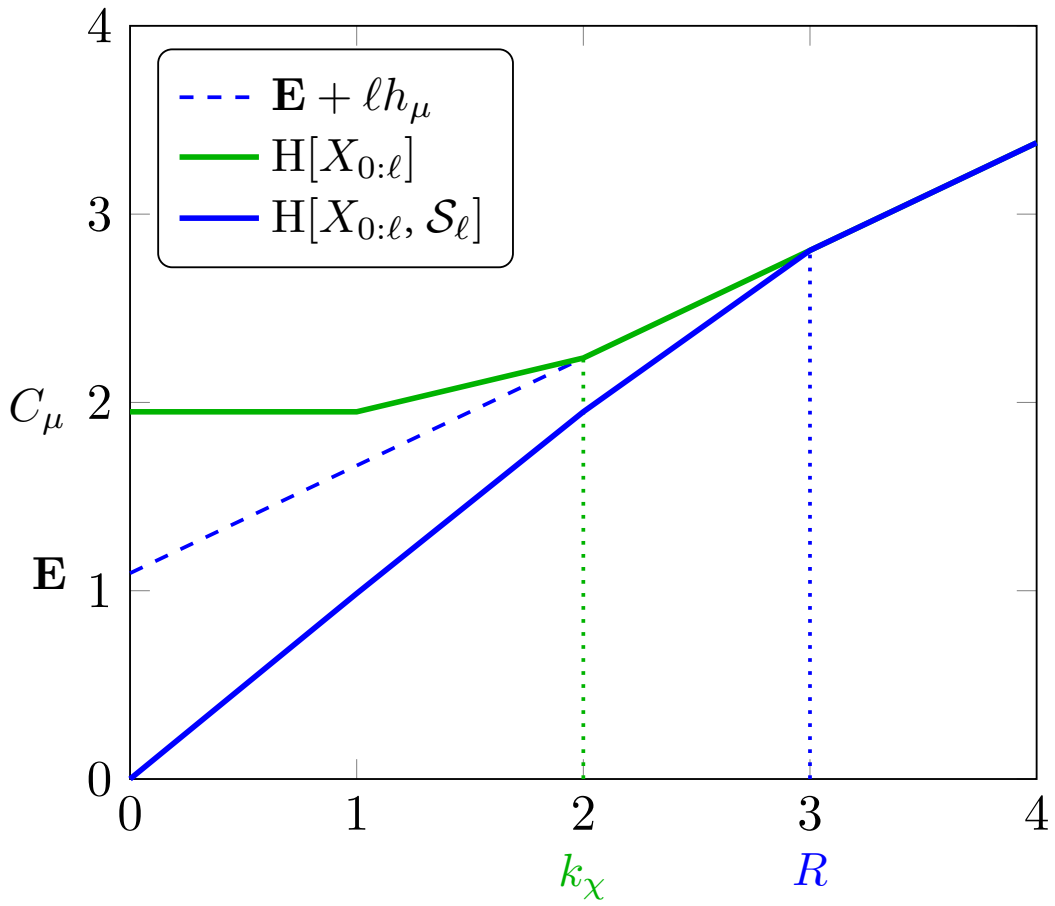


Figure 2.1: Block entropy and block state entropy. The block entropy curve has reached its asymptotic behavior at $\ell = 3$ and is therefor Markov order 3. The block-state entropy curve reaches the same asymptote at $\ell = 2$ and so the process is cryptic order 2.

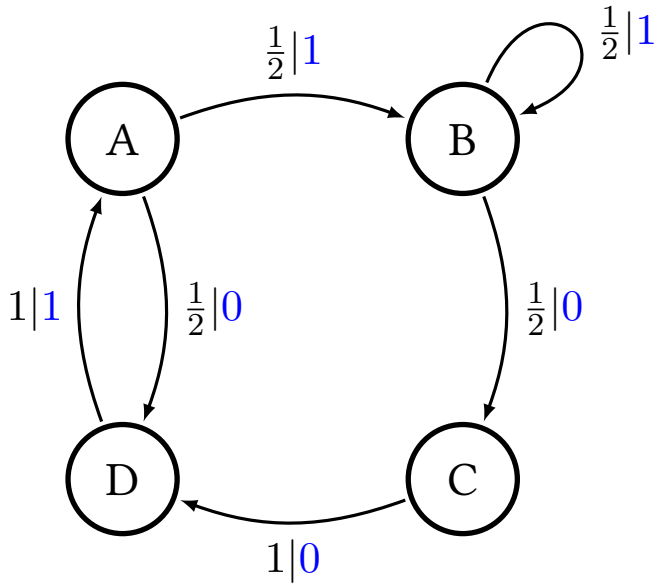


Figure 2.2: 4-state, 2-symbol process #12810. Edges are labeled $p|s$ where p is the probability of this edge being followed and s is the symbol emitted upon that edge being traversed.

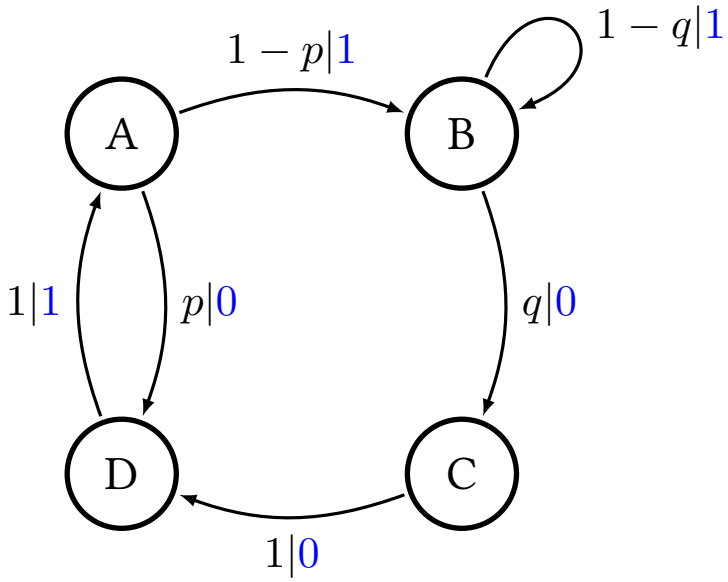


Figure 2.3: Process #12810 again, but the transition probabilities have been generalized.

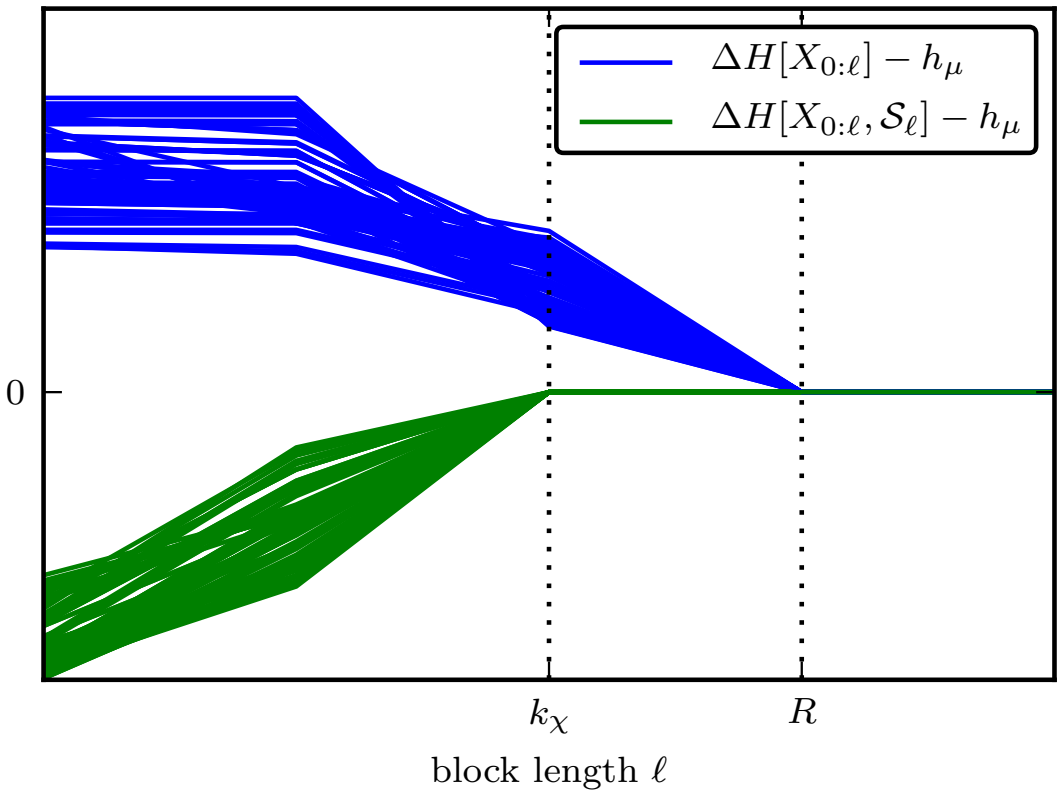


Figure 2.4: Entropy convergence curves versus block length ℓ for the family of processes seen in Fig. 2.3, with random values for p and q. The linear asymptotic behavior has been subtracted out of each curve (see inset). The Markov and cryptic orders (the lengths at which the blue (dark) and green (lighter) lines, respectively, become flat) are independent of the selected probabilities.

This provides us with an improved method of determining the Markov order. Lexicographically enumerate words until they have synchronized to a single state. The longest such word will be the Markov order, because by that point *any* word will have synchronized, and therefore the causal states will be determined uniquely by words of that length. This process has been done for process #12810 in Fig. 2.5. It can be verified that at lengths 0, 1, and 2 it is possible to still have ambiguity as to what state the system is in, for example if the two symbols observed are 10. In such a situation it is possible that the system is in either state C or state D. One more observation is required to determine which it was. Therefore, as observed previously, the Markov order for this process is 3.

This method has improved our situation somewhat. Now neither $\mathbf{E}$ nor h_μ are needed at all, nor do we need to concern ourselves with the details of comparing nearly equal floating point values. We have not, however, alleviated the issue of infinities: it is quite possible for a process to have an infinite number of these suffix-free synchronizing words, and therefore it is not feasible to enumerate them and identify the longest. To fix this problem, we will turn to a formal language theory.

2.7 Subset Construction

The remaining problem is how to find the longest suffix-free synchronizing word without having to enumerate them all. This can be accomplished with a very standard algorithm from the study of finite automata. We will construct an object known as the *power automata* (PA), named such because its states are elements of the power set of the states from another automata.

Begin constructing the power automata with a single state: the set of all states from the ϵ-machine, this will be called the *start* state. Next, recursively, for each state in the PA consider which states of the ϵ-machine are successors to those in the state of the PA on each symbol, and add a new state to the PA consisting of those successors and an edge connecting the two with that symbol. Once the successors to each state in the PA have been computed there will be a subgraph of the automata which is isomorphic to that of the ϵ-machine. This subgraph is the *recurrent* component. It is also the largest strongly connected component of the graph with no outgoing edges. The rest of the graph is *transient*.

Each path in the power automata originating from the start state and traversing only edges in the transient part of the graph, and ending in a recurrent state of the PA, are synchronizing. To find the longest synchronizing word, weight each edge in the transient part of the PA -1 and each edge of the recurrent part 0. The Bellman-Ford algorithm (or Floyd-Warshall, see Ref. [Cor09] for details regarding both) can

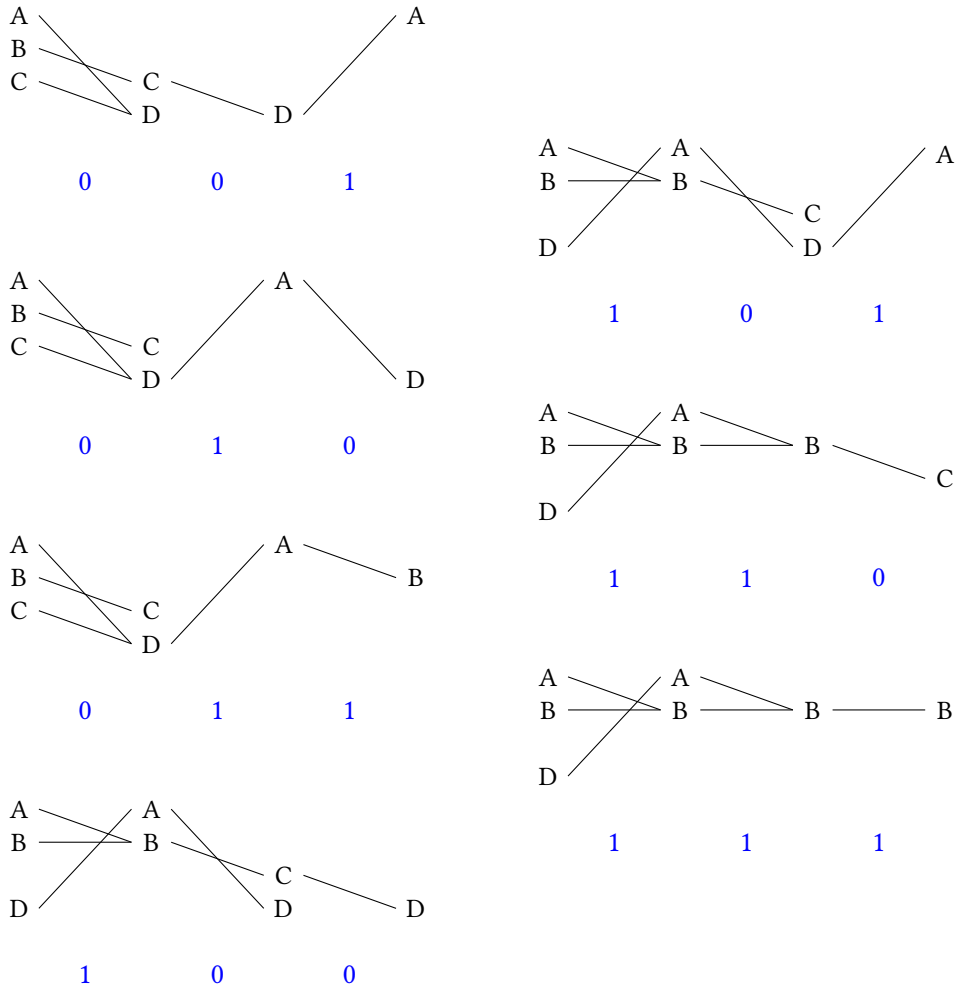


Figure 2.5: All observable words of length 3 for process #12810. Each word has been annotated with the paths through which that word invokes synchrony. It is not until the observation of three symbols that in all cases there is only a single possible state. There are, however, some words which induce synchrony more quickly.

then be employed to discover the path of least weight from the start state to any recurrent state. By the construction of the weights, the path of least weight will be the longest. The Bellman-Ford algorithm was chosen for two reason: first, it works on graphs with negative weight, and second it can detect negative weight cycles. A negative weight cycle implies that the longest path is arbitrary (infinite) in length.

We now have a complete method for computing the Markov order efficiently and accurately. First, construct the power automata. Then weight the edges according to their status as transient or recurrent. Lastly, find the path of least weight from the start to a recurrent. This algorithm runs in $O(\mathcal{A}2^{2N})$ time, which is exponential but finite, and depends only on integer calculations, and has thus alleviated all the computational difficulties of the naive approach.

2.7.1 Examples

A variety of qualitatively different behaviors can be exhibited by processes under the Markov order algorithm. Acting on process #12810, the algorithm produces a fairly simple transient structure consisting of three nodes, $ABCD$, AB, and CD, seen in Fig. 2.6. There are two longest paths starting from $ABCD$ and ending in a recurrent node: $ABCD \xrightarrow{1} AB \xrightarrow{0} CD \xrightarrow{1} A$ which is traversed with the word 101, and $ABCD \xrightarrow{1} AB \xrightarrow{0} CD \xrightarrow{0} D$ traversed with the word 100. This means that the longest synchronizing words are 101 and 100, both of length three, and therefore the Markov order is 3.

The second example we will look at is shown in Fig. 2.7. This process has a slightly more complicated transient structure than that of process #12810. Of particular note is the loop on state AB: this is because states A and B transition to each other upon production of a 0, and we can not determine which is actually the state of the system until a 1 is produced. This inability to synchronize on some words results in a non-Markovian process; that is, $R = \infty$.

The Nemo process, Fig. 2.8, is our third and final example of the Markov order algorithm. Its transient structure is particularly simple: a single state representing all the recurrent states. Since the recurrent states simply permute upon emitting a zero, the word 0000... will never allow one to determine what state the system is in. This once again means that the process is non-Markovian, $R = \infty$.

2.8 Cryptic Order

We now turn our attention to the calculation of the cryptic order. Consider Eq. 2.3, and note that it conditions on the *infinite* future. With probability 1, each infinite future synchronizes [Tra11]. We can then

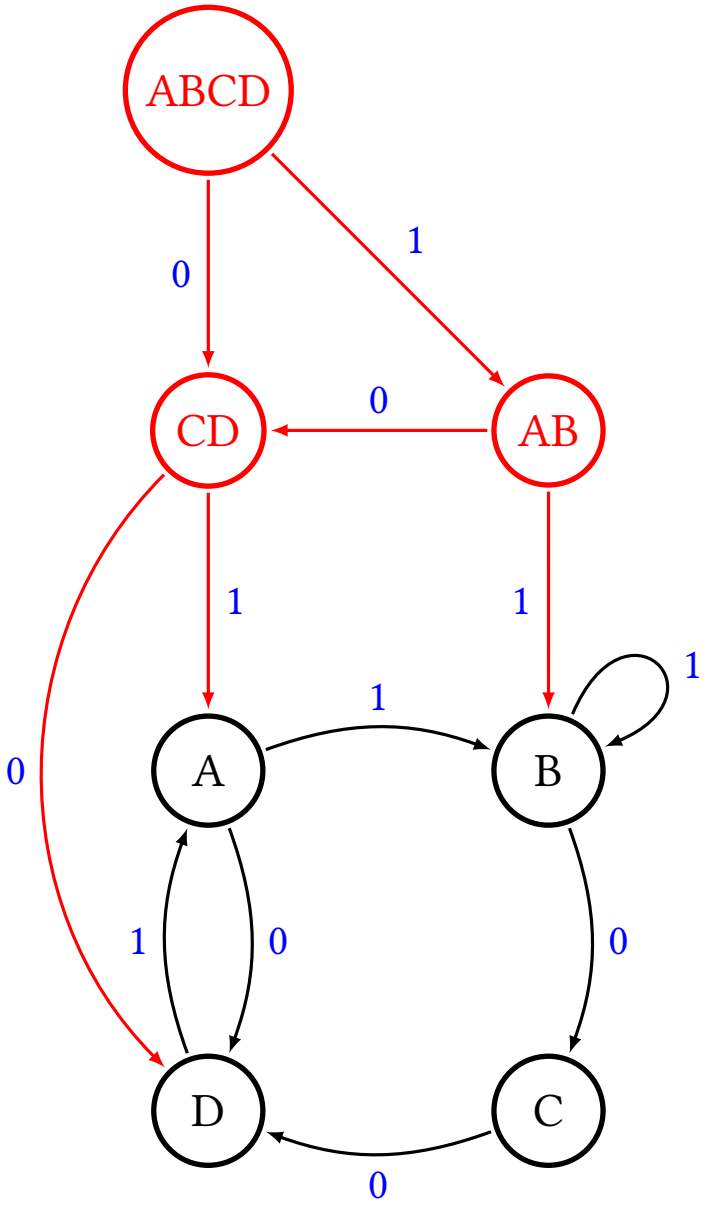


Figure 2.6: The power automata for process #12810. The longest path beginning from state $ABCD$, traversing transient (red) edges, and ending in a recurrent (black) node is of length 3: $ABCD \xrightarrow{1} AB \xrightarrow{0} CD \xrightarrow{1} A$(or $\xrightarrow{0} D$).

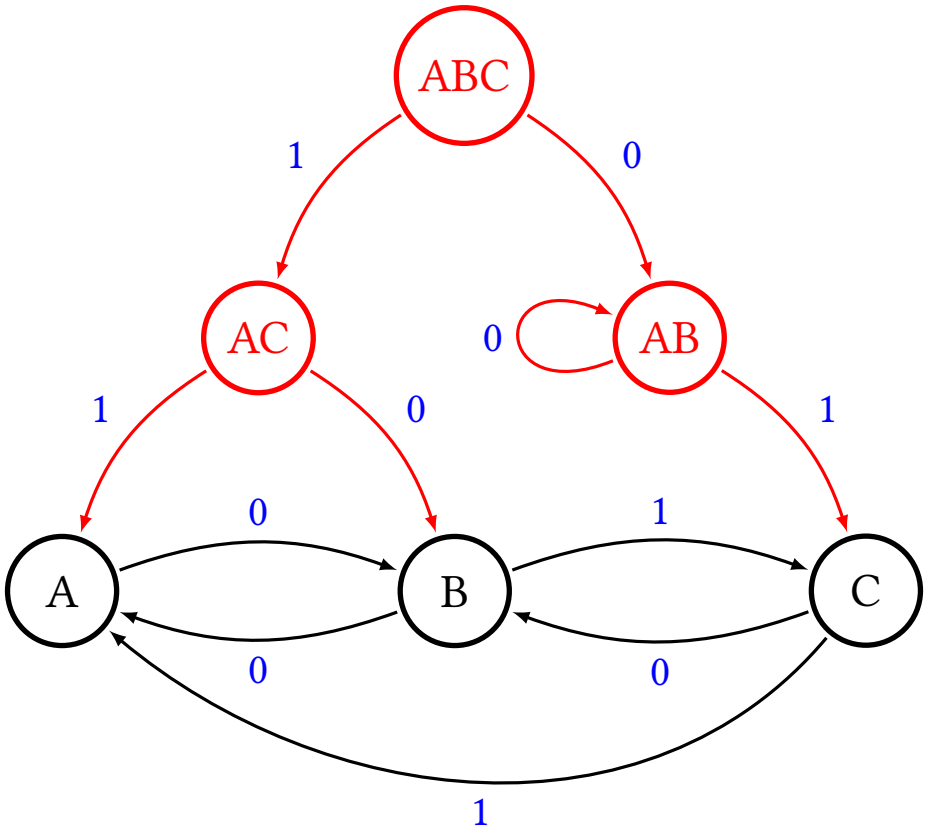


Figure 2.7: This process is a typical example of non-Markovian behavior. The signature of this is a loop in the transient structure, here the loop $AB \xrightarrow{0} AB$. This means there is the possibility of an arbitrarily long series of observations which never reveal which state the system is truly in. It is exactly this which causes the Markov order to be infinite.

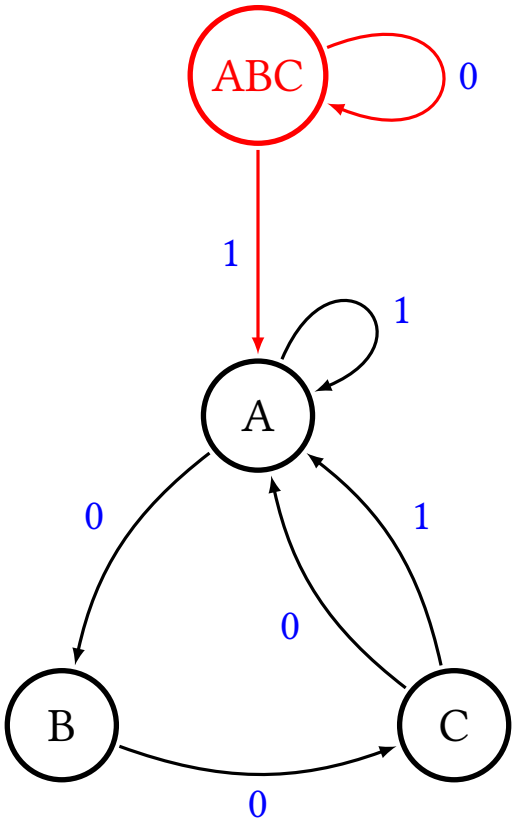


Figure 2.8: The Nemo process is, like the process in Fig. 2.7, non-Markovian. It is perhaps more clear here, however, as the recurrent states permute each other upon production of a 0. The transient structure makes that clear: ABC maps back to itself on a zero.

consider the problem of calculating k_χ to be that of determining as much of a state history as possible, given a synchronizing word and the state it synchronizes to. The maximum number of states we can not "retrodict" will be the cryptic order. Figure 2.9 depicts this process. The synchronizing paths from Fig. 2.5 are reproduced, but those paths which can not actually be traversed upon production of that word have been removed. What remains are the paths which can actually occur for the given word. From this we consider how many symbols into each word we must go before it is unambiguous as to which state the system is in. The maximum of this lengths is the cryptic order.

We are again faced with the issue that there may be an infinite number of these synchronizing words. As with the Markov order algorithm, a better method is possible. As before, we begin by constructing the power automata. We now consider the veracity of each transient edge. Take as an example the edge $ABC \xrightarrow{1} A$ in Fig. 2.8: it states that upon producing a 1 from the superposition of states A, B, and C, the system will transition to state A. Upon inspection however, the system could only have been in either state A or state C if it were to emit a 1. The core of the cryptic order algorithm is to address each transient edge in the power automata and adapt to honestly reflect the dynamics of the system. In this instance it would create a state AC which transitions to state A instead of ABC.

After the creation of a state, the automata must be made consistent. To do this the subset construction should be run on any newly added states. This will generally create new edges as well, and those too must be analyzed by the cryptic order algorithm. Once every edge has been processed by the algorithm, some transient structure will remain. Once again the longest path is the key, and the same edge weighting method with Bellman-Ford is employed.

2.8.1 Examples

The ways in which the cryptic order algorithm can modify the power automata are broad. Each example from Section 3.8 provides a different perspective into its behavior. First we consider the behavior of the algorithm on process #12810, the result of which can be seen in Fig. 2.10. The edge $CD \xrightarrow{0} D$ can be removed: to get to D on a 0, one must come from states A or C. However, due to that particular path of synchronization, we know that the system must be in either state C or D. The intersection of those two, state C, is therefore the only possible state the system could have been in, and the edge can be removed. This is not all, however – we must maintain the provenance, and so the edges $ABCD \xrightarrow{0} C$ and $AB \xrightarrow{0} C$ need to be added, since those are the edges that would have been traversed immediately prior to $CD \xrightarrow{0} D$.

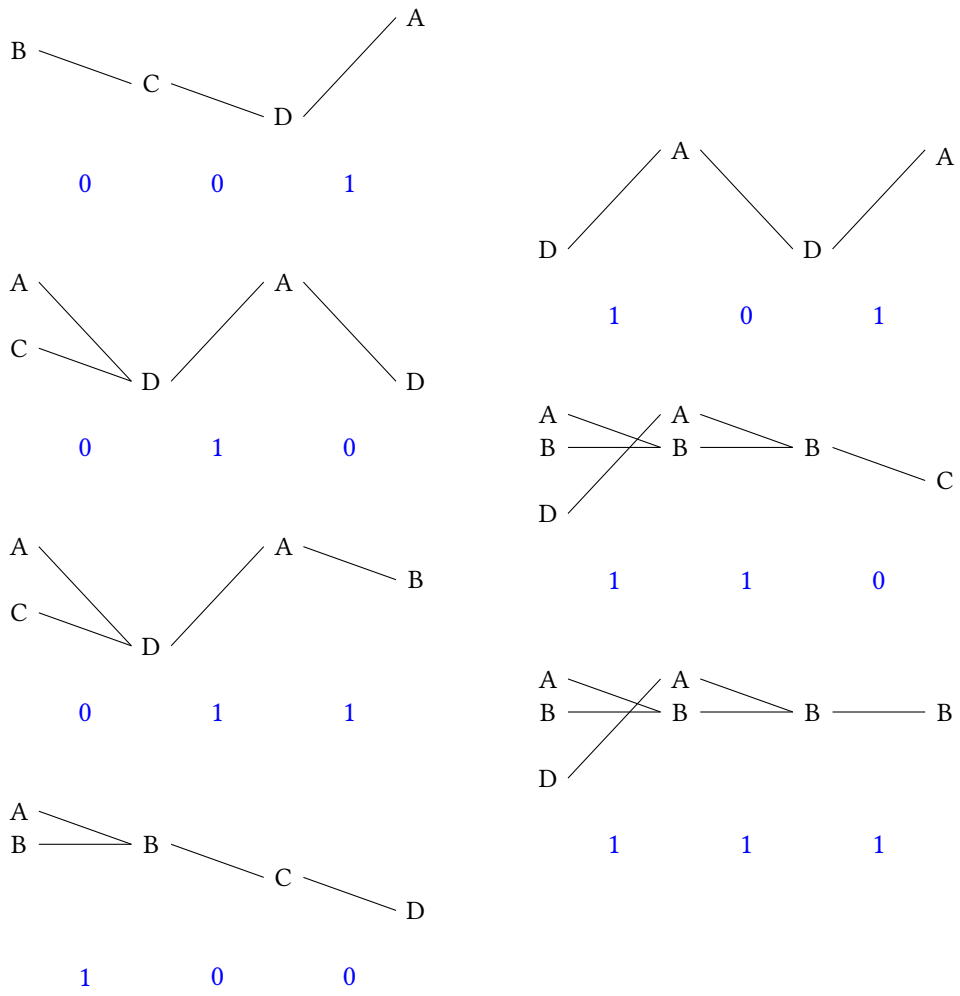


Figure 2.9: Here we see Fig. 2.5 again, except from the annotation we have removed paths which did not survive to the end of the word. It is these surviving paths which give us the cryptic order. They have each arrived at a single state by $\ell = 2$, and therefor $k_\chi = 2$.

In the end we see that the longest path from a start state to the recurrents is 2, and therefore $k_\chi = 2$, one less than the Markov order.

We next consider the example from Fig. 2.7. The output of the cryptic order algorithm on this process can be seen in Fig. 2.11. The power automata for this process consists of two major branches: one with a maximum depth of 2, and the other with containing a loop. By application of the cryptic order algorithm it is discovered that the branch with a loop is completely retrodictable. $AB \xrightarrow{1} C$ is actually $B \xrightarrow{1} C$, and this creates edges $AB \xrightarrow{0} B$ and $ABC \xrightarrow{0} B$, again to maintain provenance. The first of these newly added edges is also retrodictable: $AB \xrightarrow{0} B$ can only actually be $A \xrightarrow{0} B$. The second, $ABC \xrightarrow{0} B$, is in reality $AC \xrightarrow{0} B$. Along this branch of the transient structure, we are thus only unable to retrodict the word 01, of which the 1 can be retrodicted, leaving us with simply $AC \xrightarrow{0} B$. The other branch is more easily processed, and leaves us with $BC \xrightarrow{1} AC \xrightarrow{0} B$, the later part of which was already in our structure from the other branch. This leaves us with a longest path of length 2, making $k_\chi = 2$. This process is an example of one with an infinite Markov order, but finite cryptic order.

The last example that will be considered is the Nemo process. Recall that it is infinite Markov, as observed in Fig. 2.8. Applying the cryptic order algorithm results in the structure shown in Fig. 2.12. In this particular example, the transient structure actually grows under the algorithm. The edge connecting the transient to the recurrent structure in the power automata, $ABC \xrightarrow{1} A$, is modified in the algorithm because B can not transition to A on a 1; The state AC is created, connected to A. Completing the power automata structure from this state results in states AB and BC being added, forming the cycle $AC \xrightarrow{0} AB \xrightarrow{0} BC \xrightarrow{0} AC \ldots$. This cycle is also accurate as far as the cryptic order is concerned: each of those states can actually be transitioned to by the recurrent states in the prior state. This cycle results in an arbitrarily long path, and therefore $k_\chi = \infty$.

2.9 Other Orders

Drawing upon the interpretation of the Markov and cryptic orders as the block length at which an information measure reaches its asymptotic behavior, we introduce five new orders associated with the information measures discussed in James et al. [JAM11]. The first order is k_I, the length at which the multivariate mutual information reaches its asymptotic behavior. Unfortunately no bounds are known for this order.

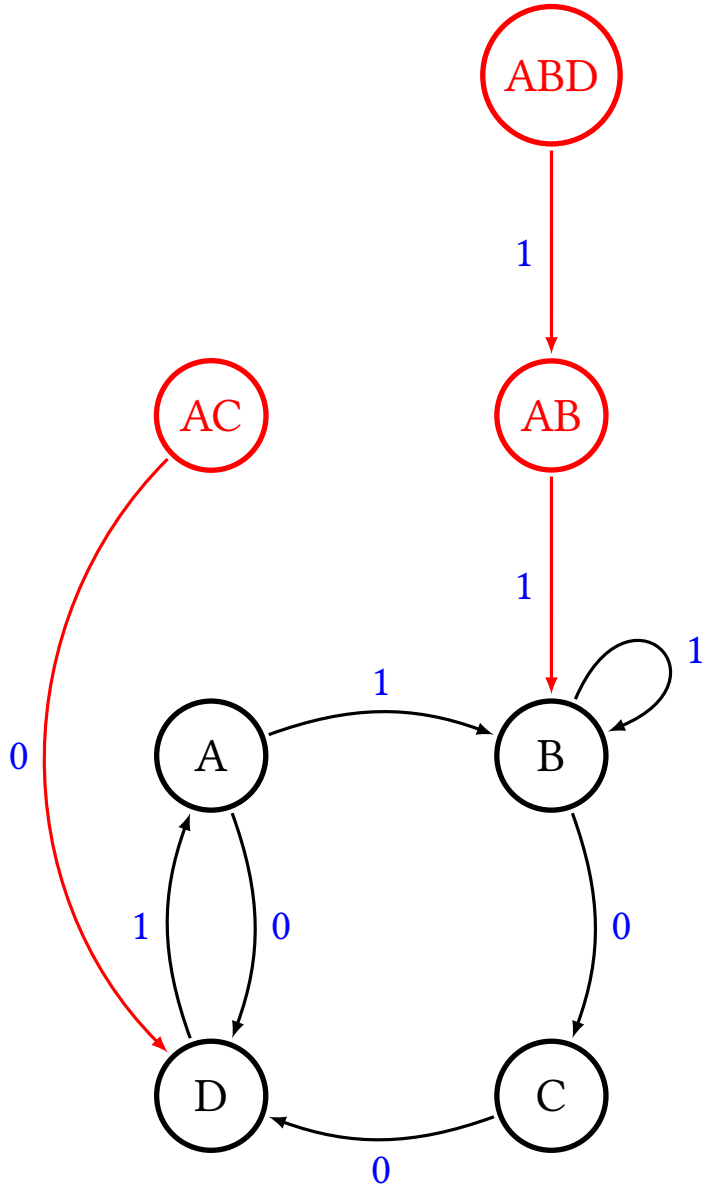


Figure 2.10: The result of the cryptic order algorithm applied to process #12810. The power automata, seen in Fig. 2.6, suggests that the word 11 could originate in any of A, B, C, or D, but a careful analysis of the recurrent structure shows that C could not be the originator of the word 11, whereas the other three states could. The cryptic order algorithm takes such constraints into account. The longest path from a transient state to a recurrent state is $ACD \xrightarrow{1} AB \xrightarrow{1} B$, and therefore $k_\chi = 2$.

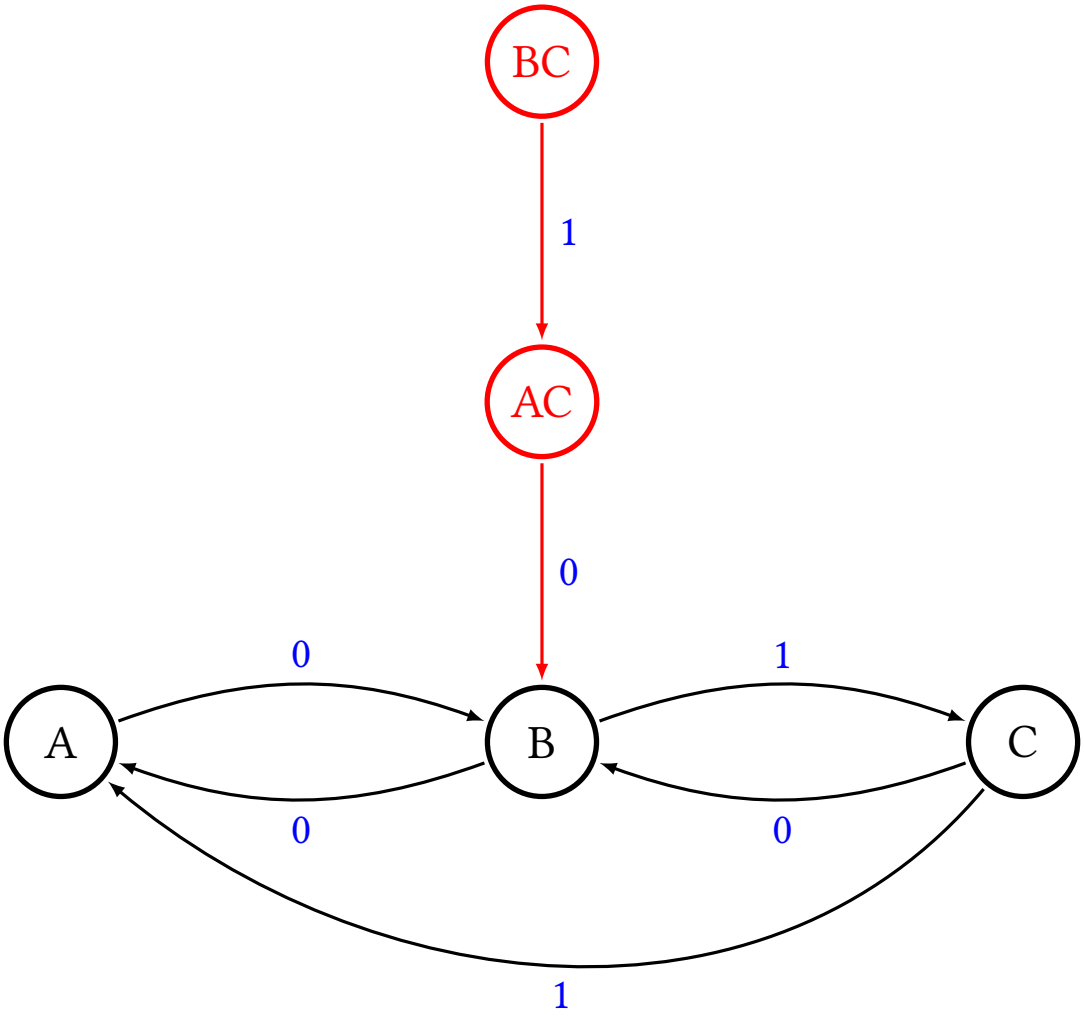


Figure 2.11: The result of the cryptic order algorithm when applied to the process seen in Fig. 2.7. The branch of the transient structure, $ABC \overset{0}{\rightarrow} (AB \overset{0}{\rightarrow} AB)^* \overset{1}{\rightarrow} C$, which witnessed the arbitrarily long synchronizing word 00^*1, can be perfectly retrodicted. Further, only a fragment of the left branch of the transient structure remains. This fragment has a length of 2, and so $k_\chi = 2$.

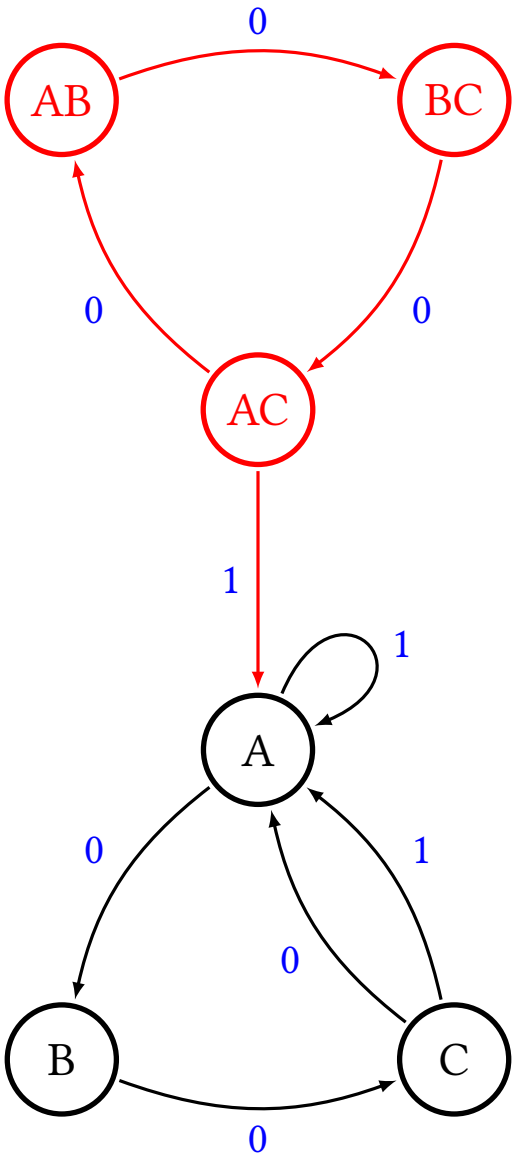


Figure 2.12: The Nemo process provides a particularly illustrative example of the cryptic order algorithm. The power automata contains the edge $ABC \xrightarrow{1} A$. However, upon inspection of the automata only states A and C can transition to A on a 1. This causes the creation of an AC state. Upon the production of a 0, AC becomes AB, and upon a second 0 that becomes BC. A third 0 completes the cycle. These edges are legitimate: the states that truly lead to AC on a 0 are BC, and those that lead to BC are AB, etc. There is a cycle in the cryptic order algorithm's transient structure, and therefore we conclude that $k_\chi = \infty$.

The others, k_R, k_B, k_Q, and k_W, are the lengths at which the residual entropy, the binding information, the enigmatic information, and the local exogenous information all reach their respective asymptotes [Jam11]. Further, these four orders are equal due to the linear interdependence of their respective measures. Here we provide lower and upper bounds for these with respect to the Markov order. Consider Figure 8 from Ref. [Jam11]: by definition $\mathrm{H}[X_{:0}]$ can be replaced with $\mathrm{H}[X_{-R:0}]$ and, if the process is stationary, $\mathrm{H}[X_{1:}]$ with $\mathrm{H}[X_{1:R+1}]$. It is therefore reasonable that it requires at least R symbols and most $2R$ symbols to accurate dissect $\mathrm{H}[X_0]$. Indeed, numerical surveys agree with these limits.

While we have defined these orders, and provided bounds for some of them, it remains to be seen if there exists an efficient method to compute them, let alone a topological interpretation.

2.10 Survey

We illustrate applying the above results and algorithms by empirically answering several simple questions. How typical are infinite Markov order and infinite cryptic order presentations in the space of finite-state processes?

Restricting ourselves to ϵ-machines, we enumerate all binary processes with a given number of states to which one can exactly synchronize [Joh10]. Using these ϵ-machines one can compute, as we just showed, their Markov and cryptic orders. The result for all of the $1,132,613$ six-state ϵ-machines is shown in Fig. 2.13.

The number of ϵ-machines that share a (R, k_χ) pair is illustrated by the size of the circle at that (R, k_χ). It is easily seen that the vast majority of processes—in fact, 98%—are non-Markovian at this state-size (6). Furthermore, most (85% to be exact) of those non-Markovian processes are also ∞-cryptic. One concludes that, in the space of finite-state processes, infinite-range correlation and infinitely diffuse internal state information are the overwhelming rule. As a consequence, it is generically difficult to synchronize.

Also of interest are the "forbidden" (R, k_χ) pairs within the space of 6-state ϵ-machines. For example, ϵ-machines with $k_\chi = 4, 5, 8, 10, 11$ do not occur with $R = 13$. In addition, in the case of infinite Markov order, but finite cryptic order, the latter appears to be bounded above by $k_\chi = 11$ despite the fact that larger finite cryptic orders exist for finite Markov order processes.

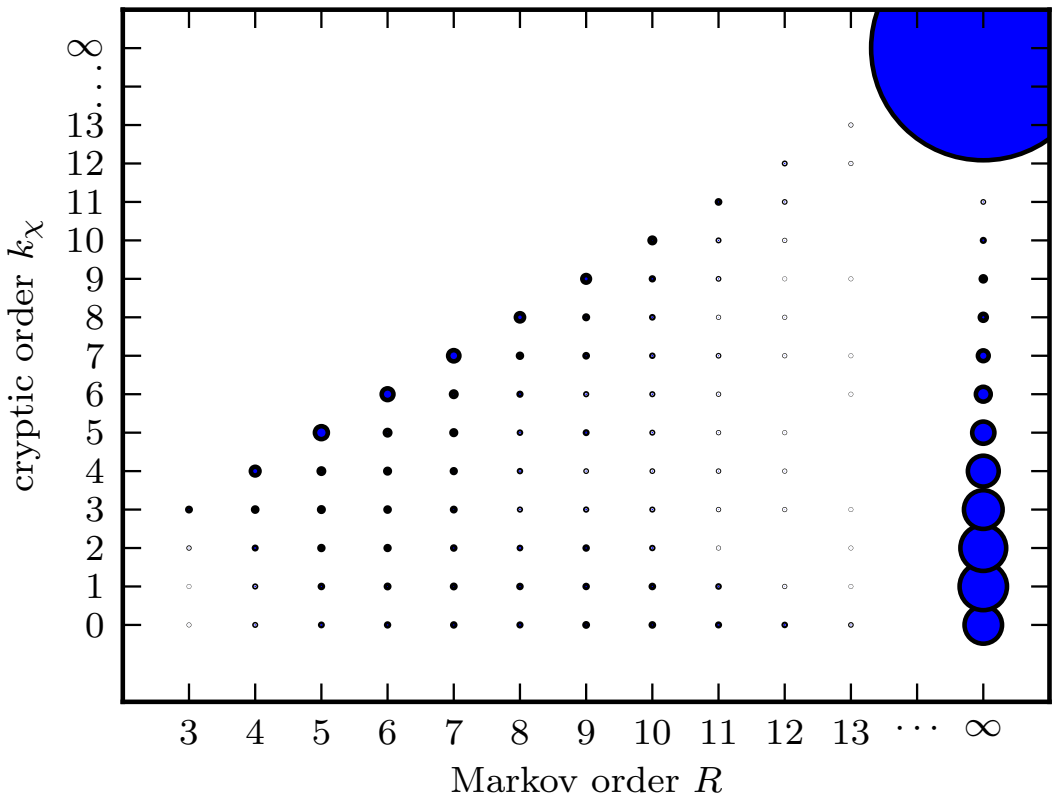


Figure 2.13: The distribution of Markov order R and cryptic order k_χ for all $1,132,613$ six-state, binary-alphabet, exactly-synchronizing ϵ-machines. The marker size is proportional to the number of ϵ-machines within this class at the same (R, k_χ) value.

2.11 Spin Systems

To draw a direct connection to physics, we now discuss how the Markov order is related to one dimensional spin chains. It has been shown by Feldman [Fel98] that the Markov order, R, of an ϵ-machine representing a spin system is equal to the range of interaction in the system's Hamiltonian.

To illustrate, consider first the ferromagnetic, one-dimensional, nearest-neighbor Ising model. The ϵ-machines for this system can be seen in Figure 2.14. As put forth above, since the system is nearest-neighbor the Markov order should therefore be 1. This is straightforward to see from the first ϵ-machine, that for a generic T. Without an observation there are two possible states the system could be in: $\uparrow$ or $\downarrow$. Once a single spin has been observed, however, the state is known exactly. The story changes somewhat at the temperature limits. At $T = 0$, the system will be in a ground state of either all up spins or all down spins, but without an external field to break this symmetry an observation must be made to determine which ground state the system is in and so the Markov order is still 1. If, however, there is an external field there will only be a single ground state — that which is aligned with the field — and no observation would be required to know which state the system is in ($R = 0$). At $T = \infty$ the system collapses to a single state where the next spin is entirely determined by thermal fluctuation, and so the Markov order is 0.

The anti-ferromagnetic, one-dimensional, nearest-neighbor Ising model is similar, and seen in Fig. 2.15. The generic temperature and high temperature limit are identical to that of the ferromagnetic case, but the low temperature is different. At $T = 0$ the spin system forms a perfect crystal of alternating spins. One must take a single observation to know what phase the crystal is in, and then the entire structure is known and so the Markov order is one. The is not a broken symmetry like ferromagnetic case, though: even with a non-zero external field an observation is still required to know which causal state the system is in.

Given the configuration of an unknown spin system, no dynamic, no ensemble, just a single typical instance from the ensemble of possible configurations, how much can be inferred about the Hamiltonian? While the technique described here does not provide coupling strengths or the like, it will give the maximum range of interactions. A variety of methods exist for inferring hidden Markov models from a sample, and any HMM can be converted to an ϵ-machine, and from there the Markov order can be directly computed.

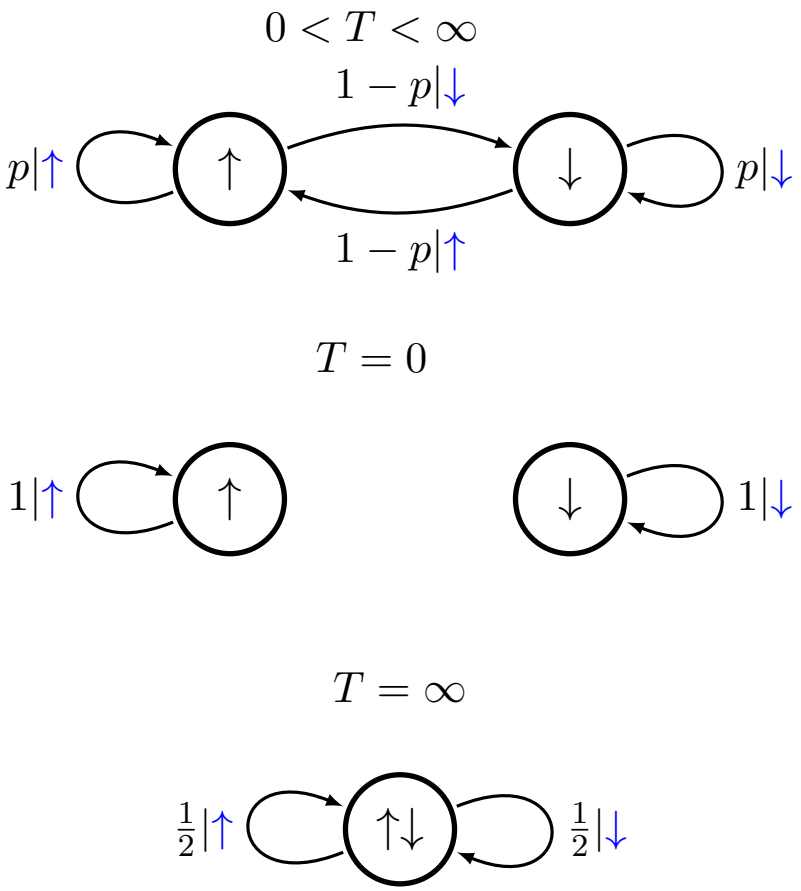


Figure 2.14: The ϵ-machines for a one dimensional ferromagnetic Ising model, where $p = \frac{1}{2}\left(1 + \tanh\beta\right)$, the external field $B = 0$, and $J = k_B = 1$.

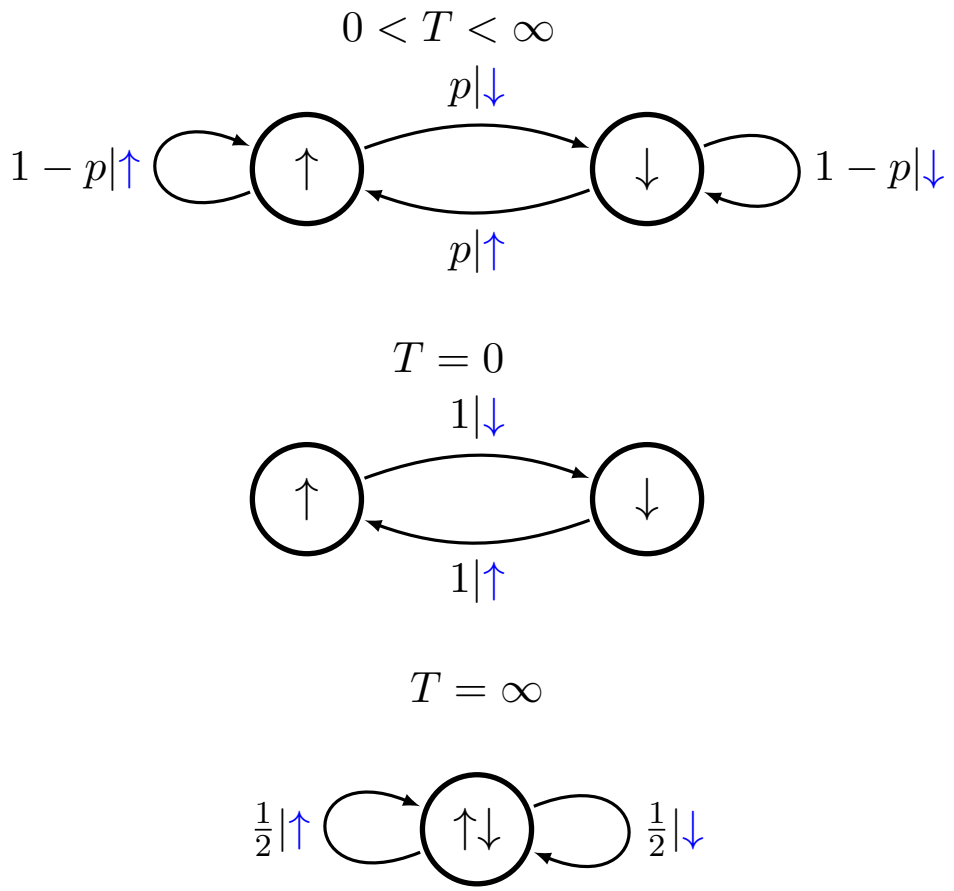


Figure 2.15: The ϵ-machines for a one dimensional anti-ferromagnetic Ising model, where $p = \frac{1}{2}\left(1 + \tanh\beta\right)$, the external field $B = 0$, and $J = k_B = 1$.

2.12 Conclusion

We began by defining two different measures of memory in complex systems. The first, the Markov order, is the length of time a system need to be observed in order for accurate predictions of its behavior to be made. The second, the cryptic order, quantifies the ability to retrodict a systems internal dynamics. Having defined these quantities, we showed that despite their statistical nature they are properties of the synchronizing words of the process's ϵ-machine. From this relationship we are able to construct efficient algorithms for their calculation.

Armed with these algorithms we survey their occurrence among all the ϵ-machineswith six states over a binary alphabet. This survey revealed a variety of interesting properties and has left a number of open questions. We have also demonstrated the behavior of the Markov order in a class of simple spin systems. From this we proposed its usage when confronted with a spin system of an unknown Hamiltonian. Although these measures are not currently widely used, we hope the advent of these algorithms encourages researchers to discover their usefulness in scientific research.

CHAPTER 3

Anatomy of a Bit: Information in a Time Series Observation

3.1 Summary

Appealing to several multivariate information measures—some familiar, some new here—we analyze the information embedded in discrete-valued stochastic time series. We dissect the uncertainty of a single observation to demonstrate how the measures' asymptotic behavior sheds structural and semantic light on the generating process's internal information dynamics. The measures scale with the length of time window, which captures both intensive (rates of growth) and subextensive components. We provide interpretations for the components, developing explicit relationships between them. We also identify the informational component shared between the past and the future that is not contained in a single observation. The existence of this component directly motivates the notion of a process's effective (internal) states and indicates why one must build

A single measurement, when considered in the context of the past and the future, contains a wealth of information, including distinct kinds of information. Can the present measurement be predicted from the past? From the future? Or, only from them together? Or not at all? How much of the measurement value is due to randomness? Does that randomness have consequences for the future or it is simply lost? We answer all of these questions and more, giving a complete dissection of a measured bit of information.

3.2 Introduction

In a time series of observations, what can we learn from just a single observation? If the series is a sequence of coin flips, a single observation tells us nothing of the past nor of the future. It gives a single bit of information about the present—one bit out of the infinite amount the time series contains. However, if the time series is periodic—say, alternating 0s and 1s—then with a single measurement in hand, the entire

observation series need not be stored; it can be substantially compressed. In fact, a single observation tells us the oscillation's phase. And, with this single bit of information, we have learned *everything*—the full bit that the time series contains. Most systems fall somewhere between these two extremes. Here, we develop an analysis of the information contained in a single measurement that applies across this spectrum.

Starting from the most basic considerations, we deconstruct what a measurement is, using this to directly step through and preview the main results. With that framing laid out, we reset, introducing and reviewing the relevant tools available from multivariate information theory including several that have been recently proposed. At that point, we give a synthesis employing information measures and the graphical equivalent of the information diagram. The result is a systematic delineation of the kinds of information that the distribution of single measurements can contain and their required contexts of interpretation. We conclude by indicating what is missing in previous answers to the measurement question above, identifying what they do and do not contribute, and why alternative state-centric analyses are ultimately more comprehensive.

3.3 A Measurement: A Synopsis

For our purposes an instrument is simply an interface between an observer and the system to which it attends. All the observer sees is the instrument's output—here, we take this to be one of k discrete values. And, from a series of these outputs, the observer's goal is to infer and to understand as much about the system as possible—how predictable it is, what are the active degrees of freedom, what resources are implicated in generating its behavior, and the like.

The first step in reaching the goal is that the observer must store at least one measurement. How many decimal digits must its storage device have? To specify which one of k instrument outputs occurred the device must use $\log_{10} k$ decimal digits. If the device stores binary values, then it must provide $\log_2 k$ bits of storage. This is the maximum for a one-time measurement. If we perform a series of n measurements, then the observer's storage device must have a capacity of $n \log_2 k$ bits.

Imagine, however, that over this series of measurements it happens that output 1 occurs n_1 times, 2 occurs n_2 times, and so on, with k occurring n_k times. It turns out that the storage device can have much less capacity; using less, sometimes substantially less, than $n \log_2 k$ bits.

To see this, recall that the number M of possible sequences of n measurements with $n_1, n_2, \ldots, n_k$

counts is given by the multinomial coefficient:

$$M = \binom{n}{n_1\ n_2\ \cdots\ n_k} = \frac{n!}{n_1! \cdots n_k!} .$$

So, to specify which sequence occurred we need no more than:

$$k \log_2 n + \log_2 M + \log_2 n + \cdots$$

The first term is the maximum number of bits to store the count n_i of each of the k output values. The second term is the number of bits needed to specify the particular observed sequence within the class of sequences that have counts $n_1, n_2, \ldots, n_k$. The third term is the number b of bits to specify the number of bits in n itself. Finally, the ellipsis indicates that we have to specify the number of bits to specify b ($\log_2 \log_2 n$) and so on, until there is less than one bit.

We can make sense of this and so develop a helpful comparison to the original storage estimate of $n \log_2 k$ bits, if we apply Stirling's approximation: $n! \approx \sqrt{2\pi n}\,(n/e)^n$. For a sufficiently long measurement series, a little algebra gives:

$$\begin{aligned} \log_2 M &\approx -n \sum_{i=1}^{k} \frac{n_i}{n} \log_2 \frac{n_i}{n} \\ &= nH[n_1/n, n_2/n, \ldots, n_k/n] . \end{aligned}$$

bits for n observations. Here, the function $H[P]$ is Shannon's *entropy* of the distribution $P = (n_1/n, n_2/n, \ldots, n_k/n)$. As a shorthand, when discussing the information in a random variable X that is distributed according to P, we also write $\mathrm{H}[X]$. Thus, to the extent that $\mathrm{H}[X] \leq \log_2 k$, as the series length n grows the observer can effectively compress the original series of observations and so use less storage than $n \log_2 k$.

The relationship between the raw measurement ($\log_2 k$) and the average-case view ($\mathrm{H}[X]$), that we just laid out explicitly, is illustrated in the contrast between Figs. 3.1(a) and 3.1(b). The difference $R_1 = \log_2 k - \mathrm{H}[X]$ is the amount of redundant information in the raw measurements. As such, the magnitude of R_1 indicates how much they can be compressed.

Information storage can be reduced further, since using $\mathrm{H}[X]$ as the amount of information in a measurement implicitly assumed the instrument's outputs were statistically independent. And this, as it turns out, leads to $\mathrm{H}[X]$ being an overestimate as to the amount of information in X. For general information

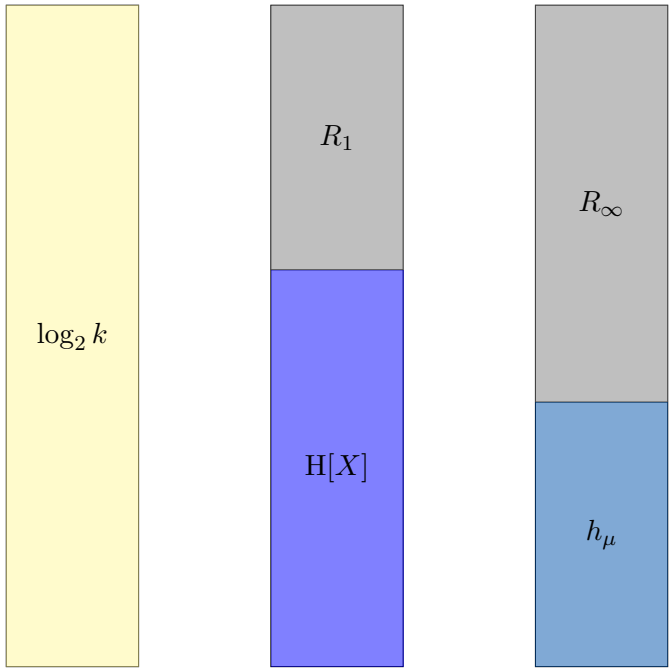


Figure 3.1: Dissecting information in a single measurement X being one of k values.

sources, there are correlations and restrictions between successive measurements that violate this independence assumption and, helpfully, we can use these to further compress *sequences* of measurements—$X_1, X_2, \ldots, X_\ell$. Concretely, information theory tells us that the irreducible information per observation is given by the Shannon *entropy rate*:

$$h_\mu = \lim_{\ell \to \infty} \frac{H(\ell)}{\ell} \,, \tag{3.1}$$

where $\mathrm{H}(\ell) = -\sum_{\{x^\ell\}} \Pr(x^\ell) \log_2 \Pr(x^\ell)$ is the *block entropy*—the Shannon entropy of the length-ℓ word distribution $\Pr(x^\ell)$.

The improved view of the information in a measurement is given in Fig. 3.1(c). Specifically, since $h_\mu \leq \mathrm{H}[X]$, we can compress even more; indeed, by an amount $R_\infty = \log_2 k - h_\mu$.

These comments are no more than a review of basic information theory [Cov06] that used a little algebra. They do, however, set the stage for a parallel, but more detailed, analysis of the information in an observation. In focusing on a single measurement, the following complements recent, more sophisticated analyses of information sources that focused on a process's hidden states [Cru10, and references therein]. In the sense that the latter is a state-centric informational analysis of a process, the following takes the complementary measurement-centric view.

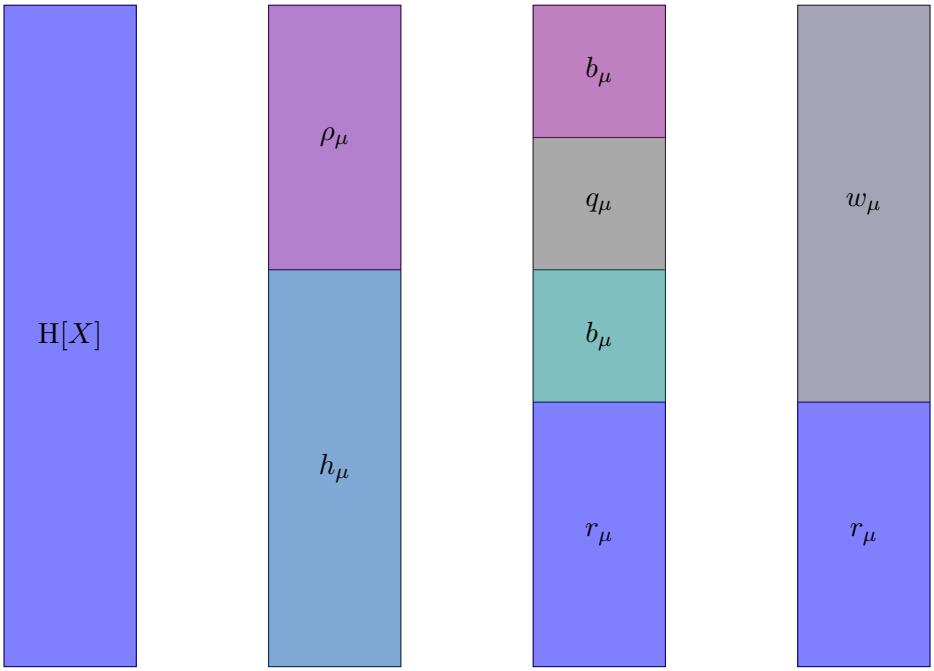


Figure 3.2: Systematic dissection of $\mathrm{H}[X]$.

Partly as preview and partly to orient ourselves on the path to be followed, we illustrate the main results in a pictorial fashion similar to that just given; see Fig. 3.2 which further dissects the information in X.

As a first cut, the information $\mathrm{H}[X]$ provided by each observation (Fig. 3.2(a)) can be broken into two pieces: one part is information ρ_μ that could be anticipated from prior observations and the other h_μ — the random component — is that which could not be anticipated. (See Fig. 3.2(b).) Each of these pieces can be further decomposed into two parts. The random component h_μ breaks into two kinds of randomness: a part b_μ relevant for predicting the future, while the remaining part r_μ is ephemeral, existing only for the moment.

The redundant portion ρ_μ of $\mathrm{H}[X]$ in turn splits into two pieces. The first part—also b_μ when the process is stationary—is shared between the past and the current observation, but its relevance stops there. The second piece q_μ is anticipated by the past, is present currently, and also plays a role in future behavior. Notably, this informational piece can be negative. (See Fig. 3.2(c).)

We can further combine all elements of $\mathrm{H}[X]$ that participate in structure — whether it be past, future, or both — into a single element w_μ. This decomposition of $\mathrm{H}[X]$ provides a very different decomposition

than h_μ and ρ_μ. It partitions $\mathrm{H}[X]$ into a piece w_μ that is structural and a piece r_μ that, as mentioned above, is ephemeral. (See Fig. 3.2(d).)

With the basic informational components contained in a single measurement laid out, we now derive them from first principles. The next step is to address information in collections of random variables, helpful in a broad array of problems. We then specialize to time series; viz., one-dimensional chains of random variables.

3.4 Information Measures

Shannon's information theory [Cov06] is a widely used mathematical framework with many advantages in the study of complex, nonlinear systems. Most importantly, it provides a unified quantitative way to analyze systems with broadly dissimilar physical substrates. It further makes no assumptions as to the types of correlation between variables, picking up multi-way nonlinear interactions just as easily as simple pairwise linear correlations.

The workhorse of information theory is the Shannon *entropy* of a random variable, just introduced. The entropy measures what would commonly be considered the amount of information learned, on average, from observing a sample from that random variable. The entropy $\mathrm{H}[X]$ of a random variable X taking on values $x \in \mathcal{A} = \{1, \ldots, k\}$ with distribution $\Pr(X = x)$ has the following functional form:

$$\mathrm{H}[X] = -\sum_{x \in \mathcal{A}} \Pr(x) \log_2 \Pr(x) \ . \tag{3.2}$$

The entropy is defined in the same manner over joint random variables—say, X and Y—where the above distribution is replaced by the joint probability $\Pr(X, Y)$.

When considering more than a single random variable, it is quite reasonable to ask how much uncertainty remains in one variable given knowledge of the other. The average entropy in one variable X given the outcome of another variable Y is the *conditional entropy*:

$$\mathrm{H}[X|Y] = \mathrm{H}[X, Y] - \mathrm{H}[Y] \ . \tag{3.3}$$

That is, it is the entropy of the joint random variable (X, Y) with the marginal entropy $\mathrm{H}[Y]$ of Y subtracted from it.

The fundamental measure of correlation between random variables is the *mutual information*. As stated before, it can be adapted to measure all kinds of interaction between two variables. It can be written

in several forms, including:

$$\mathrm{I}[X;Y] = \mathrm{H}[X] + \mathrm{H}[Y] - \mathrm{H}[X,Y] \tag{3.4}$$

$$= \mathrm{H}[X,Y] - \mathrm{H}[X|Y] - \mathrm{H}[Y|X] \,. \tag{3.5}$$

Two variables are generally considered *independent* if their mutual information is zero.

Like the entropy, the mutual information can also be conditioned on another variable, say Z, resulting in the *conditional mutual information*. Its definition is a straightforward modification of Eq. (3.4):

$$\mathrm{I}[X;Y|Z] = \mathrm{H}[X|Z] + \mathrm{H}[Y|Z] - \mathrm{H}[X,Y|Z] \,. \tag{3.6}$$

For example, consider two random variables X and Y that take the values 0 or 1 independently and uniformly, and a third $Z = X$ **XOR** Y, the exclusive-or of the two. There is a total of two bits of information among the three variables: $\mathrm{H}[X,Y,Z] = 2$ bits. Furthermore, the variables X and Y share a single bit of information with Z, their parity. Thus, $\mathrm{I}[X,Y;Z] = 1$ bit. Interestingly, although X and Y are independent, $\mathrm{I}[X;Y] = 0$, they are not conditionally independent: $\mathrm{I}[X;Y|Z] = 1$.

3.5 Multivariate Information Measures

We now turn to a difficult problem: How does one quantify interactions among an arbitrary set of variables? As just noted, the mutual information provides a very general, widely applicable method of measuring dependence between *two*, possibly composite, random variables. The challenge comes in the fact that there exist several distinct methods for measuring dependence between more than two random variables.

In the following, we will use the word *information* to name measures which involve only subsets of entropy which are shared by multiple variables, and the word *entropy* for those measures which do not. The distinction is made visually clear in Fig. 3.7. Further, although we are aware of the arbitrariness of the names of these measures and the connotations the names may or may not carry, we do not wish the reader to dwell on them. Rather, we would like to stress that it is the quantities themselves we are interested in, and that the I-diagrams of Sec. 3.7.1 provide an unbiased and unencumbered method of understanding them.

Consider a finite set $\mathcal{A}$ and random variables X_i taking on values $x_i \in \mathcal{A}$ for all $i \in \mathbb{Z}$. The vector of N random variables $X_{0:N} = \{X_0, X_1, \ldots, X_{N-1}\}$ takes on values in $\mathcal{A}^N$. A straightforward generalization

of Eq. (3.2) yields the *joint entropy*:

$$\mathrm{H}[X_{0:N}] = -\sum_{\{x_{0:N}\}} \Pr(x_{0:N}) \log_2 \Pr(x_{0:N}) \,, \tag{3.7}$$

which measures the total amount of information contained in the joint distribution. By this we mean that if we wanted to convey the particular realization this random variable took, we would need to transmit $\mathrm{H}[X_{0:N}]$ bits. From here onward, we suppress notating the set $\{x_{0:N}\}$ of realizations over which the sums are taken.

In generalizing the mutual information to arbitrary sets of variables, we make use of power sets. We let $\Omega_N = \{0, 1, \ldots, N-1\}$ denote the universal set over the variable indices and define $P(N) = \mathcal{P}\big(\Omega_N\big)$ as the power set over Ω_N. Then, for any set $A \in P(N)$, its complement is denoted $\bar{A} = \Omega_N \setminus A$ and its cardinality is denoted $|A|$. Finally, we use a shorthand to refer to the set of random variables corresponding to index set A:

$$X_A \equiv \{X_i : i \in A\} \,. \tag{3.8}$$

There are at least three extensions of the two-variable mutual information, each based on a different interpretation of what its original definition intended. The first is the *multivariate mutual information* or *co-information* [Bel03]: $\mathrm{I}[X_0; X_1; \ldots; X_{N-1}]$. Denoted $\mathrm{I}[X_{0:N}]$, it is the amount of mutual information to which *all* variables contribute:

$$\begin{aligned}
\mathrm{I}[X_{0:N}] &= -\sum \Pr(x_{0:N}) \log_2 \left(\prod_{A \in P(N)} \Pr(x_A)^{-1^{|A|}} \right) \\
&= -\sum_{A \in P(N)} (-1)^{|A|} \, \mathrm{H}[X_A] \qquad (3.9) \\
&= \mathrm{H}[X_{0:N}] - \sum_{\substack{A \in P(N) \\ 0 < |A| < N}} \mathrm{I}[X_A | X_{\bar{A}}] \,, \qquad (3.10)
\end{aligned}$$

where, e.g., $\mathrm{I}[X_{\{1,3,4\}}|X_{\{0,2\}}] = \mathrm{I}[X_1; X_3; X_4|X_0, X_2]$ and $\mathrm{I}[X_{\{1\}}|X_{\{0,2\}}] = \mathrm{H}[X_1|X_0, X_2]$. It can be verified that Eq. (3.9) is a generalization of Eq. (3.4), adding and subtracting all possible entropies according to the number of random variables they include. We will now demonstrate the usage of our notation by way of an example, first using Eq. 3.9:

$$\begin{aligned}\mathrm{I}[X_0; X_1; X_2] = {} & \mathrm{H}[X_0] + \mathrm{H}[X_1] + \mathrm{H}[X_2] \\ & - \mathrm{H}[X_0, X_1] - \mathrm{H}[X_0, X_2] - \mathrm{H}[X_1, X_2] \\ & + \mathrm{H}[X_0, X_1, X_2]\end{aligned}$$

and using Eq. 3.10:

$$\begin{aligned}\mathrm{I}[X_0; X_1; X_2] = {} & \mathrm{H}[X_0, X_1, X_2] - \mathrm{H}[X_0|X_1, X_2] \\ & - \mathrm{H}[X_1|X_0, X_2] - \mathrm{H}[X_2|X_0, X_1] \\ & - \mathrm{I}[X_0; X_1|X_2] - \mathrm{I}[X_0; X_2|X_1] \\ & - \mathrm{I}[X_1; X_2|X_0]\end{aligned}$$

The multivariate mutual information has several interesting properties. First, it can be negative, though a consistent interpretation of what this means is still lacking in the literature. Second, this measure vanishes if *any two* variables in the set are completely independent. (That is, they are independent and also conditionally independent with respect to all subsets of the other variables.) This is true regardless of interdependencies among the other variables. The multivariate mutual information has been used in the study of gene-environment interactions [Cha07].

In the second interpretation, the mutual information is seen as the relative entropy between a joint distribution and the product of its marginals. Specifically, the starting point is:

$$\mathrm{I}[X; Y] = \sum \Pr(x, y) \log_2 \frac{\Pr(x, y)}{\Pr(x) \Pr(y)} \,, \tag{3.11}$$

which is simply a rewriting of Eq. (3.4). When generalized from this form, we obtain the *total correlation* [Wat60]:

$$\begin{aligned}\mathrm{T}[X_{0:N}] &= \sum \Pr(x_{0:N}) \log_2 \left(\frac{\Pr(x_{0:N})}{\Pr(x_0) \ldots \Pr(x_N)} \right) \\ &= \sum_{\substack{A \in P(N) \\ |A|=1}} \mathrm{H}[X_A] - \mathrm{H}[X_{0:N}] \,. \end{aligned} \tag{3.12}$$

The total correlation is sometimes referred to as the "multi-information" or the "integration", though we refrain from using these ambiguous terms. It differs from the prior measure in many fundamental ways. To begin with, it is nonnegative. It also differs in that if X_0 is independent of the others, then

$T[X_{0:N}] = T[X_{1:N}]$. Finally, it compares only the individual variables to the entire set. This leads it to include redundant correlations in the set, giving a possibly misleading representation of the dependencies contained in the system. Indeed, this is a common problem. The total correlation and the net measure miss, or at best conflate unique $(n > 2)$-way interactions. The total correlation has been used to analyze frequency data [Han80].

The last extension stems from the view that mutual information is the joint entropy minus all (single-variable) unshared information—that is, we start from Eq. (3.5). When interpreted this way, the generalization is called the *binding information* [Abd12]:

$$\mathrm{B}[X_{0:N}] = \mathrm{H}[X_{0:N}] - \sum_{\substack{A \in P(N) \\ |A|=1}} \mathrm{H}[X_A | X_{\bar{A}}] \ . \tag{3.13}$$

Like the total correlation, the binding information is nonnegative and independent random variables do not change its value. Note that $\mathrm{B}[X_{0:N}]$ is a first approximation to the multivariate information of Eq. (3.9) when the sets A are restricted to singleton sets. Binding information was derived in the process of defining a possible measure of musical "interestingness" [Abd10].

We next define three additional multivariate information measures that have not been studied previously, but appear following a similar strategy. First, we have the amount of information in individual variables that is not shared in any way. This is the *residual entropy*:

$$\begin{aligned} \mathrm{R}[X_{0:N}] &= \mathrm{H}[X_{0:N}] - \mathrm{B}[X_{0:N}] \\ &= \sum_{\substack{A \in P(N) \\ |A|=1}} \mathrm{H}[X_A | X_{\bar{A}}] \ . \end{aligned} \tag{3.14}$$

In a sense, it is an anti-mutual information: It measures the total amount of randomness localized to an individual variable and so not correlated to that in its peers.

Second, we can sum the total correlation and the binding information. Then we have the *local exoge-*

nous information:

$$\mathrm{W}[X_{0:N}] = \mathrm{B}[X_{0:N}] + \mathrm{T}[X_{0:N}] \tag{3.15}$$

$$= \sum_{\substack{A \in P(N) \\ |A|=1}} \left(\mathrm{H}[X_A] - \mathrm{H}[X_A|X_{\bar{A}}]\right) \tag{3.16}$$

$$= \sum_{\substack{A \in P(N) \\ |A|=1}} \mathrm{I}[X_A; X_{\bar{A}}] \ . \tag{3.17}$$

It is the amount of information in each variable that comes from its peers. It is a "very mutual" information, one that discounts for the randomness produced locally—that randomness inherent in each variable individually.

$\mathrm{W}[X_{0:N}]$ is close to the binding information, except that it uses the sum of marginals not the joint entropy. As such, it seems to more consistently capture the role of single variables within a set than $\mathrm{B}[X_{0:N}]$, which compares the set's joint entropy to individual residual uncertainties.

Third and finally, there is a measure which, for lack of a better name, we call the *enigmatic information*:

$$\mathrm{Q}[X_{0:N}] = \mathrm{T}[X_{0:N}] - \mathrm{B}[X_{0:N}] \ . \tag{3.18}$$

Like the multivariate mutual information—which it equals when $N = 3$—it can be negative. Its operational meaning will become clear on further discussion.

3.6 Time Series

We now adapt the general multivariate measures to analyze discrete-valued, discrete-time series generated by a stationary process. That is, rather than analyzing sets of random variables, we specialize to a one-dimensional chain of them. In this setting, the measures are most appropriately applied to successively longer blocks of consecutive observations. This allows us to study the asymptotic block-length behavior of each, mimicking the approach of Ref. [CRU03, CRU10]. For the class of processes known as *finitary* (defined shortly), each of these measures tend to a linear asymptote characterized by a subextensive component and an extensive component controlled by an asymptotic growth rate.

Let's first state more precisely and introduce the notation for the class of processes that are the object of study. We consider a bi-infinite chain $\ldots X_{-1}X_0X_1 \ldots$ of random variables. Each $X_t, t \in \mathbb{Z}$, takes on a finite set of values $x_t \in \mathcal{A}$. We denote contiguous subsets of the time series with $X_{A:B}$ where the left

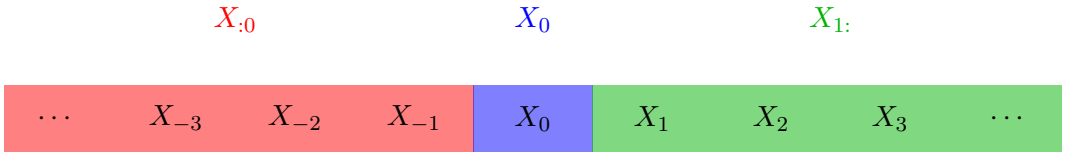


Figure 3.3: A process's time series: Time indices less than zero refer to the past $X_{:0}$; index 0 to the present X_0; and times after 0 to the future $X_{1:}$.

index is inclusive and the right is exclusive. By leaving one of the indices off the subset is partially infinite in that direction. We divide this bi-infinite chain into three segments. First we single out the *present* X_0. All the symbols prior to the present are the *past* $X_{:0}$. The symbols following the present are the *future* $X_{1:}$. Figure 3.3 illustrates the setting.

Our focus is on the ℓ-blocks $X_{t:t+\ell} = X_t X_{t+1} \cdots X_{t+\ell-1}$. The associated *process* is specified by the set of length-ℓ word distributions: $\{\Pr(X_{t:t+\ell}) : t \in \mathbb{Z}, \ell \in \mathbb{N}\}$. We consider only *stationary* processes for which $\Pr(X_{t:t+\ell}) = \Pr(X_{0:\ell})$. And so, we drop the absolute-time index t. More precisely, the word probabilities derive from an underlying time-shift invariant, ergodic measure μ on the space of bi-infinite sequences.

In the following, an information measure $\mathcal{F}$ applied to to the process's length-ℓ words is denoted $\mathcal{F}[X_{0:\ell}]$ or, as a shorthand, $\mathcal{F}(\ell)$.

3.6.1 Block Entropy versus Total Correlation

We begin with the long-studied block entropy information measure $\mathrm{H}(\ell)$ [Cru83, Eri87]. (For a review and background to the following see Ref. [Cru03].) The block entropy curve defines two primary features. First, its growth rate limits to the *entropy rate* h_μ. Second, its subextensive component is the *excess entropy* $\mathbf{E}$:

$$\mathbf{E} = \mathrm{I}[X_{:0}; X_{0:}] \ , \tag{3.19}$$

which expresses the totality of information shared between the past and future.

The entropy rate and excess entropy, and the way in which they are approached with increasing block length, are commonly used quantifiers for complexity in many fields. They are complementary in the sense that, for finitary processes, the block entropy for sufficiently long blocks takes the form:

$$\mathrm{H}(\ell) \sim \mathbf{E} + \ell h_\mu \ . \tag{3.20}$$

Recall that $\mathrm{H}(0) = 0$ and that $\mathrm{H}(\ell)$ is monotone increasing and concave down. The finitary processes, mentioned above, are those with finite $\mathbf{E}$.

Next, we turn to a less well studied measure for time series—the *block total correlation* $\mathrm{T}(\ell)$. Adapting Eq. (3.12) to a stationary process gives its definition:

$$\mathrm{T}(\ell) = \ell\,\mathrm{H}[X_0] - \mathrm{H}(\ell) \ . \tag{3.21}$$

Note that $\mathrm{T}(0) = 0$ and $\mathrm{T}(1) = 0$. Effectively, it compares a process's block entropy to the case of independent, identically distributed random variables. In many ways, the block total correlation is the reverse side of an information-theoretic coin for which the block entropy is the obverse. For finitary processes, its growth rate limits to a constant ρ_μ and its subextensive part is a constant that turns out to be $-\mathbf{E}$:

$$\mathrm{T}(\ell) \sim -\mathbf{E} + \ell\rho_\mu \ . \tag{3.22}$$

That is, $\rho_\mu = \lim_{\ell\to\infty} \mathrm{T}(\ell)/\ell$. Finally, $\mathrm{T}(\ell)$ is monotone increasing, but concave up. All of this is derived directly from Eqs. (3.20) and (3.21), by using well known properties of the block entropy.

The block entropy and block total correlation are plotted in Fig. 3.4. Both measures are 0 at $\ell = 0$ and from there approach their asymptotic behavior, denoted by the dashed lines. Though their asymptotic slopes appear to be the same, they in fact differ. Numerical data for the asymptotic values can be found in Tables 3.1 and 3.2 under the heading NRPS (defined later).

There is a persistent confusion in the neuroscience, complex systems, and information theory literatures concerning the relationship between block entropy and block total correlation [1]. This can be alleviated by explicitly demonstrating a partial symmetry between the two in the time series setting and by highlighting a weakness of the total correlation.

We begin by showing how, for stationary processes, the block entropy and the block total correlation contain much the same information. From Eqs. (3.7) and (3.12) we immediately see that:

$$\mathrm{H}(\ell) + \mathrm{T}(\ell) = \ell\,\mathrm{H}(1) \ . \tag{3.23}$$

Furthermore, by substituting Eqs. (3.20) and (3.22) in Eq. (3.23) we note that the righthand side has no subextensive component. This gives further proof that the subextensive components of Eqs. (3.20) and

[1] For example, see the conflated usage of I for both the total correlation (here, T) and its rate (here, ρ_μ) in [ERB04]

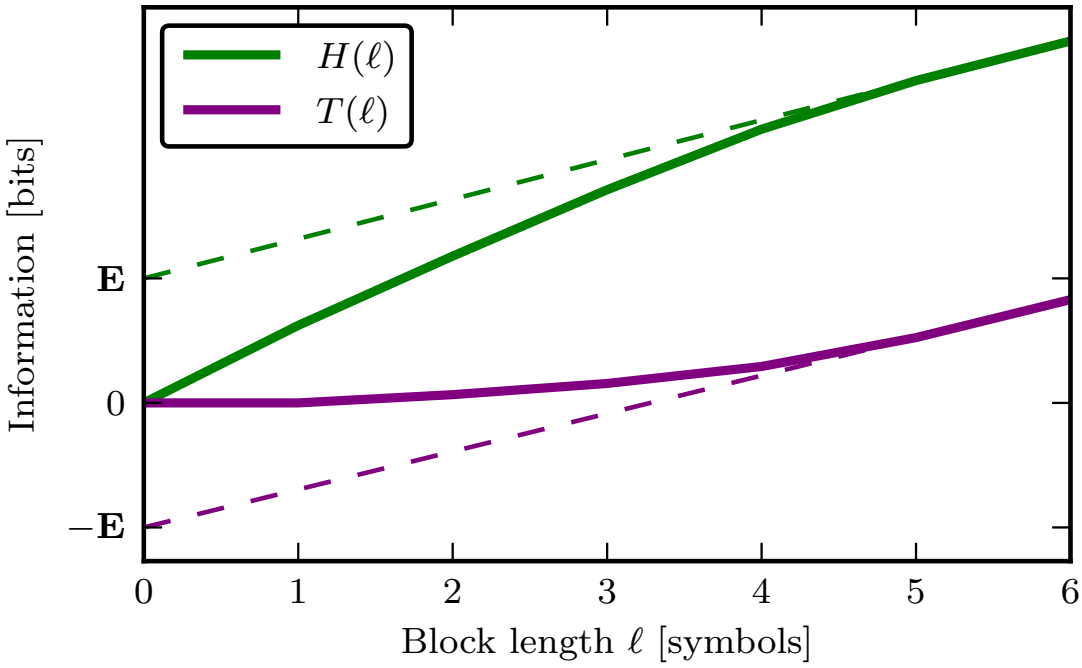


Figure 3.4: Block entropy $\mathrm{H}(\ell)$ and block total correlation $\mathrm{T}(\ell)$ illustrating their behaviors for the NRPS Process.

(3.22) must be equal and opposite, as claimed. Moreover, by equating individual ℓ-terms we find:

$$h_\mu + \rho_\mu = \mathrm{H}(1) \ . \tag{3.24}$$

And, this is the decomposition given in Fig. 3.2(b): the lefthand side provides two pieces comprising the single-observation entropy $\mathrm{H}(1)$.

Continuing, either information measure can be used to obtain the excess entropy. In addition, since the block entropy provides h_μ as well as intrinsically containing $\mathrm{H}(1)$, ρ_μ can be directly obtained from the block entropy function by taking $\mathrm{H}(1) - h_\mu$, yielding ρ_μ. The same is not true, however, for the total correlation. Though ρ_μ can be computed, one cannot obtain h_μ from $\mathrm{T}(\ell)$ alone—$\mathrm{H}(1)$ is required, but not available from $\mathrm{T}(\ell)$, since it is subtracted out.

There are further parallels between the two quantities that can be drawn. First, following Ref. [Cru03], we define discrete derivatives of the block measures at length ℓ:

$$h_\ell = \mathrm{H}(\ell) - \mathrm{H}(\ell - 1) \tag{3.25}$$

$$\rho_\ell = \mathrm{T}(\ell) - \mathrm{T}(\ell - 1) \ . \tag{3.26}$$

These approach h_μ and ρ_μ, respectively. From them we can determine the subextensive components by discrete integration, while subtracting out the asymptotic behavior. We find that:

$$\mathbf{E} = \sum_{\ell=1}^{\infty} (h_\ell - h_\mu) \tag{3.27}$$

and also that

$$\mathbf{E} = -\sum_{\ell=1}^{\infty} (\rho_\ell - \rho_\mu) \ . \tag{3.28}$$

Second, these sums are equal term by term.

The first sum, however, indirectly brings us back to Eq. (3.24). Since $h_1 = \mathrm{H}(1)$, we have:

$$\mathbf{E} = \rho_\mu + \sum_{\ell=2}^{\infty} (h_\ell - h_\mu) \ . \tag{3.29}$$

Finally, it has been said that the total correlation ("multi-information") is the first term in $\mathbf{E}$ [Erb04]. This has perhaps given the impression that the total correlation is only useful as a crude approximation. Equation (3.29) shows that it is actually the total correlation *rate* ρ_μ that is $\mathbf{E}$'s first term. As we just showed, the total correlation is more useful than being a first term in an expansion. Its utility is ultimately limited, though, since its properties are redundant with that of the block entropy which, in addition, gives the process's entropy rate h_μ.

3.6.2 A Finer Decomposition

We now show how, in the time series setting, the binding information, local exogenous information, enigmatic information, and residual entropy constitute a refinement of the single-measurement decomposition provided by the block entropy and the total correlation [Abd12, Abd10]. To begin, their block equivalents are, respectively:

$$\mathrm{B}(\ell) = \mathrm{H}(\ell) - \mathrm{R}(\ell) \tag{3.30}$$

$$\mathrm{Q}(\ell) = \mathrm{T}(\ell) - \mathrm{B}(\ell) \tag{3.31}$$

$$\mathrm{W}(\ell) = \mathrm{B}(\ell) + \mathrm{T}(\ell) \ , \tag{3.32}$$

where $\mathrm{R}(\ell)$ does not have an analogously simple form. Their asymptotic behaviors are, respectively:

$$\mathrm{R}(\ell) \sim \mathbf{E}_R + \ell r_\mu \tag{3.33}$$

$$\mathrm{B}(\ell) \sim \mathbf{E}_B + \ell b_\mu \tag{3.34}$$

$$\mathrm{Q}(\ell) \sim \mathbf{E}_Q + \ell q_\mu \tag{3.35}$$

$$\mathrm{W}(\ell) \sim \mathbf{E}_W + \ell w_\mu \ . \tag{3.36}$$

Their associated rates break the prior two components (h_μ and ρ_μ) into finer pieces. Substituting their definitions into Eqs. (3.7) and (3.21) we have:

$$\mathrm{H}(\ell) = \mathrm{B}(\ell) + \mathrm{R}(\ell) \tag{3.37}$$

$$= (\mathbf{E}_B + \mathbf{E}_R) + \ell(b_\mu + r_\mu) \tag{3.38}$$

$$\mathrm{T}(\ell) = \mathrm{B}(\ell) + \mathrm{Q}(\ell) \tag{3.39}$$

$$= (\mathbf{E}_B + \mathbf{E}_Q) + \ell(b_\mu + q_\mu) \ . \tag{3.40}$$

The rates in Eqs. (3.38) and (3.40) corresponding to h_μ and ρ_μ, respectively, give the decomposition laid out in Fig. 3.2(c) above. Two of these components (b_μ and r_μ) were defined in Ref. [Abd12] and the third (q_μ) is a direct extension. We defer interpreting them to Sec. 3.7.2 which provides greater understanding by appealing to the semantics afforded by the process information diagram developed there.

The local exogenous information, rather than refining the decomposition provided by the block entropy and the total correlation, provides a different decomposition:

$$\mathrm{W}(\ell) = \mathrm{B}(\ell) + \mathrm{T}(\ell) \tag{3.41}$$

$$= (\mathbf{E}_B - \mathbf{E}) + \ell(b_\mu + \rho_\mu) \ . \tag{3.42}$$

So, $w_\mu = h_\mu + \rho_\mu$, as mentioned in Fig. 3.2(d).

Similar to Eq. (3.23), we can take the local exogenous information together with the residual entropy and find:

$$\mathrm{R}(\ell) + \mathrm{W}(\ell) = \ell\, \mathrm{H}(1) \ . \tag{3.43}$$

This implies that $\mathbf{E}_R = -\mathbf{E}_W$ and that r_μ and w_μ are yet another partitioning of $\mathrm{H}[X]$, as shown earlier in Fig. 3.2(d).

Figure 3.5 illustrates these four block measures for a generic process. Each of the four measures

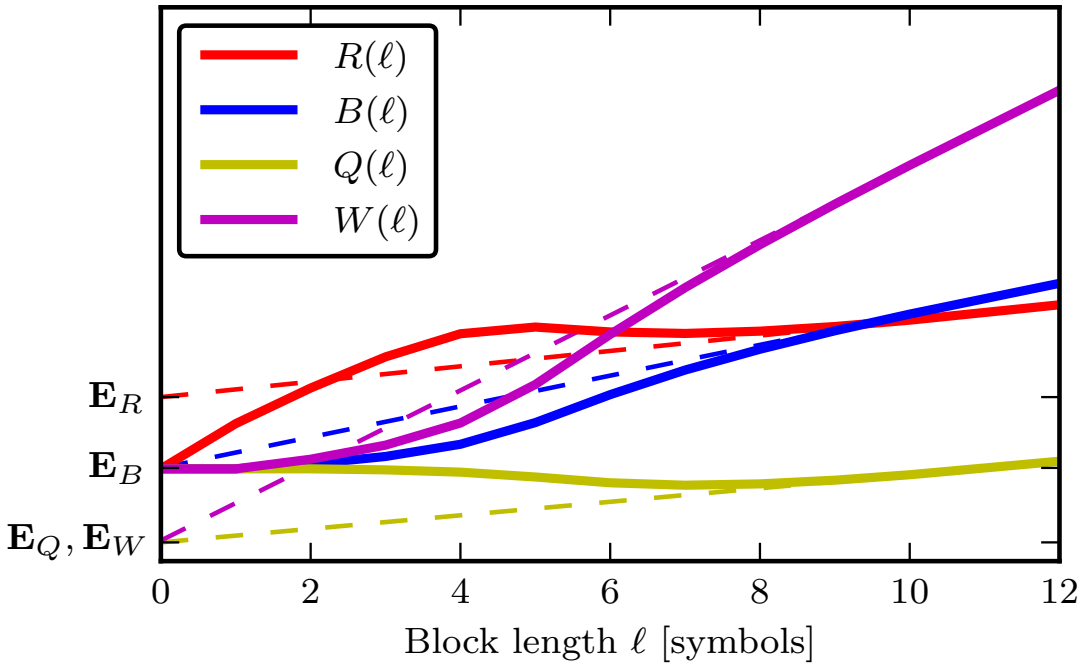


Figure 3.5: Block equivalents of the residual entropy $\mathrm{R}(\ell)$, binding information $\mathrm{B}(\ell)$, enigmatic information $\mathrm{Q}(\ell)$, and local exogenous information $\mathrm{W}(\ell)$ for a generic process (same as previous figure).

reaches asymptotic linear behavior at a length of $\ell = 9$ symbols. Once there, we see that they each possess a slope that we just showed to be a decomposition of the slopes from the measures in Fig. 3.4. Furthermore, each has a subextensive component that is found as the y-intercept of the linear asymptote. These subextensive parts provide a decomposition of the excess entropy, discussed further below in Sec. 3.7.2.

3.6.3 Multivariate Mutual Information

Lastly, we come to the block equivalent of the multivariate mutual information $\mathrm{I}[X_{0:N}]$:

$$\mathrm{I}(\ell) = \mathrm{H}(\ell) - \sum_{\substack{A \in P(\ell) \\ 0<|A|<\ell}} \mathrm{I}[X_A | X_{\bar{A}}] \ . \tag{3.44}$$

Superficially, it scales similarly to the other measures:

$$\mathrm{I}(\ell) \sim \mathbf{I} + \ell i_\mu \ , \tag{3.45}$$

with an asymptotic growth rate i_μ and a constant subextensive component $\mathbf{I}$. Yet, it has differing implications regarding what it captures in the process. This is drawn out by the following propositions, whose proofs appear elsewhere.

The first concerns the subextensive part of $\mathrm{I}(\ell)$.

Proposition 1. *For all finite-state processes:*

$$h_\mu > 0 \quad \Rightarrow \quad \lim_{\ell \to \infty} \mathrm{I}(\ell) = 0 \; . \tag{3.46}$$

The intuition behind this is fairly straightforward. For $\mathrm{I}(\ell)$ to be nonzero, no two observations can be independent. Finite-state processes with positive h_μ are stochastic, however. So, observations become (conditionally) decoupled exponentially fast. Thus, for arbitrarily long blocks, the first and the last observations tend toward independence exponentially and so $\mathrm{I}(\ell)$ limits to 0.

The second proposition regards the growth rate i_μ.

Proposition 2. *For all finite-state processes:*

$$i_\mu = 0 \; . \tag{3.47}$$

The intuition behind this follows from the first proposition. If $h_\mu > 0$, then it is clear that since $\mathrm{I}(\ell)$ tends toward 0, then the slope must also tend toward 0. What remains are those processes that are finite state but for which $h_\mu = 0$. These are the periodic processes. For them, i_μ also vanishes since, although $\mathrm{I}(\ell)$ may be nonzero, there is a finite amount of information contained in a bi-infinite periodic sequence. Once all this information has been accounted for at a particular block length, then for all blocks larger than this there is no additional information to gain. And so, i_μ decays to 0.

The final result concerns the subextensive component $\mathbf{I}$.

Proposition 3. *For all finite-state processes with $h_\mu > 0$:*

$$\mathbf{I} = 0 \; . \tag{3.48}$$

This follows directly from the previous two propositions.

Thus, the block multivariate mutual information is qualitatively different from the other block measures. It appears to be most interesting for infinitary processes with infinite excess entropy.

Figure 3.6 demonstrates the general behavior of $\mathrm{I}(\ell)$, illustrating the three propositions. The dashed line highlights the asymptotic behavior of $\mathrm{I}(\ell)$: both $\mathbf{I}$ and i_μ vanish. We further see that $\mathrm{I}(\ell)$ is not

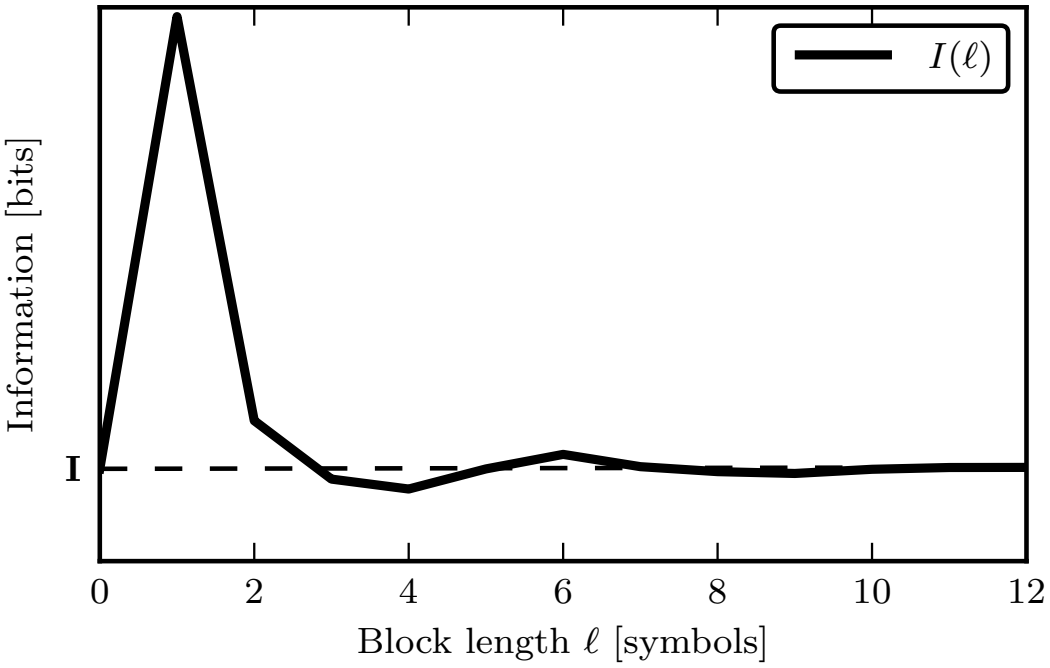


Figure 3.6: Block multivariate mutual information $\mathrm{I}(\ell)$ for the same example process as before.

restricted to positive values. It oscillates about 0 until length $\ell = 11$ where it finally vanishes.

3.7 Information Diagrams

Information diagrams [Yeu91] provide a graphical and intuitive way to interpret the information-theoretic relationships among variables. In construction and concept, they are very similar to Venn diagrams. The key difference is that the measure used is a Shannon entropy rather than a set size. Additionally, an overlap is not set intersection but rather a mutual information. The irreducible intersections are, in fact, elementary *atoms* of a sigma-algebra over the random-variable event space. An atom's size reflects the magnitude of one or another *Shannon information measure*—marginal, joint, or conditional entropy or mutual information.

3.7.1 Four-Variable Information Diagrams

Using information diagrams we can deepen our understanding of the multivariate informations defined in Sec. 3.5. Fig. 3.7 illustrates them for four random variables—X_1, X_2, X_3, X_4. There, an atom's shade of gray denotes how much weight it carries in the overall value of its measure. Consider for example the total correlation I-diagram in Fig. 3.7(c). From the definition of the total correlation, Eq. (3.12), we see that

each variable provides one count to each of its atoms and then a count is removed from each atom. Thus, the atom associated with four-way intersection $X_1 \cap X_2 \cap X_3 \cap X_4$ contained in each of the four variables carries a total weight $\mathrm{I}[X_1; X_2; X_3; X_4] = 4 - 1 = 3$. Those atoms contained in three variables carry a weight of 2, those shared among only two variables a weight of 1, and information solely contained in one variable is not counted at all.

Utilizing the I-diagrams in Fig. 3.7, we can easily visualize and intuit how these various information measures relate to each other and the distributions they represent. In Fig. 3.7(a), we find the joint entropy. Since it represents all information contained in the distribution with no bias to any sort of interaction, we see that it counts each and every atom once. The residual entropy, Fig. 3.7(e), is equally easy to interpret: it counts each atom which is not shared by two or more variables.

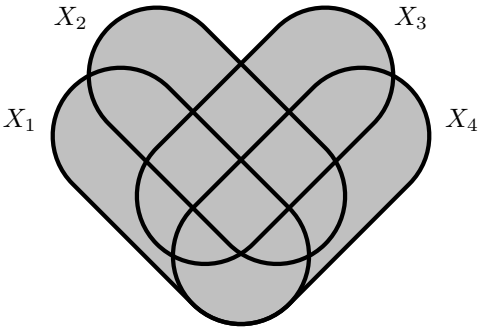
The distinctions in the menagerie of measures attempting to capture interactions among N variables can also be easily seen. The multivariate mutual information, Fig. 3.7(b), stands out in that it is isolated to a single atom, that contained in all variables. This makes it clear why the independence of any two of the variables leads to a zero value for this measure. The total correlation, Fig. 3.7(c), contains all atoms contained in at least two variables and gives higher weight to those contained in more variables. The local exogenous information, Fig. 3.7(f), is similar. It counts the same atoms as the total correlation does, but it gives them higher weight. Lastly, the binding information, Fig. 3.7(d), also counts the same atoms, but only weights each of them once regardless of how many variables they participate in.

The lone enigmatic information, Fig. 3.7(g), counts only those variables that participate in at least three variables and, similar to the total correlation, it counts those that participate in more variables more heavily.

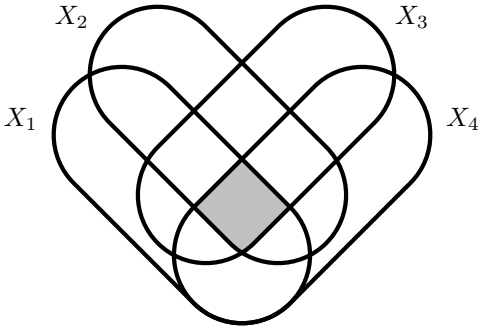
3.7.2 Process Information Diagrams

Following Ref. [Cru09] we adapt the multivariate I-diagrams just laid out to tracking information in finitary stationary processes. In particular, we develop process I-diagrams to explain the information in a single observation, as described before in Fig. 3.2. The resulting process I-diagram is displayed in Fig. 4.1. As we will see, exploring the diagram gives a greater, semantic understanding of the relationships among the process variables and, as we will emphasize, of the internal structure of the process itself.

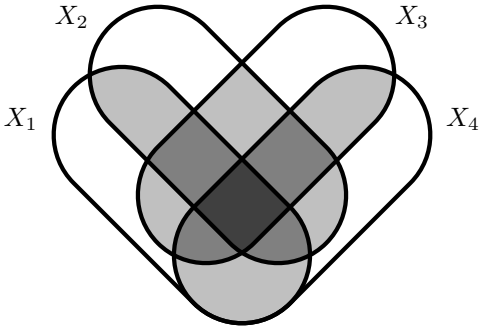
For all measures, except the multivariate mutual information, the extensive rate corresponds to one or more atoms in the decomposition of $\mathrm{H}[X_0]$. To begin, we allow $\mathrm{H}[X_0]$ to be split in two by the past.



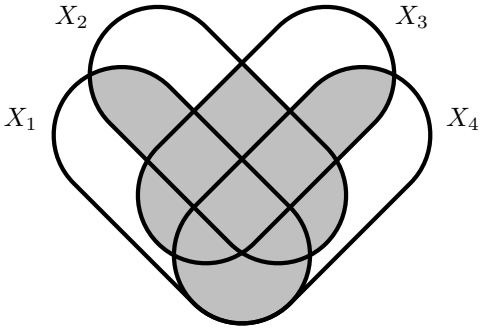
(a) Entropy



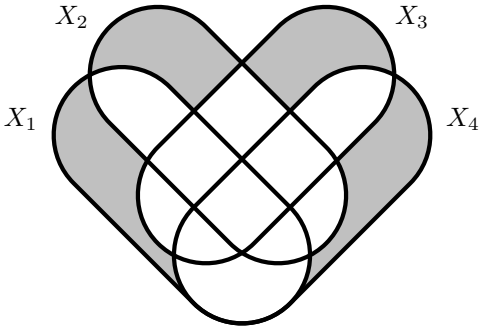
(b) Multivariate Mutual Information



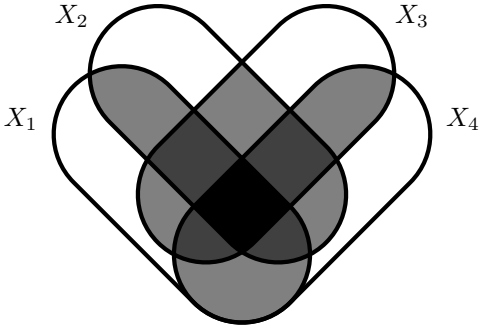
(c) Total Correlation



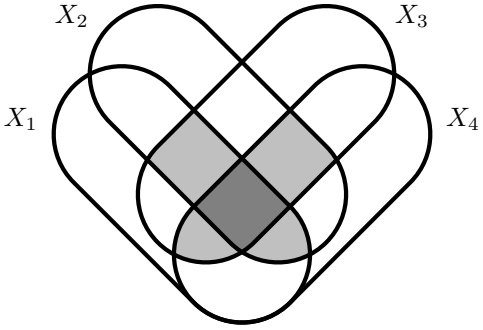
(d) Binding Information



(e) Residual Entropy



(f) Local Exogenous Information



(g) Enigmatic Information

Figure 3.7: Four-variable information diagrams for the multivariate information measures of Sec. 3.5. Darker shades of gray denote heavier weighting in the corresponding informational sum. For example, the atoms to which all four variables contribute are added thrice to the total correlation and so the central atom's weight $\mathrm{I}[X_1; X_2; X_3; X_4] = 3$.

This exposes two pieces: h_μ, the part exterior to the past, and ρ_μ, the part interior. This partitioning has been well studied in information theory due to how it naturally arises as one observes a sequence. This decomposition is displayed in Fig. 3.9(a).

Taking a step back and including the future in the diagram, we obtain a more detailed understanding of how information is transmitted in a process. The past and the future together divide $\mathrm{H}[X_0]$ into four parts; see Fig. 3.9(b). We will discuss each part shortly. First, however, we draw out a different decomposition—that into r_μ and w_μ as seen in Fig. 3.9(c). From this diagram it is easy to see the semantic meaning behind the decomposition: r_μ being divorced from any temporal structure, while w_μ is steeped in it.

We finally turn to the partitioning shown in Fig. 3.9(b). The process I-diagram makes it rather transparent in which sense r_μ is an amount of *ephemeral* information: its atom lies outside both the past and future sets and so it exists only in the present moment, having no repercussions for the future and being no consequence of the past. It is the amount of information in the present observation neither communicated to the future nor from the past. Ref. [Abd12] referred to this as the residual entropy *rate*, as it is the amount of uncertainty that remains in the present even after accounting for every other variable in the time series.

Ref. [Abd12] also proposed to use b_μ as a measure of structural complexity [Abd12], and we tend to agree. The argument for this is intuitive: b_μ is an amount of information that is present now, is not explained by the past, but has repercussions in the future. That is, it is the portion of the entropy rate h_μ that has consequences. In some contexts one may prefer to employ the ratio b_μ/h_μ when b_μ is interpreted an indicator of complex behavior since, for a fixed b_μ, larger h_μ values imply less temporal structure in the time series.

Due to stationarity, the mutual information $\mathrm{I}[X_0; X_{1:}|X_{:0}]$ between the present X_0 and the future $X_{1:}$ conditioned on the past $X_{:0}$ is the same as the mutual information $\mathrm{I}[X_0; X_{:0}|X_{1:}]$ between X_0 and the past $X_{:0}$ conditioned on the future $X_{1:}$. Moreover, both are b_μ. This lends a symmetry to the process I-diagram that does not exist for nonstationary processes. Thus, b_μ atoms in Fig. 4.1 are the same size.

There are two atoms remaining in the process I-diagram that have not been discussed in literature. Both merit attention. The first is q_μ—the information shared by the past, the present, and the future. Notably, its value can be negative and we discuss this further below in Sec. 3.7.2. The other piece, denoted σ_μ, is a component of information shared between the past and the future that *does not exist in the present* observation. This piece is vital evidence that attempting to understand a process without using a model

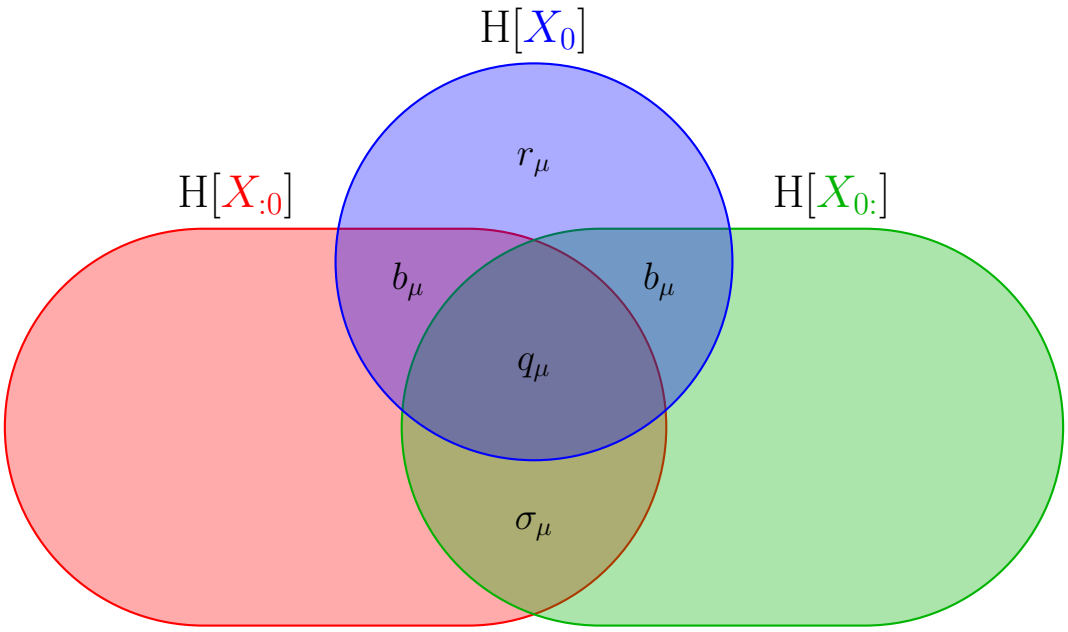


Figure 3.8: I-diagram anatomy of $\mathrm{H}[X_0]$ in the full context of time: The past $X_{:0}$ partitions $\mathrm{H}[X_0]$ into two pieces: h_μ and ρ_μ. The future $X_{0:}$ then partitions those further into r_μ, two b_μs, and q_μ. This leaves a component σ_μ, shared by the past and the future, that is not in the present X_0.

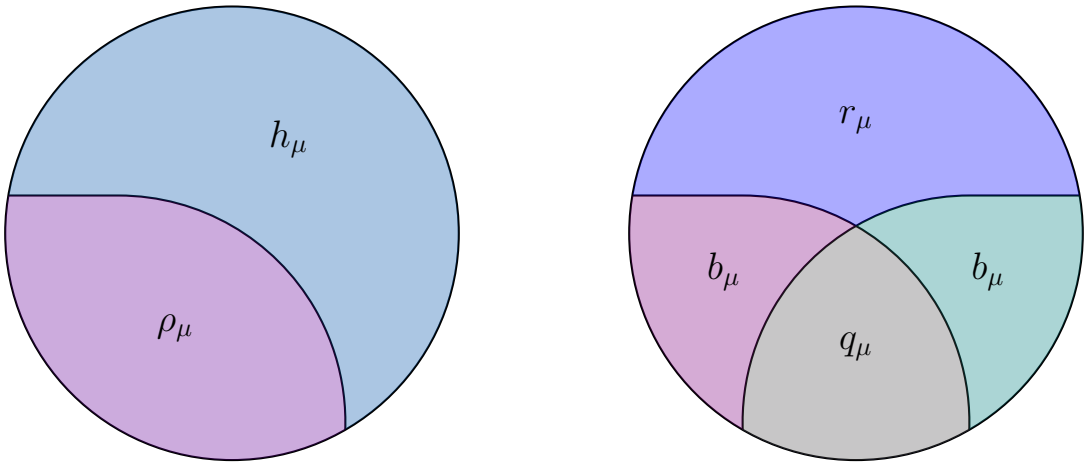
for its generating mechanism is ultimately incomplete. We discuss this point further in Sec. 3.7.2 below.

Negativity of q_μ

The sign of q_μ holds valuable information. To see what this is we apply the partial information decomposition [Wil10] to further analyze $w_\mu = \mathrm{I}[X_0; X_{:0}, X_{1:}]$—that portion of the present shared with the past and future. By decomposing w_μ into four pieces—three of which are unique—we gain greater insight into the value of q_μ and also draw out potential asymmetries between the past and the future.

The partial information lattice provides us with a method to isolate (i) the contributions $\Pi_{\{X_{:0}\}\{X_{1:}\}}$ to w_μ that both the past and the future provide redundantly, (ii) parts $\Pi_{\{X_{:0}\}}$ and $\Pi_{\{X_{1:}\}}$ that are uniquely provided by the past and the future, respectively, and (iii) a part $\Pi_{\{X_{:0},X_{1:}\}}$ that is synergistically provided by both the past and the future. Note that, due to stationarity, $\Pi_{\{X_{:0}\}} = \Pi_{\{X_{1:}\}}$. We refer to this as the *uniquity* and denote it ι.

Using Ref. [Wil10] we see that q_μ is equal to the *redundancy* minus the *synergy* of the past and the future, when determining the present. Thus, if $q_\mu > 0$, the past and future predominantly contribute information to the present. When $q_\mu < 0$, however, considering the past and the future separately in determining the present misses essential correlations. The latter can be teased out if the past and future



(a) Human

(b) Tralfamadorian

(c) Structural

(d) Intuitive

Figure 3.9: Four decompositions of $\mathrm{H}[X]$ from Fig. 3.2, each corresponding to different interpretations of a process.

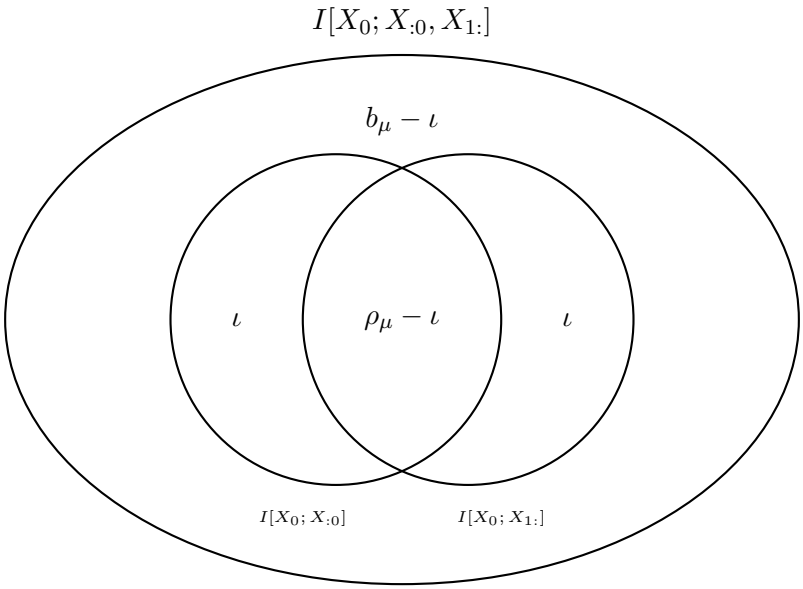


Figure 3.10: Partial information decomposition of $w_\mu = \mathrm{I}[X_0; X_{:0}, X_{1:}]$. The multivariate mutual information q_μ is given by the redundancy $\Pi_{\{X_{:0}\}\{X_{1:}\}}$ minus the synergy $\Pi_{\{X_{:0}, X_{1:}\}}$. $w_\mu = \rho_\mu + b_\mu$ is the sum of all atoms in this diagram.

are considered together.

The process I-diagram (Fig. 4.1) showed that the mutual information between the present and either the past or the future is ρ_μ. One might suspect from this that the past and the future provide the same information to the present, but this would be incorrect. Though they provide the same *quantity* of information to the present, *what* that information conveys can differ. This is evidence of a process's structural irreversibility; cf. Refs. [Cru09, Ell09]. In this light, the redundancy $\Pi_{\{X_{:0}\}\{X_{1:}\}}$ between the past and future when considering the present is $\rho_\mu - \iota$. Furthermore, the synergy $\Pi_{\{X_{:0}, X_{1:}\}}$ provided by the past and the future is equal to $b_\mu - \iota$.

Taking this all together, we find what we already knew: that $q_\mu = \rho_\mu - b_\mu$, The journey to this conclusion, however, provided us with deeper insight into what negative q_μ means and into the structure of w_μ and the process as a whole.

Consequence of σ_μ: Why we model

Notably, the final piece of the process I-diagram is not part of $\mathrm{H}[X_0]$—not a component of the information in a single observation. This is σ_μ, which represents information that is transmitted from the past to the future, but does not go through the currently observed symbol X_0. This is readily understood and leads to an important conclusion.

If one believes that the process under study is generated according to the laws of physics, then the process's internal physical configuration must store all the information from the past that is relevant for generating future behavior. Only when the observed process is order-1 Markov is it sufficient to keep track of just the current observable. For the plethora of processes that are not order-1 or that are non-Markovian altogether, we are faced with the fact that information relevant for future behavior must be stored somehow. And, this fact is reflected in the existence of σ_μ. When $\sigma_\mu > 0$, a complete description of the process requires accounting for this internal configurational or, simply, *state information*. This is why we build models and cannot rely on only collecting observation sequences.

The amount of information shared between $X_{:0}$ and $X_{1:}$, but ignoring X_0, was previously discussed in Ref. [Bal10]. We now see that the meaning of this information quantity — there denoted $\mathcal{I}_1$ — is easily gleaned from its components: $\mathcal{I}_1 = q_\mu + \sigma_\mu$.

Furthermore, in Refs. [Abd12], [Abd10], and [Bal10], efficient computation of b_μ and $\mathcal{I}_1$ were not provided and the brute force estimates are inaccurate and very compute intensive. Fortunately, by a direct extension of the methods developed in Ref. [Ell09] on bidirectional machines, we can easily compute both $r_\mu = \mathrm{H}[X_0|\mathcal{S}_0^+, \mathcal{S}_1^-]$ and $\mathcal{I}_1 = \mathrm{I}[\mathcal{S}_0^+, \mathcal{S}_1^-]$. This is done by constructing joint probabilities of forward-time and reverse-time causal states — $\{\mathcal{S}^+\}$ and $\{\mathcal{S}^-\}$, respectively — at different time indices employing the dynamic of the bidirectional machine. This gives closed-form, exact methods of calculating these two measures, provided one constructs the process's forward and reverse ϵ-machines. b_μ follows directly in this case since it is the difference of h_μ and r_μ; the former is also directly calculated from the ϵ-machine.

Decompositions of $\mathbf{E}$

Using the process I-diagram and the tools provided above, three unique decompositions of the excess entropy, Eq. (3.19), can be given. Each provides a different interpretation of how information is transmitted from the past to the future.

The first is provided by Eqs. (3.37)-(3.40). The subextensive parts of the block entropy and total cor-

relation there determine the excess entropy decomposition. We have:

$$\begin{aligned}
\mathbf{E} =& \mathbf{E}_B + \mathbf{E}_R && (3.49)\\
=& -\mathbf{E}_B - \mathbf{E}_Q && (3.50)\\
=& \frac{1}{2}(\mathbf{E}_R - \mathbf{E}_Q) && (3.51)\\
=& -\frac{1}{2}(\mathbf{E}_W + \mathbf{E}_Q)\ . && (3.52)
\end{aligned}$$

We leave the meaning behind these decompositions as an open problem, but do note that they are distinct from those discussed next.

The second and third decompositions both derive directly from the process I-diagram of Fig. 4.1. Without further work, one can easily see that the excess entropy breaks into three pieces, all previously discussed:

$$\mathbf{E} = b_\mu + q_\mu + \sigma_\mu\ . \tag{3.53}$$

And, finally, one can perform the partial information decomposition on the mutual information $\mathrm{I}[X_{:0}; X_0, X_{1:}]$. The result gives an improved understanding of (i) how much information is uniquely shared with the either the immediate or the more distant future and (ii) how much is redundantly or synergistically with both.

The decompositions provided by the atoms of the process I-diagram and those provided by the subextensive rates of block-information curves are conceptually quite different. It has been shown [CRU03] that the subextensive part of the block entropy and the mutual information between the past and the future, though equal for one dimensional processes, differ in two dimensions. We believe the semantic differences shown here are evidence that the degeneracy of alternate $\mathbf{E}$-decompositions breaks in higher dimensions.

3.8 Examples

We now make the preceding concrete by calculating these quantities for three different processes, selected to illustrate a variety of informational properties. Figure 3.11 gives each process via it's ϵ-machine [SHA01]: the Even Process, the Golden Mean Process, and the Noisy Random Phase-Slip (NRPS) Process. A process's ϵ-machine consists of its *causal states*—a partitioning of infinite pasts into sets that give rise to the same predictions about future behavior. The state transitions are labeled $p|s$ where s is the observed symbol and p is the conditional probability of observing that symbol given the state the process is in. The ϵ-machine

representation for a process is its minimal unifilar presentation.

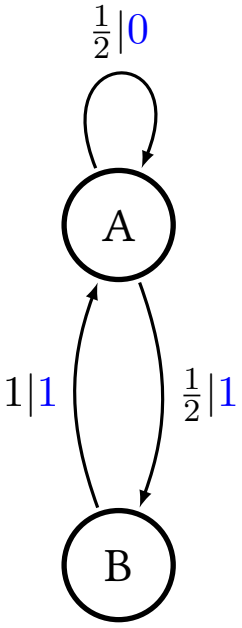
Table 3.1 begins by showing the single-observation entropy $\mathrm{H}[1]$ followed by h_μ and ρ_μ. Note that the Even and the Golden Mean Processes cannot be differentiated using these measures alone. The table then follows with the finer decomposition. We now see that the processes can be differentiated. We can understand fairly easily that the Even Process, being infinite-order Markovian, and consisting of blocks of 1s of even length separated by one or more 0s, exhibits more structure than the Golden Mean Process. (This is rather intuitive if one recalls that the Golden Mean Process has only a single restriction: it cannot generate sequences with consecutive 0s.) We see that, for the Even Process, r_μ is 0. This can be understood by considering a bi-infinite sample from the Even Process with a single gap in it. The structure of this process is such that we can always and immediately identify what that missing symbol must be.

These two processes are further differentiated by q_μ, where it is negative for the Even Process and positive for the Golden Mean Process. On the one hand, this implies that there is a larger amount of synergy than redundancy in the Even Process. Indeed, it is often the case, when appealing only to the past or the future, that one cannot determine the value of X_0, but when taken together the possibilities are limited to a single symbol. On the other hand, since q_μ is positive for the Golden Mean Process we can determine that its behavior is dominated by redundant contributions. That w_μ is larger for the Even Process than the Golden Mean Process is consonant with the impression that the former is, overall, more structured.

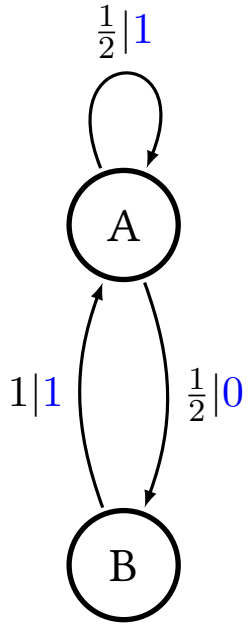
The next value in the table is σ_μ, the amount of state information not contained in the current observable. This vanishes for the Golden Mean Process, as it is order-1 Markovian. The Even Process, however, has a significant amount of information stored that is not observable in the present.

Last in the table is a partial information decomposition of $\mathrm{I}[X_0; X_{:0}, X_{1:}]$. q_μ is given by $\Pi_{\{X_{:0}\}\{X_{1:}\}} - \Pi_{\{X_{:0}, X_{1:}\}}$. Of note here is that the NRPS process's nonzero uniquity $\iota = 0.02437$. For the Even and Golden Mean Processes it vanishes. That is, in the NRPS Process information is uniquely communicated to the present from the past and an equivalent in magnitude, but different, information is communicated to the future. Thus, the NRPS Process illustrates a subtle asymmetry in statistical structure.

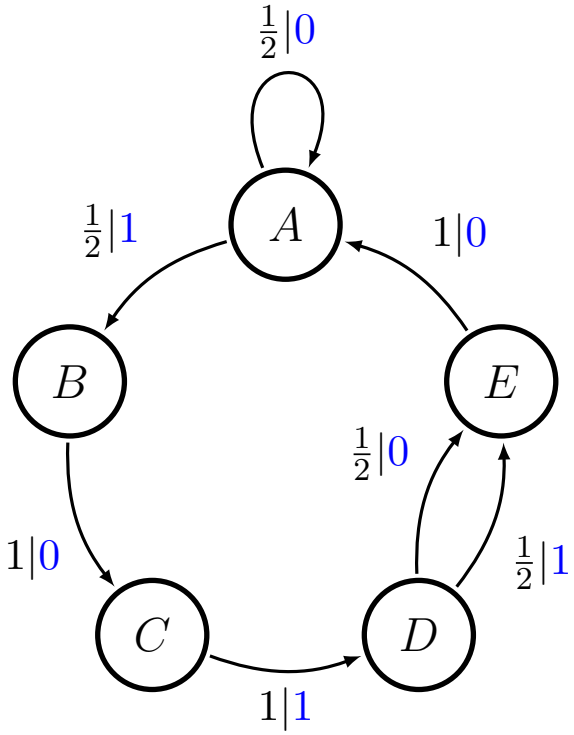
Table 3.2 then provides an alternate breakdown of $\mathbf{E}$ for each prototype process. We use this here to only highlight how much the processes differ in character from one another. The consequences of the first decomposition of excess entropy—$\mathbf{E} = b_\mu + q_\mu + \sigma_\mu$—follow directly from the previous table's discussion. The second and third decompositions into $\mathbf{E}_R + \mathbf{E}_B$ and $-\mathbf{E}_B - \mathbf{E}_Q$ vary from one another significantly.



(a) Even Process



(b) Golden Mean Process



(c) Noisy Random Phase-Slip Process

Figure 3.11: ϵ-Machine presentations for the three example processes.

	Even	Golden Mean	NRPS
$H[1]$	0.91830	0.91830	0.97987
h_μ	0.66667	0.66667	0.50000
ρ_μ	0.25163	0.25163	0.47987
r_μ	0.00000	0.45915	0.16667
b_μ	0.66667	0.20752	0.33333
q_μ	-0.41504	0.04411	0.14654
w_μ	0.91830	0.45915	0.81320
σ_μ	0.66667	0.00000	1.09407
$\Pi_{\{X_{:0}\}\{X_{1:}\}}$	0.25163	0.25163	0.45550
$\iota : \Pi_{\{X_{:0}\}}, \Pi_{\{X_{1:}\}}$	0.00000	0.00000	0.02437
$\Pi_{\{X_{:0},X_{1:}\}}$	0.66667	0.20752	0.30896

Table 3.1: Information measure analysis of three processes.

	Even	Golden Mean	NRPS
$\mathbf{E}$	0.91830	0.25163	1.57393
b_μ	0.66667	0.20752	0.33333
q_μ	-0.41504	0.04411	0.14654
σ_μ	0.66667	0.00000	1.09407
$\mathbf{E}_R$	4.48470	0.41504	1.55445
$\mathbf{E}_B$	-3.56640	-0.16341	0.01948
$\mathbf{E}_Q$	2.64810	-0.08822	-1.59342
$\mathbf{E}_W$	-4.48470	-0.41504	-1.55445
$\Pi_{\{X_0\}\{X_{1:}\}}$	0.25163	0.04411	0.47987
$\Pi_{\{X_0\}}$	0.00000	0.20752	0.00000
$\Pi_{\{X_{1:}\}}$	0.00000	0.00000	0.76073
$\Pi_{\{X_0,X_{1:}\}}$	0.66667	0.00000	0.33333

Table 3.2: Alternative decompositions of excess entropy $\mathbf{E}$ for the three prototype processes.

The Even Process has much larger values for these pieces than the total $\mathbf{E}$, whereas the NRPS process has two values nearly equal to $\mathbf{E}$ and one very small. The Golden Mean Process falls somewhere between these two.

The final excess entropy breakdown is provided by the partial information decomposition of $\mathrm{I}[X_{:0}; X_0, X_{1:}]$. Here, we again see differing properties among the three processes. The Even Process consists only of redundancy $\Pi_{\{X_{:0}\}\{X_{1:}\}}$ and synergy $\Pi_{\{X_{:0},X_{1:}\}}$. The Golden Mean Process contains no synergy, a small amount of redundancy, and most of its information sharing is with the present uniquely. The NRPS Process possesses both synergy and redundancy, but also a significant amount of information shared solely with the more distant future.

And, finally, Fig. 3.12 plots how h_μ partitions into r_μ and b_μ for the Golden Mean family of processes.

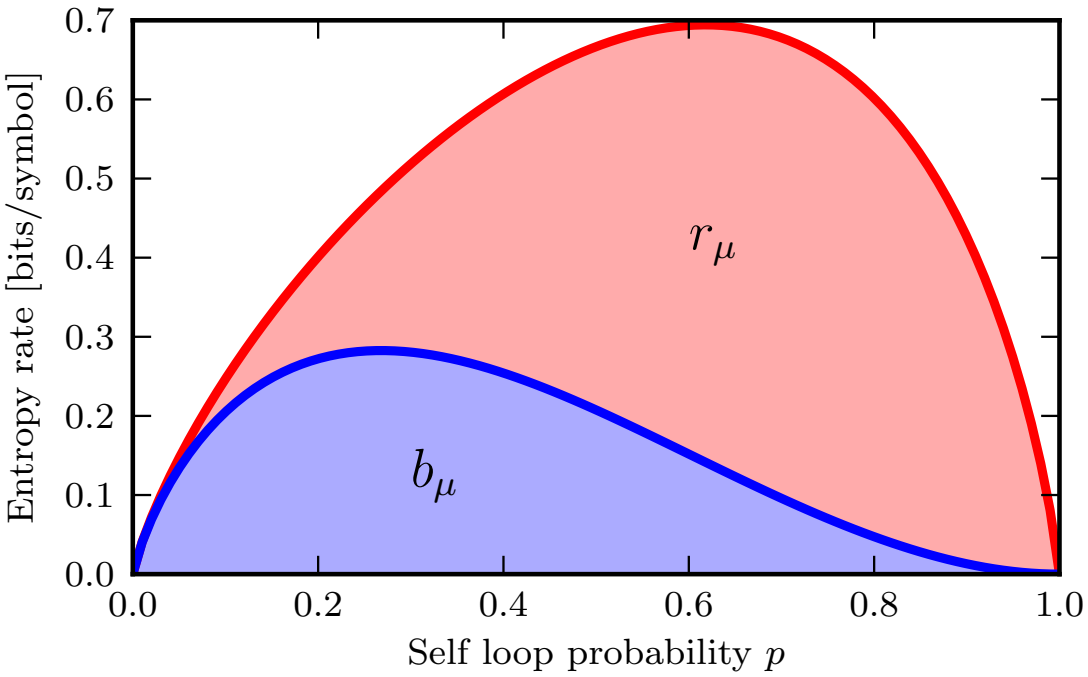


Figure 3.12: The breakdown of h_μ for the Golden Mean Process. The self-loop probability was varied from 0 to 1, adjusting the other edge's probability accordingly.

This family consists of all processes with ϵ-machine structure given in Fig. 3.11(b), but where the outgoing transition probabilities from state **A** are parametrized. We can easily see that for small self-loop transition probabilities, the majority of h_μ is consumed by b_μ. This should be intuitive since, when the self-loop probability is small, the process is nearly periodic and r_μ should be nearly zero. On the other end of the spectrum, when the self-loop probability is large, h_μ is mostly consumed by r_μ. This is again intuitive since observations from that process are dominated by 1s and the occasional 0—which provides all the entropy for h_μ—has no effect on structure.

3.9 Concluding Remarks

We began by outlining a conceptual decomposition of a single observation in a time series: a single observation contains a hierarchy of informational components. We then made the decomposition concrete using a variety of multivariate information measures. Adapting them to time series, we showed that their asymptotic growth rates are identified with the hierarchical decomposition. To unify the various competing views, we provided the measurement-centric process I-diagram, demonstrating that it concisely reveals the semantic meaning behind each component in the hierarchy.

Once the measurement-centric process I-diagram was available, we isolated two components, analyzing in detail their meaning. We utilized the partial information lattice [Wil10] to refine our understanding of when the past and the future redundantly and synergistically inform the present. This allowed us to explain a subtle statistical asymmetry—the directionality in the difference between ρ_μ and $\Pi_{\{X_{:0}\}\{X_{1:}\}}$.

The other atom we singled out in the process I-diagram was σ_μ. It is the most compelling evidence that analyzing a process from its measurements alone, without constructing a state-based model, is ultimately limited. We also quickly stated how a model of the type studied in computational mechanics can be used to give closed-form expressions for many of the quantities discussed which allow for exact calculation of their values.

Next, we discussed how the different methods and measures relate to one of the most widely used complexity measures—the past-future mutual information or excess entropy. In particular, we showed how they yield four distinct decompositions and, in some cases, give useful interpretations of what these decompositions mean operationally.

Then, we calculated all the measures for three different prototype processes, each highlighting particular features of the information-theoretic decompositions. We gave interpretations of negative mutual informations, as seen in q_μ. The interpretations were consistent, understandable, and insightful. In light of the partial information decomposition, there is nothing untoward about negative informations.

By adapting it to the time series setting, we highlighted a key weakness of the total correlation (or multi-information). This undoubtedly explains the lack of interest in using it in the time series setting, though the weakness still holds when it is used to analyze any group of random variables. The weakness has led to persistent over-interpretations of what it describes. It also may have eclipsed the importance of its more complete analog, such as the block entropy, in the settings of networked random variables.

We believe that the decompositions detailed here provide useful tools for information-theoretic time series analysis, and that r_μ and b_μ potentially have many interesting uses due to their data-centric nature. For example, it is possible that b_μ is useful in constructing error-correcting codes since a time series with small $\frac{r_\mu}{h_\mu}$ has a high recovery rate from single-site deletions. This is because if there is no entropy in the value of a site given its neighbors (the past and the future), there is only one value that could have been in that site while still respecting the correlations present. This can be seen clearly with the Even process.

In closing, we take a longer view. There is an exponential number of possible atoms for N-way information measures. In addition, there is a similarly large number possible partial information decom-

positions for N variables. This diversity presents the possibility of a large number of independent efforts to define and uniquely motivate why one or the other information measure is the best. Indeed, many of these yet-to-be-explored measures may be useful. In this light, there is a bright future for developing information measures adapted to a wide range of nonlinear, complex systems. And, helpfully, a unifying framework appears to be emerging.

Chapter 4

Forgetting and Remembering in Chaotic Dynamical Systems: A Novel Measure of Information Processing in Chaos

4.1 Summary

One of the most fundamental measures of a chaotic dynamical system is its Lyapunov exponent. Utilizing Pesin's relation and a recently introduced measure of stochastic complexity, we show that the Lyapunov exponent can be naturally decomposed in to two semantically meaningful components and so is not as fundamental as it may seem. Of the information generation in a chaotic system, we show that some of it is "forgotten" and some is "remembered". We associate the remembered component with intrinsic computation. This framework is then applied to the logistic, tent, and Lozi maps demonstrating hitherto unknown features.

4.2 Introduction

Many systems generate information. Ants form intricate, structured nests. Magnets form complex domain structures. Music compositions weaves theme and structure with surprise and innovation. The population dynamics of species can be chaotic. Here we focus on one- and two-dimensional discrete time chaotic maps which can model many real-world systems.

While many systems generate information via *stochasticity* — thermal fluctuations, for example — deterministic chaotic dynamical systems generate information via the stretching and folding of their phase space leading to their signature sensitivity to initial conditions. However, this information generation has long eluded a more rigorous understanding as to its *quality*.

To ground what we mean by this, consider two systems. The first, a coin: each flip independent of the others leading to simple, uncorrelated randomness. The second system, a stock: while its price

may fluctuate, the direction and magnitude of those fluctuations tell us much of the systems following behavior. Here we make this distinction rigorous, dividing the entropy generation of a chaotic map into a component that is relevant to temporal structure and a component which is divorced from it. We believe the first component — the part that is structurally relevant — is the observation of the systems internal information processing, and therefore is of practical interest to those wishing to harness such systems for useful means (See [Dit10], for example).

The organization of this paper is as follows: in section 4.3 we define and provide intuition for the information measures. Then in section 4.4 we dive directing into their calculation for a selection of discrete time maps: the logistic map, the tent map, and the Lozi map. Finally in section 4.5 we summarize our findings and speculate about their implications and future applications.

4.3 Anatomy

In an effort to remain as general as possible we analyze systems as *processes*, bi-infinite sequences of random variables with shift-invariant statistics: $\ldots, X_{-2}, X_{-1}, X_0, X_1, X_2, \ldots$. Specifically for our domain of chaotic dynamical systems, a standard generating partition will produce a process. We denote a contiguous block of observations beginning at index t and extending for a length ℓ by $X_{t:t+\ell}$. Note that this notation is closed on the left and open on the right. To denote semi-infinite blocks we will leave one index blank, for example $X_{:t}$ is all symbols up to but excluding t. Similarly, $X_{t:}$ is all observations from t onward. This implies that the process itself can be denoted $X_{:}$.

Here we wish to segment our bi-infinite sequence into three components. We first isolate a single observation, for simplicity X_0, and refer to it as the *present*. Everything prior to this, $X_{:0}$, is therefore the *past*, and everything after, $X_{0:}$, is the *future*. The information-theoretic relationships between these three random variables are expressed in Fig. 4.1, known as an I-diagram [Cov06].

The rate of information generation in a process is the amount of new information in an observation given all the information in prior observations [Cov06]:

$$h_\mu = \mathrm{H}[X_0|X_{:0}] \tag{4.1}$$

This quantity arises in many contexts and goes by many names: the Shannon entropy rate, the Kolmogorov-Sinai entropy, and the metric entropy, for example. The dual of the entropy rate is the predicted informa-

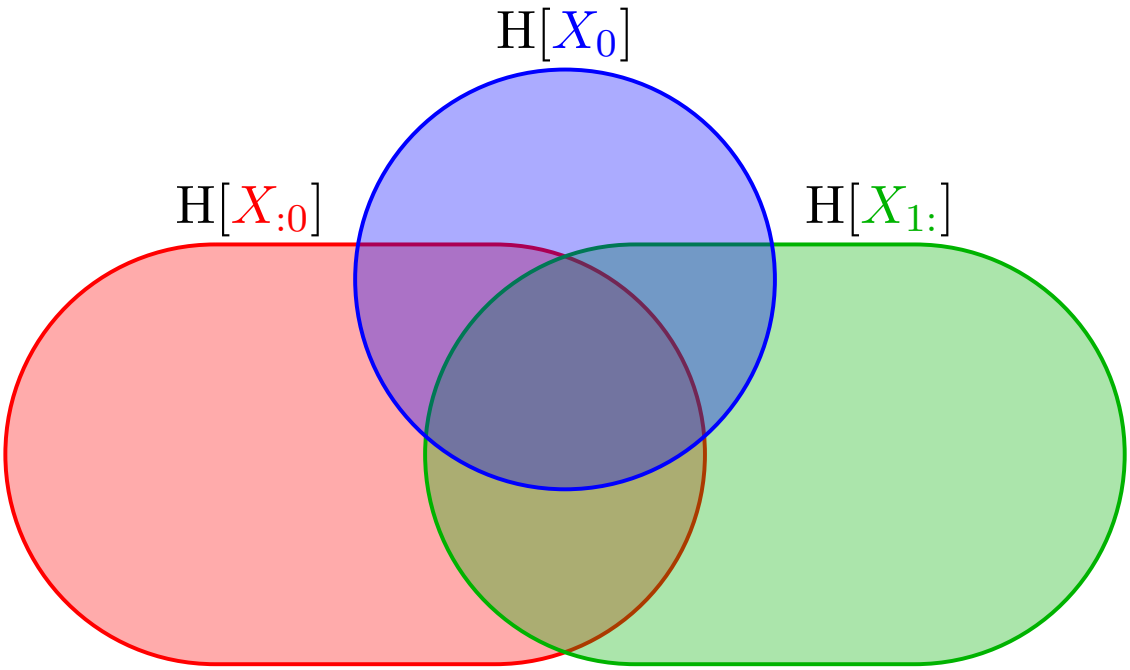


Figure 4.1: The I-Diagram of the past, present and future. There exist seven distinct ways that information can be partitioned among the three temporal variables. Here we focus on the four regions contained in the *present*, X_0, represented as a blue circle here.

tion:

$$\rho_\mu = \mathrm{I}[X_{:0} : X_0] \tag{4.2}$$

It is the information in present that can be predicted from prior observations.

Although counter-intuitive, by simple application of a chain rule it can be shown that:

$$\mathrm{H}[X_0|X_{:0}] = \mathrm{I}[X_0 : X_{0:}|X_{:0}] + \mathrm{H}[X_0|X_{:0}, X_{0:}] \tag{4.3}$$

$$h_\mu = b_\mu + r_\mu \tag{4.4}$$

That is, the entropy rate is equal to the information shared by the present and the future given the past, plus the information in the present given both the past and the future.

This decomposition was at least partially studied first by Verdú and Weissman[Ver08] where they defined the *erasure entropy* (there denoted H^-, here r_μ) for the analysis of erasure channels. Here we refer to this quantity as the *ephemeral information*, as it is information that exists only in a single moment. The other component of the decomposition, *predictive information rate* b_μ, here referred to as the *bound*

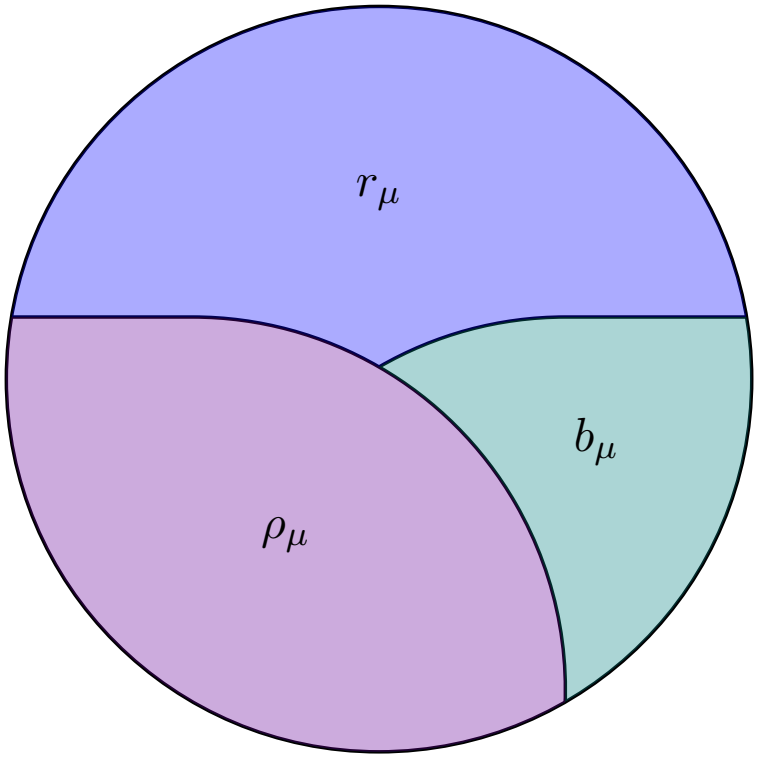


Figure 4.2: Decomposition of the *present*, X_0, into r_μ, b_μ, and ρ_μ. ρ_μ, the component which overlaps with the *past*, $X_{:0}$, and represents *redundant* information, which could be predicted. r_μ is the component which is outside both the *past*, $X_{:0}$, and the *future*, $X_{0:}$, and represents *ephemeral* information, which does not contribute to temporal structure. The remaining component, b_μ, is the component in the future but not in the past, which represents new, structural information — we refer to this as *bound*.

information[1], was first studied as a measure of interestingness in computational musicology by Abdallah and Plumbley[Abd12], where they also suggested its use as a general measure of complexity. For a more complete analysis of this decomposition as well as methods of computation and related measures, see James et al[Jam11].

Isolating the information contained in the present, $\mathrm{H}[X_0]$, from Fig. 4.1 and labeling the decomposition, we get Fig. 4.2. This gives us a particularly intuitive way of thinking about the information contained in an observation. Some of the behavior can be predicted, ρ_μ, the rest can not be predicted (h_μ). Of that which can not be predicted, some plays a role in the future behavior, b_μ, and some does not, r_μ. This is a natural decomposition provided by the time series resulting in a semantic understanding of the entropy rate.

By way of example, let us consider a few simple processes and how they decompose into these three

[1]We refer to it as the bound information due to criticisms of the term *predictive information* which are beyond the scope of this paper.

values. A periodic process of alternating 0s and 1s has $\mathrm{H}[X_0] = 1$ since 0s and 1s occur equally often. Given a prior observation, one can accurately predict exactly which symbol will occur next, and so $\mathrm{H}[X_0] = \rho_\mu = 1$. On the other extreme is a coin flip. Again each outcome is equally likely and so $\mathrm{H}[X_0] = 1$. In this case, however, each flip is independent of all others and so $\mathrm{H}[X_0] = r_\mu = 1$.

In between lie the interesting processes: those with *stochastic structure*. These typically will have all three components of the decomposition non-zero. These patterns are what separate meaning from noise. Not being purely predictable nor independently random, these patterns are captured by b_μ. The more intricate the pattern, the larger b_μ. The generation of these patterns requires computation, and so we propose the use of b_μ as a simple method of discovering intrinsic computation: where there are intricate patterns, there is sophisticated processing.

4.4 Maps

We now turn our attention to discrete time chaotic dynamical systems. To enable us to analyze these systems we call upon Pesin's theorem[Pes77] which states that the entropy rate of a system is equal to the sum of its positive Lyapunov exponents. Since the maps considered here have but a single positive Lyapunov exponent, that alone is equal to the entropy rate. Since we know how to decompose the entropy rate into these semantically meaningful, we can therefore decompose the Lyapunov exponent.

For each of these example systems we utilize the standard generating partition of:

$$x_n = \begin{cases} 0 & \text{if } s_n < \frac{1}{2} \\ 1 & \text{if } s_n \geq \frac{1}{2} \end{cases} \tag{4.5}$$

and generate an extremely long sample ($\approx 10^{10}$). From this long sequence, we use a sliding window technique with a window size appropriate for the given Lyapunov exponent (e.g. when the system has a low Lyapunov exponent we use a much longer window size than when the system is fully chaotic), with a minimum window size of 31. From these windows we partition them into a past, present, and future. From this partition we approximate both r_μ and b_μ. The minimum window size and scaling of the window size were chosen such that the numerical values obtained vary by less than 0.01% when the window size is incremented.

The logistic map is perhaps the most studied chaotic system. It is defined by the following mapping:

$$x_{n+1} = r x_n (1 - x_n) \tag{4.6}$$

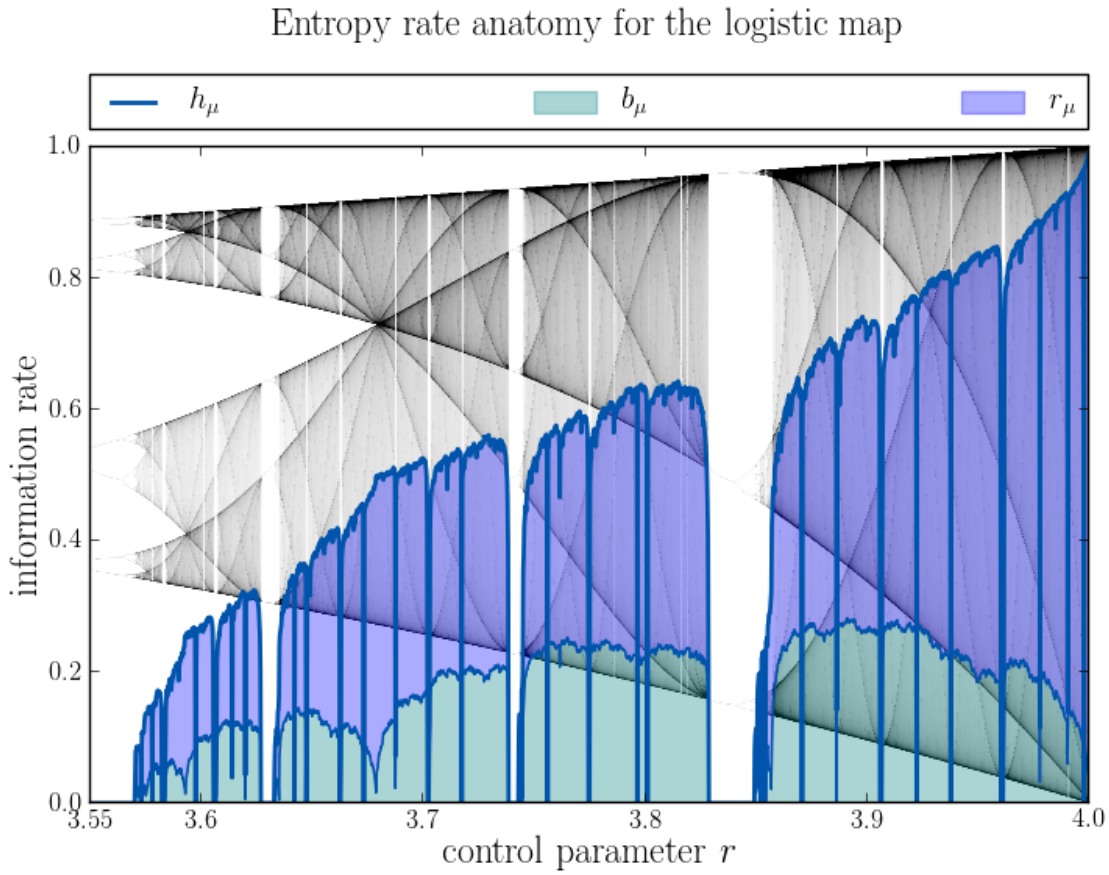


Figure 4.3: The anatomy decomposition applied to the logistic map. The map's bifurcation diagram is in the background for reference. The lower component, green in color, is b_μ, the portion of h_μ that is *bound* and structurally relevant. The upper component, blue in color, is r_μ, the portion of h_μ that is *ephemeral* and structurally irrelevant.

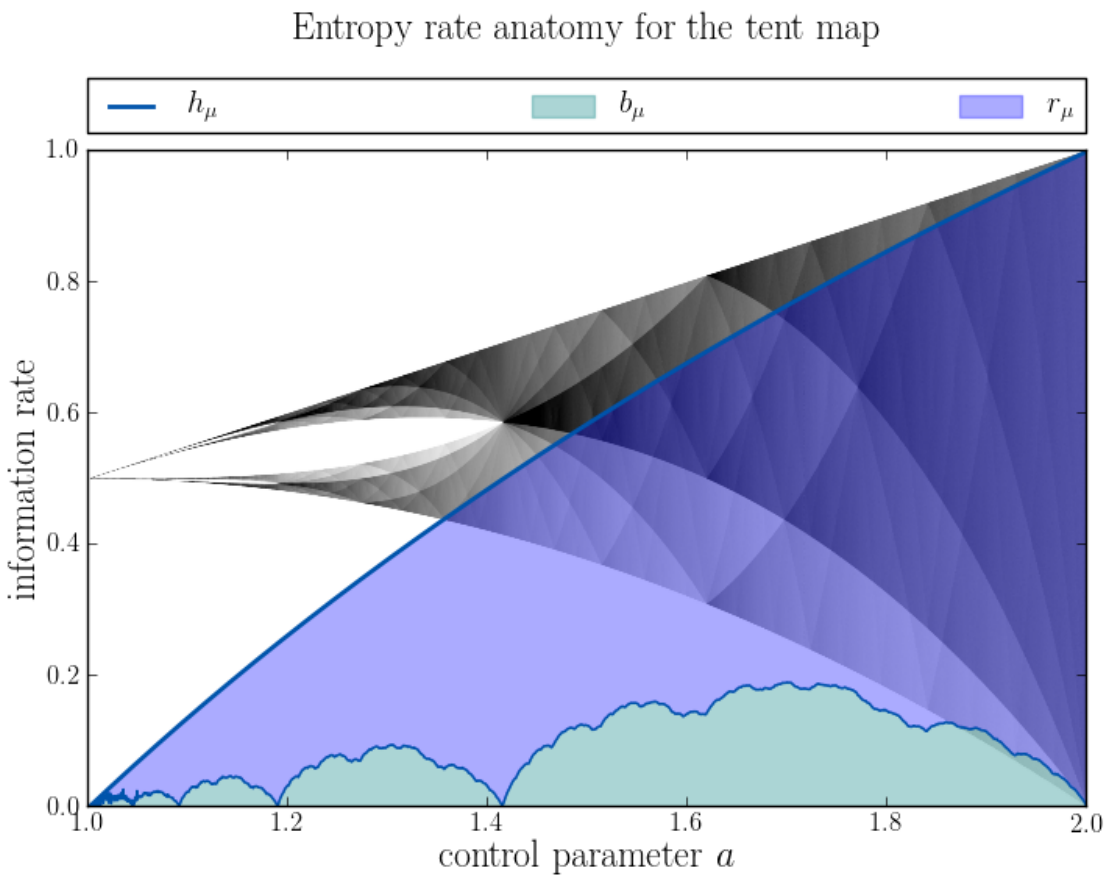


Figure 4.4: The anatomy decomposition applied to the tent map. The map's bifurcation diagram is in the background for reference. The lower component, green in color, is b_μ, the portion of h_μ that is *bound* and structurally relevant. The upper component, blue in color, is r_μ, the portion of h_μ that is *ephemeral* and structurally irrelevant. Note that although the sum of both r_μ and b_μ, h_μ, is a smooth function of a, their decomposition is not.

where r is a control parameter between 0 and 4, and x_n is drawn from the unit interval. The results of our analysis can be seen in Fig. 4.3, where a number of interesting features need to be pointed out.

Firstly, the information generation of a chaotic system is, in fact, a mixture of ephemeral and bound, consequential information at nearly all parameter values. For the most part, the division into these two components varies in non-trivial ways as a function of r. The boundary between the two seems to be non-differentiable, but this in and of itself should not be surprising given that h_μ $(= \lambda)$ is non-differentiable. Furthermore, where band-merging occurs are the only non-trivial parameter values where b_μ is zero.

To discover if the r_μ - b_μ boundary is complex and non-differentiable due to the nature of these measures, or if it is simply a consequence of the fact that the Lyapunov exponent has these properties,

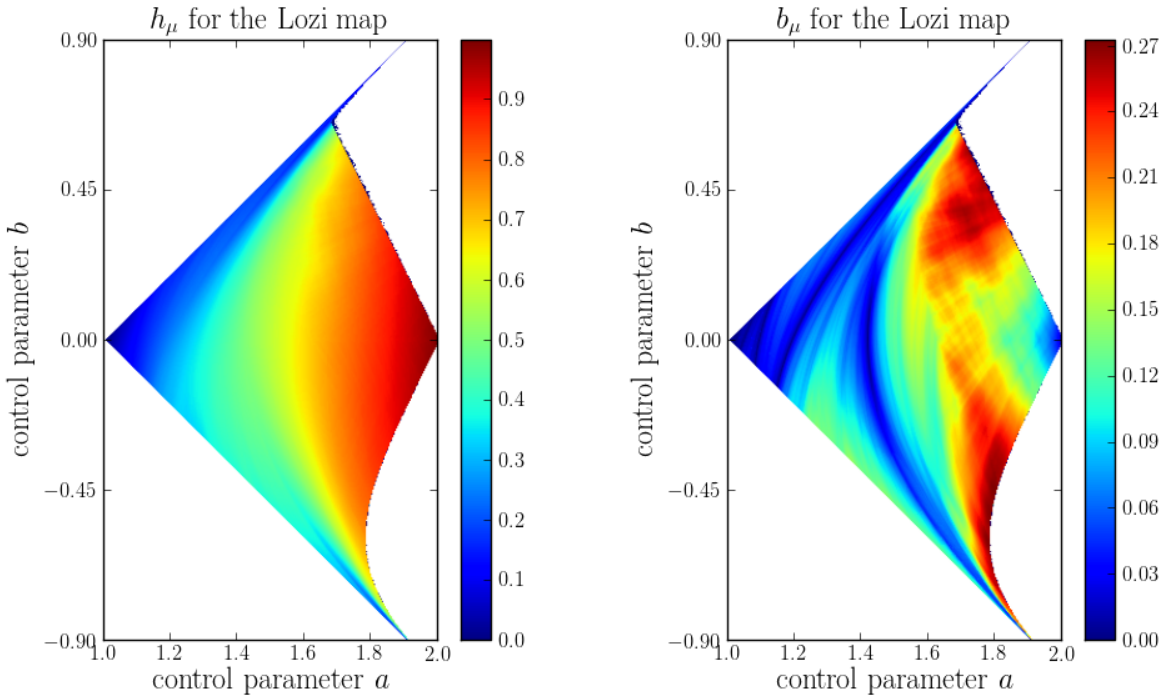


Figure 4.5: The anatomy decomposition of the Lozi map. On the left is h_μ, the positive Lyapunov exponent of the map, and on the right is b_μ (different color scales). The areas of parameter space where b_μ is maximized are on the upper-right and lower-right edges of the attractor region.

we turn to the tent map. The tent map is a (mostly) linearized variant of the logistic map defined by the mapping:

$$x_{n+1} = a\left(1 - 2\left|x_n - \frac{1}{2}\right|\right) \tag{4.7}$$

where a is a control parameter between 0 and 2. The Lyapunov exponent for the tent map is simply $\lambda = \log_2 a$, and so if the decomposition into r_μ and b_μ were a simple function of λ, we'd see a simple division between them in the tent map.

Figure 4.4, however, demonstrates that this can not be true. We see here that, despite the simplicity of h_μ, r_μ and b_μ are again non-differentiable. This fairly definitively implies that these quantities are extracting something both quantitatively and qualitatively new from the system. Once again we find that $h_\mu = r_\mu$ at parameter values that correspond to band mergings in the bifurcation diagram.

These analysis can be extended into two dimensions by considering the Lozi map. The Lozi map is a

two-dimensional extension of the tent map defined by the following mappings:

$$\begin{aligned} x_{n+1} &= 1 - a\,|x_n| + y_n \\ y_{n+1} &= b x_n \end{aligned} \tag{4.8}$$

Note that when $b = 0$ the system is isomorphic to the tent map. The map has a vaguely diamond-shaped attractor region in the region $1.0 < a < 2.0$ and $-0.9 < b < 0.9$.

It is this region we focus on in Fig. 4.5. The entropy rate, h_μ, varies smoothly over the attractor region, whereas b_μ varies greatly. There are swaths of low b_μ that correspond to "fuzzy" band mergings in the bifurcation diagram. It is also interesting to note that the regions of maximal b_μ are off the $b = 0$, which is where maximal h_μ exists.

4.5 Conclusion

We have shown that the entropy rate of a process can be decomposed via a chain rule into two semantically meaningful components. These components, the ephemeral r_μ and the bound b_μ, provided a new insight into the behavior of a system without any modeling or other domain-specific knowledge. We also argue that b_μ is a strong though indirect measurement of internal computation. Due to Pesin's relation, we can import this decomposition into chaotic dynamical systems and break the Lyapunov exponent into those two components.

Applying this decomposition logistic, tent, and Lozi maps details the topography of their internal-computation landscape. We propose that the analysis and usage of this landscape can lead to improved engineering of natural systems such as Chaogates [Dit10], DNA computing, and the like. While a simple geometric interpretation, such as that which exists of the Lyapunov exponent, has thus far eluded us for r_μ and b_μ, we strongly believe that such an interpretation exists do to the natural and straightforward definition and interpretation of these quantities.

4.A Computing b_μ Analytically

Due to the large amount of data required for an accurate calculation of b_μ, it is advantageous to have a method for computing it analytically so as to verify the accuracy of the numerical approximations. This is possible if one can construct a model of the discretized dynamics in both the forward and backwards

directions. From these models b_μ can be computed via:

$$b_\mu = \mathrm{I}[X_0 : \mathcal{S}_0^+ | \mathcal{S}_1^-] \tag{4.9}$$

where $\mathcal{S}_0^+$ is the forward model's state distribution at time 0 and $\mathcal{S}_1^-$ is the backwards model's state distribution at time 1.

We will now explicitly do this calculation for one parameter value of the tent map. In particular, consider the Misiurewicz point where $f^4(\frac{1}{2}) = f^5(\frac{1}{2})$. Solving this gives us:

$$\begin{aligned} a &= \sqrt[3]{\sqrt{\frac{19}{27}} + 1} + \frac{2}{3\sqrt[3]{\sqrt{\frac{19}{27}} + 1}} \\ &= 1.76929235\ldots \end{aligned} \tag{4.10}$$

At this parameter value the tent map admits a Markov partition as seen in Figure. 4.6

As noted in Eq. 4.9, we require a particular form of model to be able to compute b_μ, and so we must transform the model in Fig. 4.7 into one that is *unifilar*, and in particular to the *ϵ-machine*. The result can be seen in Fig. 4.8

Lastly, we construct the *bidirectional* model of the process, seen in Fig. 4.9. From this model we construct $\Pr(\mathcal{S}_0^+, \mathcal{S}_0^-, X_0, \mathcal{S}_1^+, \mathcal{S}_1^-)$ which allows for the calculation of $\mathrm{H}[X_0|\mathcal{S}_0^+, \mathcal{S}_1^-]$ and $\mathrm{I}[X_0 : \mathcal{S}_1^-|\mathcal{S}_0^+]$ which are r_μ and b_μ respectively. Utilizing this structure we find that, for the tent map at a:

$$h_\mu = \log_2 a = \log_2\left(\frac{\sqrt[3]{9+\sqrt{57}} + \sqrt[3]{9-\sqrt{57}}}{3^{\frac{2}{3}}}\right) = 0.823172\ldots \tag{4.11}$$

$$r_\mu = \frac{1}{4}\left(3 - \frac{2}{a+1} - \frac{4}{a+2} + \frac{9}{2a+3}\right) \tag{4.12}$$

$$= \frac{1}{9}\left(\frac{\sqrt[3]{207\sqrt{57} - 1349}}{19^{2/3}} - \frac{32}{\sqrt[3]{19\left(207\sqrt{57} - 1349\right)}} + 7\right) = 0.648258\ldots \tag{4.13}$$

$$b_\mu = h_\mu - r_\mu = 0.174915\ldots \tag{4.14}$$

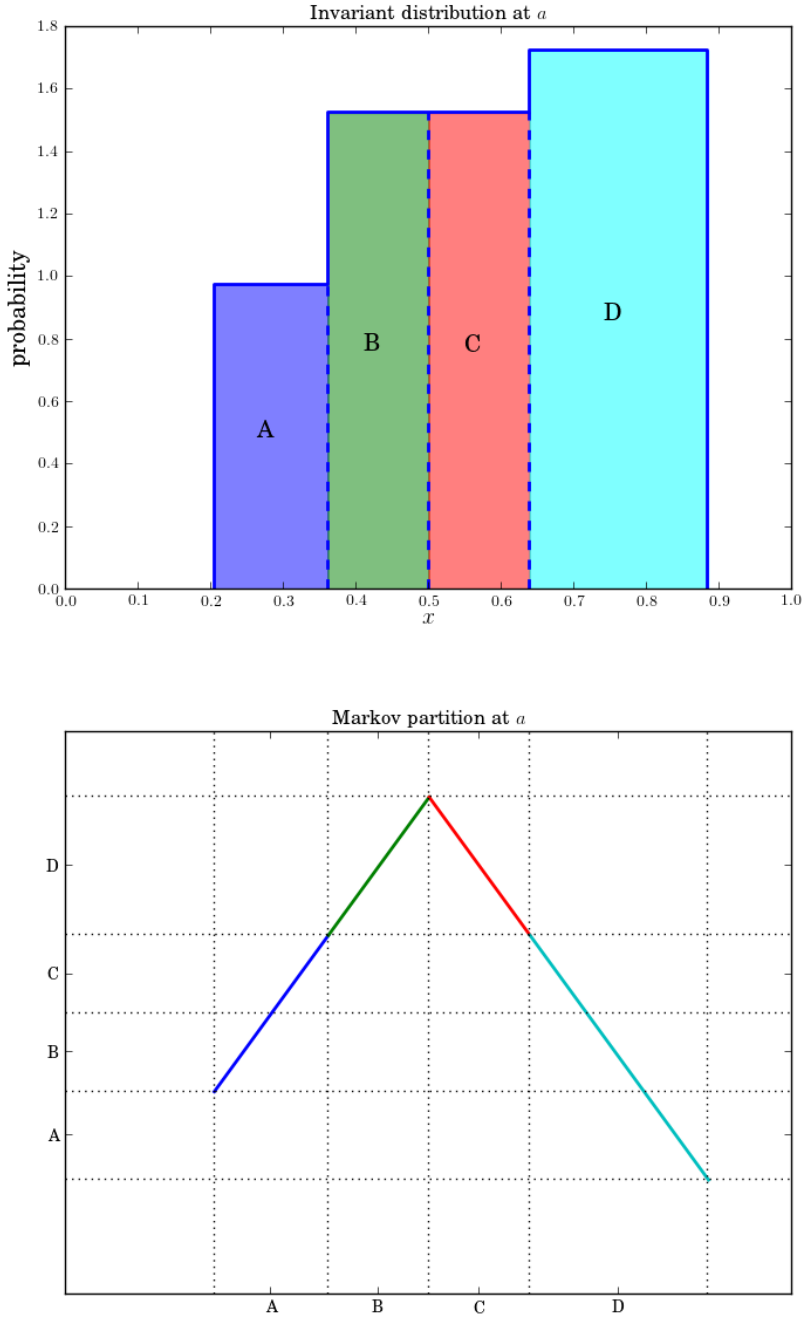


Figure 4.6: The tent map's invariant distribution at parameter a as defined in Eq. 4.10. It consists of three contiguous uniform parts, each colored differently for clarity. Those same colors are superimposed on the map itself along with guides to show that these uniform parts do indeed form a Markov partition.

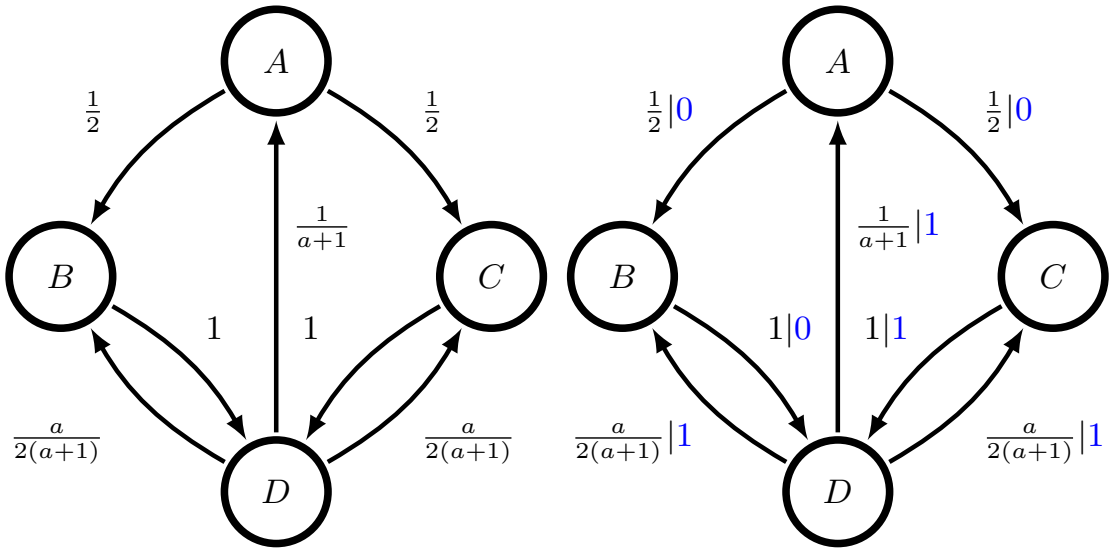


Figure 4.7: On the left is the Markov chain induced by the Markov partition. Once the standard generating partition has been applied the edges become labeled resulting in the hidden Markov model on the right.

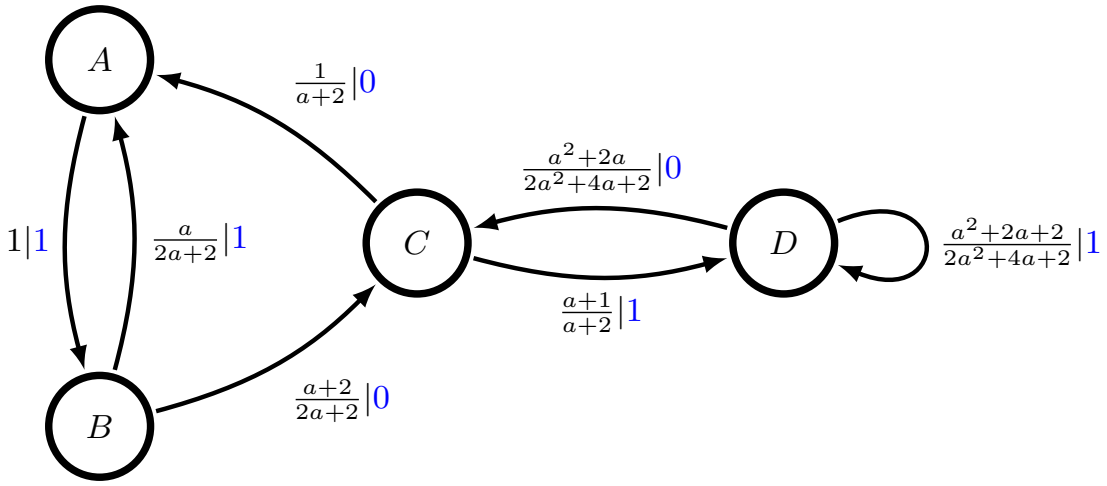


Figure 4.8: The ϵ-machine form of the hidden Markov model seen in Fig. 4.7. Note that for each state, there is at most a single outgoing edge with each symbol. This makes the states a function of $X_{:0}$, and allows for the calculation we need.

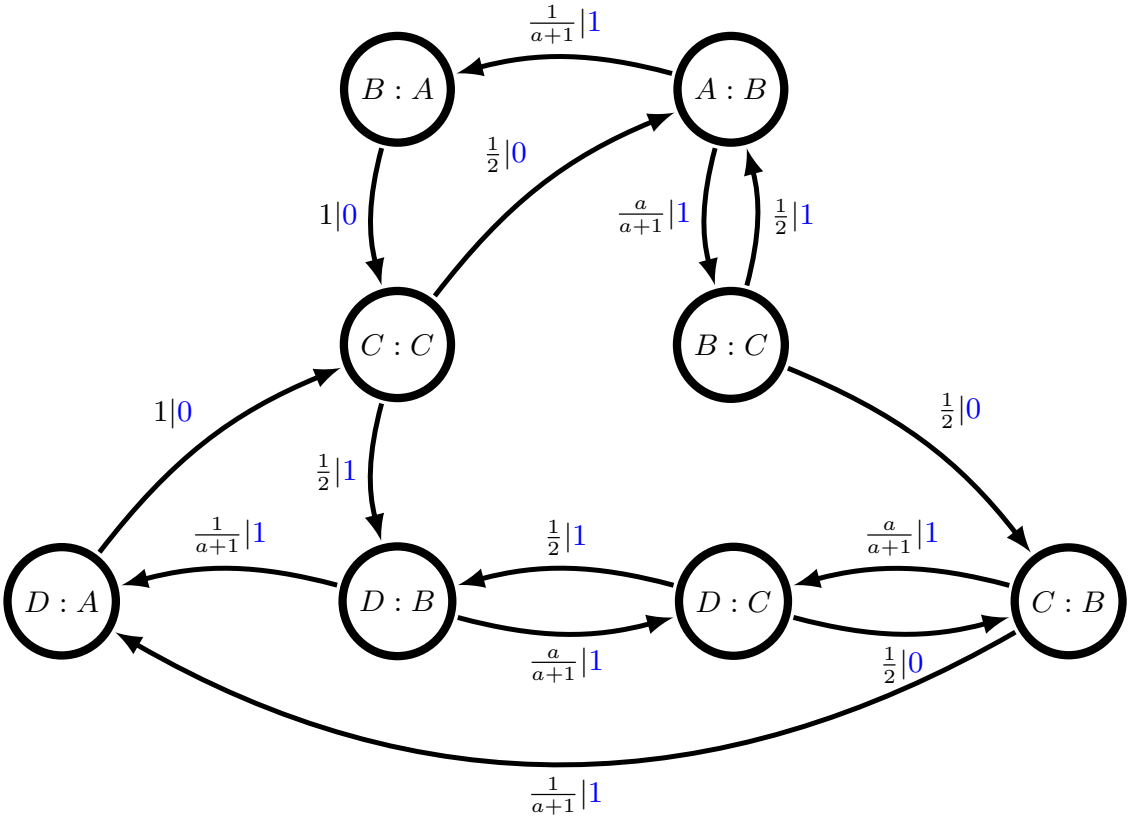


Figure 4.9: The bidirectional model of the process generated by the tent map at the parameter a. The states of this model are pairs $\mathcal{S}_0^+, \mathcal{S}_0^-$. This model will allow us to compute $\mathrm{H}[X_0|\mathcal{S}_0^+, \mathcal{S}_1^-]$ and thus r_μ and b_μ.

Chapter 5

Conclusions

Perhaps the single most important thing I have learned while studying complex systems is that the interpretation of measures is a very subtle and precarious thing. I have made a supreme effort throughout my work thus far to be reserved in the claims I make regarding the meaning and importance of the measures I work with. Unfortunately due to the somewhat sterile definitions, interpretation is necessary.

By way of example, consider $\mathrm{I}[X_{:0} : X_{0:}]$. It is reasonable to call this the information shared between the past and the future of a system. This, however, is not a particularly useful interpretation. We would like a more operational description. One common name for this quantity is *predictive information*. This title is fraught with peril. While it is true that this is the amount of information about the future that can be predicted from the past, this is *not*, however, the amount of information required to make this prediction.

Why this discrepancy? The simple, though non-intuitive answer, is that the mutual information between two random variables is not itself the entropy of a "shared" random variable. Therefore if one wishes to make the best predictions possible, you must find a random variable which contains $\mathrm{I}[X_{:0} : X_{0:}]$ as a portion of its entropy. The entropy of the smallest such random variable is known as the *generative complexity*[Loh09], and can be arbitrarily larger than $\mathrm{I}[X_{:0} : X_{0:}]$.

The story does not end there, however. The generative complexity is the smallest entropy of *any* random variable that captures $\mathrm{I}[X_{:0} : X_{0:}]$. Operationally though, one can not in general construct this random variable. What is needed is a random variable that is a function of the observables thus far, $X_{:0}$. The entropy of the smallest random variable that is a function of $X_{:0}$ is known as the *statistical complexity*. Again, this can be arbitrarily larger than not only $\mathrm{I}[X_{:0} : X_{0:}]$, but also than the generative complexity.

This narrative is unfortunately just a small sample of the subtleties that arise when using, interpreting, and defining complexity measures. It is my sincerest hope that I have not added to this collective uncertainty and frustration. The Markov and cryptic orders are, luckily, difficult to misinterpret. The ephemeral and bound information are unfortunately ripe for interpretational difficulties.

As my tenure as a graduate student comes to a close I find myself continuing to be drawn toward information theory with interpretational difficulties, namely the issue of negative information. This has been a relatively hot topic in the field of late, sparked by the work of Williams and Beer[Wil10]. This work, though groundbreaking, is considered lacking by many. Although a number of solutions have been proposed, none have been commonly accepted. This is one direction I wish to head in.

The other direction of primary interest to me is applications of r_μ and b_μ. These measures are of broad appeal and applicability, providing insight into nearly any system where the entropy rate is studied. Possibilities include the analysis of DNA, the study of pseudorandom number generators, analysis of written text, cryptography and cyphertexts, just to name a few. In time I suspect that researchers will be less interested in how much information a system generates (h_μ), but rather how much relevant, or bound, information it generates (b_μ).

Bibliography

Abd10 Abdallah, Samer A. and Mark D. Plumbley. "Predictive Information, Multi-Information, and Binding Information." *Technical Report C4DM-TR10-10, Centre for Digital Music, Queen Mary University of London*, 2010.

Abd12 Abdallah, Samer A. and Mark D. Plumbley. "A measure of statistical complexity based on predictive information with application to finite spin systems." *Physics Letters A*, *volume 376* (4), January 2012: pages 275–281.

Bal10 Ball, Robin C., Marina Diakonova, and Robert S. MacKay. "Quantifying Emergence in Terms of Persistent Mutual Information." *Advances in Complex Systems*, *volume 13* (3), 2010: pages 327–338.

Bel03 Bell, Anthony J. "The Co-Information Lattice." In N. Murata S. Amari, A. Cichocki , S. Makino (Editor) *Proceedings of the Fifth International Workshop on Independent Component Analysis and Blind Signal Separation*. Springer, New York, ica 2003 edition, 2003, pages 921–926.

Cha07 Chanda, Pritam, Aidong Zhang, Daniel Brazeau, Lara Sucheston, Jo L Freudenheim, Christine Ambrosone, and Murali Ramanathan. "Information-theoretic metrics for visualizing gene-environment interactions." *American journal of human genetics*, *volume 81* (5), November 2007: pages 939–63.

Cor09 Cormen, Thomas H., Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to algorithms*. The MIT Press, Cambridge, Massachusetts, 3rd edition, September 2009.

Cov06 Cover, Thomas M. and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, second edition, 2006.

Cru83 Crutchfield, James P. and N. H. Packard. "Symbolic Dynamics of Noisy Chaos." *Physica*, *volume 7D*, 1983: pages 201–223.

Cru03 Crutchfield, J. P. and D. P. Feldman. "Regularities Unseen, Randomness Observed: Levels of Entropy Convergence." *CHAOS*, *volume 13* (1), 2003: pages 25–54.

Cru09 Crutchfield, James P., Christopher J. Ellison, and John R. Mahoney. "Time's Barbed Arrow: Irreversibility, Crypticity, and Stored Information." *Physical Review Letters*, *volume 103* (9), 2009: page 094101.

Cru10 Crutchfield, James P., Christopher J. Ellison, Ryan G. James, and John R. Mahoney. "Synchronization and Control in Intrinsic and Designed Computation: An Information-Theoretic Analysis of Competing Models of Stochastic Computation." *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *volume 20* (3), July 2010: page 037105.

Dit10 Ditto, William L., A. Miliotis, K. Murali, Sudeshna Sinha, and Mark L. Spano. "Chaogates: Morphing logic gates that exploit dynamical patterns." *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *volume 20* (3), 2010: 037107.

Ell09 Ellison, Christopher J., John R. Mahoney, and James P. Crutchfield. "Prediction, Retrodiction, and The Amount of Information Stored in the Present." *Journal of Statistical Physics*, *volume 136* (6), 2009: pages 1005–1034.

Erb04 Erb, Ionas and Nihat Ay. "Multi-Information in the Thermodynamic Limit." *Journal of Statistical Physics*, *volume 115* (3/4), May 2004: pages 949–976.

Eri87 Eriksson, Karl-Erik and Kristian Lindgren. "Structural Information in Self-Organizing Systems." *Physica Scripta*, *volume 35* (3), March 1987: pages 388–397.

Fel98 Feldman, D. P. and J. P. Crutchfield. "Discovering Noncritical Organization: Statistical Mechanical, Information Theoretic, and Computational Views of Patterns in One-Dimensional Spin Systems." *J. Stat. Phys. (submitted)*, 1998.

Han80 Han, Te Sun. "Multiple mutual informations and multiple interactions in frequency data." *Information and Control*, *volume 46* (1), July 1980: pages 26–45.

Hop01 Hopcroft, John E., Rajeev Motwani, and Jeffrey D. Ullman. "Introduction to automata theory, languages, and computation, 2nd edition.", March 2001.

Jam11 James, Ryan G., Christopher J. Ellison, and James P. Crutchfield. "Anatomy of a bit: Information in a time series observation." *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *volume 21* (3), 2011: 037109.

Joh10 Johnson, B. D., James P. Crutchfield, Christopher J. Ellison, and Carl S. McTague. "Enumerating Finitary Processes." *CoRR*, *volume abs/1011.0036*, 2010.

Lin99 Lind, Douglas and Brian Marcus. *An introduction to Symbolic Dynamics and Coding*. Cambridge University Press, 1999.

Loh09 Lohr, Wolfgang and Nihat Ay. "On the Generative Nature of Prediction." *Advances in Complex Systems*, *volume 12* (02), 2009: pages 169–194.

Mah09 Mahoney, John R, Christopher J Ellison, and James P Crutchfield. "Information accessibility and cryptic processes." *Journal of Physics A: Mathematical and Theoretical*, *volume 42* (36), 2009: page 362002.

Pes77 Pesin, YB. "Characteristic Lyapunov exponents and smooth ergodic theory." *Russian Mathematical Surveys*, *volume 32* (4), August 1977: pages 55–114.

Sha01 Shalizi, Cosma Rohilla and James P. Crutchfield. "Computational Mechanics: Pattern and Prediction, Structure and Simplicity." *Journal of Statistical Physics*, *volume 104*, 2001: pages 817–879.

Tra11 Travers, Nicholas F. and James P. Crutchfield. "Exact Synchronization for Finite-State Sources." *Journal of Statistical Physics*, *volume 145* (5), 2011: pages 1181–1201.

Ver08 Verdu, Sergio and Tsachy Weissman. "The Information Lost in Erasures." *IEEE Transactions on Information Theory*, *volume 54* (11), November 2008: pages 5030–5058.

Wat60 Watanabe, Satosi. "Information Theoretical Analysis of Multivariate Correlation." *IBM Journal of Research and Development*, *volume 4* (1), January 1960: pages 66–82.

Wil10 Williams, Paul L. and Randall D. Beer. "Nonnegative Decomposition of Multivariate Information." *arXiv:1004.2515*, April 2010.

Yeu91 Yeung, Raymond W. "A New Outlook on Shannon's Information Measures." *IEEE Transactions on Information Theory*, *volume 37* (3), 1991: pages 466–474.