Dissertation Title

By

JOHN RIES MAHONEY, III
B.S. (University of California at Chico) 2001

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:

Professor James P. Crutchfield (Chair)

Professor Raissa D'Souza

Professor Gergely Zimanyi

Committee in Charge

2010

Lorem ipsum dolor sit amet, consectetuer adipiscing elit.

Abstract

words

words

and more words

Acknowledgments

Thanks for all the fish

Table of Contents

List of Figures

List of Tables

Chapter 1
Beginning

§1.1 Life of a theoretical physicist

We are born. We cry for food and the arms of our mother. We grow quickly, curious about our world. We go to graduate school and study physics to learn about the world. We formulate theories generalizing and supplanting the theories we have learned. We vie for funding to research our theories. We grow old and are either pleased or perturbed by our students' theories. We die.

Somehow, in all of this kaleidoscope, and apart from our physical needs to eat and sleep, to maintain our corporeal selves, we, as individuals and as a collective, learn.

What is it that we learn? How amazing that we find things worthy of the name 'quark'?

What if we could do all of this, but without the eating, sleeping and being born and dying? What if we didn't have theories to learn and generalize? What if life was an infinite experiment in the perfect knowing of one thing?

- Motivation - prediction. Causal prediction is not exact prediction.
- Motivation - 'structural understanding'. structure $! =$ lower entropy
- Process language
- Modeling -> ϵ-machine
- Venn diagrams.

  Basics, Markov fan, foliation rules
- Calc E via $+$ - intuitive
- Calc E via mixed states

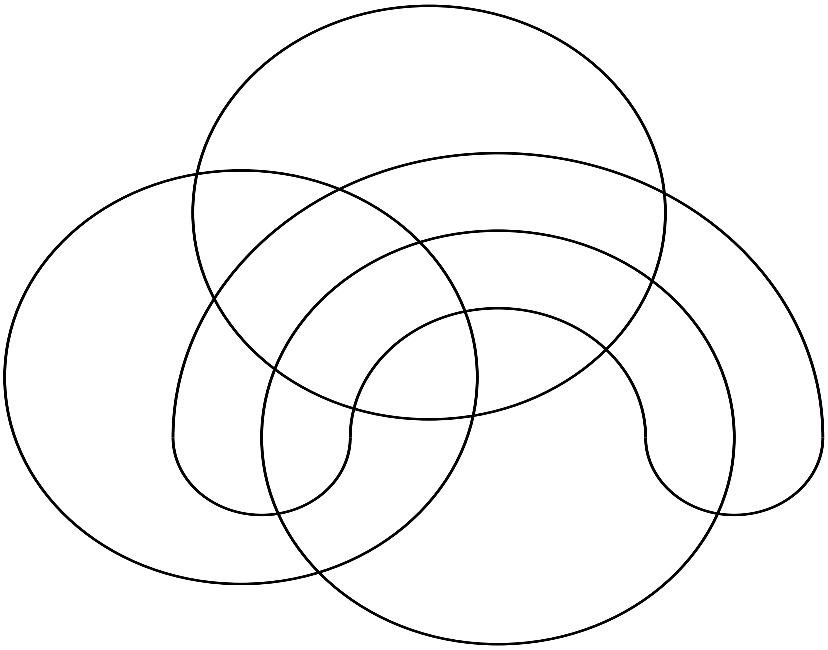

Figure 1.1: Example figure.

- Crypticity
- cryptic expansion / order
- Irreversibility
- Applications of all of this stuff.
- Appendix - Information theory.
- Appendix - Partitions and covers. Comment on topological refinement / shalizi / unifilar
- Appendix - Operational Computational Mechanics   What does it mean?

  what is C_μ, $\mathbf{E}$ (not really amnesia), operation of emachine compared to other types
- Jacquard

§1.2 Wer

werwer

§1.2.1 Subsection Name

wer

§1.2.2 Subsection Name

werwer

§1.3 Section Name

werwer

CHAPTER 2
PRATISP

§2.1 Introduction

"Predicting time series" encapsulates two notions of directionality. *Pre*diction—making a claim about the future based on the past—is directional. *Time* evokes images of rivers, clocks, and actions in progress. Curiously, though, when one writes a time series as a lattice of random variables, any necessary dependence on time's inherent direction is removed; at best it becomes convention. When we analyze a stochastic process to determine its correlation function, block entropy, entropy rate, and the like, we already have shed our commitment to the idea of *forward* by virtue of the fact that these quantities are defined independently of any perceived direction of the process.

Here we explore this ambivalence. In making it explicit, we consider not only predictive models, but also retrodictive models. We then demonstrate that it is possible to unify these two viewpoints and, in doing so, we discover several new properties of stationary stochastic dynamical systems. Along the way, we also rediscover, and recast, old ones.

We first review minimal causal representations of stochastic processes, as developed by *computational mechanics* [**?**, **?**]. We extend its (implied) forward-time representation to reverse-time. Then, we prove that the mutual information between a process's past and future—the *excess entropy*—is the mutual information between its forward- and reverse-time representations.

Excess entropy, and related mutual information quantities, are widely used diagnostics for complex systems. They have been applied to detect the presence of organization in dynamical systems [**?**, **?**, **?**, **?**], in spin systems [**?**, **?**, **?**], in neurobiological systems [**?**, **?**], and even in language, to mention only a few applications. For example, in natural language the excess entropy (**E**) diverges with the number of characters L as $\mathbf{E} \propto L^{1/2}$. The claim is that this reflects the long-range and strongly non-ergodic organization necessary for human communication [**?**, **?**].

The net result is a unified view of information processing in stochastic processes. For the first time, we give an explicit relationship between the internal (causal) state information—the statistical complexity [**?**]—and the observed information—the excess entropy. Another consequence is that the forward and reverse representations are two projections of a unified time-symmetric representation. From the latter it becomes clear there are important system properties that control how accessible internal state information is and how irreversible a process is. Moreover, the methods are sufficiently constructive that one can calculate the excess entropy in closed-form for finite-memory processes.

Before embarking, we clarify the present work's role in a collection of recent work. An announcement paper appeared in Ref. [**?**], and Ref. [**?**] will provide complementary results, on the measure-theoretic relationships between the above information quantities. A new classificiation scheme of stochastic processes appears in Ref. [**?**]. Here we lay out the theory in detail, giving step-by-step proofs of the main results and the calculational methods.

§2.2 Optimal Causal Models

The approach starts with a simple analogy. Any process, $\mathcal{P}$, is a joint probability distribution over the past and future observation symbols, $\Pr(\overleftarrow{X}, \overrightarrow{X})$. This distribution can be thought of as a *communication channel* with a specified input distribution $\Pr(\overleftarrow{X})$ [1]: It transmits information from the *past* $\overleftarrow{X} = \ldots X_{-3}X_{-2}X_{-1}$ to the *future* $\overrightarrow{X} = X_0X_1X_2\ldots$ by storing it in the present. X_t is the random variable for the measurement outcome at time t.

Our goal is also simply stated: We wish to predict the future using information from the past. At root, a prediction is probabilistic, specified by a distribution of possible futures $\overrightarrow{X}$ given a particular past $\overleftarrow{x}$: $\Pr(\overrightarrow{X}|\overleftarrow{x})$. At a minimum, a good predictor needs to capture *all* of the information I shared between the past and future: $\mathbf{E} = I[\overleftarrow{X};\overrightarrow{X}]$—the process's *excess entropy*. Note that there are several equivalent forms for $\mathbf{E}$, such as $\mathbf{E} = \lim_{L\to\infty} \left(H[X_0^L] - Lh_\mu \right)$ [**?**, and references therein].

[1] Throughout, we follow the notation and definitions of Refs. [**?**, **?**]. In addition, when we say $\overrightarrow{X}$, for example, this should be interpreted as a shorthand for using $\overrightarrow{X}^L$ and then taking an appropriate limit, such as $\lim_{L\to\infty}$ or $\lim_{L\to\infty} 1/L$.

Consider now the goal of modeling—building a representation that allows not only good prediction but also expresses the mechanisms producing a system's behavior. To build a model of a structured process (a memoryful channel), computational mechanics [**?**] introduced an equivalence relation $\overleftarrow{x} \sim \overleftarrow{x}'$ that groups all histories which give rise to the same prediction [2]:

$$\epsilon(\overleftarrow{x}) = \{\overleftarrow{x}' : \Pr(\overrightarrow{X}|\overleftarrow{x}) = \Pr(\overrightarrow{X}|\overleftarrow{x}')\}\ . \tag{2.1}$$

In other words, for the purpose of forecasting the future, two different pasts are equivalent if they result in the same prediction. The result of applying this equivalence gives the process's *causal states* $\boldsymbol{\mathcal{S}} = \Pr(\overleftarrow{X}, \overrightarrow{X})/\sim$, which partition the space $\overleftarrow{\mathbf{X}}$ of pasts into sets that are predictively equivalent. The set of causal states [3] can be discrete, fractal, or continuous; see, e.g., Figs. 7, 8, 10, and 17 in Ref. [**?**].

State-to-state transitions are denoted by matrices $T^{(x)}_{\mathcal{S}\mathcal{S}'}$ whose elements give the probability $\Pr(X = x, \mathcal{S}'|\mathcal{S})$ of transitioning from one state $\mathcal{S}$ to the next $\mathcal{S}'$ on seeing measurement x. The resulting model, consisting of the causal states and transitions, is called the process's ϵ*-machine*. Given a process $\mathcal{P}$, we denote its ϵ-machine by $M(\mathcal{P})$.

Causal states have a Markovian property that they render the past and future statistically independent; they *shield* the future from the past [**?**]:

$$\Pr(\overleftarrow{X}, \overrightarrow{X}|\mathcal{S}) = \Pr(\overleftarrow{X}|\mathcal{S}) \Pr(\overrightarrow{X}|\mathcal{S})\ . \tag{2.2}$$

Moreover, they are optimally predictive [**?**] in the sense that knowing which causal state a process is in is just as good as having the entire past: $\Pr(\overrightarrow{X}|\mathcal{S}) = \Pr(\overrightarrow{X}|\overleftarrow{X})$. In other words, causal shielding is equivalent to the fact [**?**] that the causal states capture all of the information shared between past and future: $I[\mathcal{S}; \overrightarrow{X}] = \mathbf{E}$.

ϵ-Machines have an important structural property called *unifilarity* [**?**, **?**]: From the start state, each symbol sequence corresponds to exactly one sequence of causal states [4]. ϵ-Machine unifiliarity underlies many of the results here. Its importance is reflected in the fact that representations without unifilarity, such as general hidden Markov models, *cannot* be used to directly calculate important system properties—including the most basic, such as, how random a process is. As a practical result, unifilarity is easy to verify: For each state, each measure-

[2] See Ref. [**?**] for a measure-theoretic discussion.

[3] A process's causal states consist of both transient and recurrent states. To simplify the presentation, we henceforth refer *only* to recurrent causal states that are discrete.

[4] Following terminology in computation theory this is referred to as *determinism* [**?**]. However, to reduce confusion, here we adopt the practice in information theory to call it the *unifilarity* of a process's representation [**?**].

ment symbol appears on at most one outgoing transition [5]. Thus, the signature of unifilarity is that on knowing the current state and measurement, the uncertainty in the next state vanishes: $H[\mathcal{S}_{t+1}|\mathcal{S}_t, X_t] = 0$. In summary, a process's ϵ-machine is its unique, minimal unifilar model.

§2.3 Information Processing Measures

Out of all optimally predictive models $\widehat{\boldsymbol{\mathcal{R}}}$—for which $I[\widehat{\mathcal{R}}; \overrightarrow{X}] = \mathbf{E}$—the ϵ-machine captures the minimal amount of information that a process must store in order to communicate all of the excess entropy from the past to the future. This is the Shannon information contained in the causal states—the *statistical complexity* [**?**]: $C_\mu \equiv H[\mathcal{S}] \leq H[\widehat{\mathcal{R}}]$. In short, $\mathbf{E}$ is the effective information transmission rate of the process, viewed as a channel, and C_μ is the sophistication of that channel.

Combined, these properties mean that the ϵ-machine is the basis against which modeling should be compared, since it captures all of a process's information at maximum representational efficiency.

Lastly, a key (and historically prior) dynamical systems invariant is the entropy rate:

$$h_\mu = \lim_{L\to\infty} \frac{H(L)}{L} , \tag{2.3}$$

where $H(L)$ is Shannon entropy of length-L sequences X^L. This is the per-measurement rate at which the process generates information—its degree of intrinsic randomness [**?**, **?**].

Importantly, due to unifilarity one can calculate the entropy rate directly from a process's ϵ-machine:

$$\begin{aligned} h_\mu &= H[X|\mathcal{S}] \\ &= -\sum_{\{\mathcal{S}\}} \Pr(\mathcal{S}) \sum_{\{x\}} T^{(x)}_{\mathcal{S}\mathcal{S}'} \log_2 T^{(x)}_{\mathcal{S}\mathcal{S}'} . \end{aligned} \tag{2.4}$$

$\Pr(\mathcal{S})$ is the asymptotic probability of the causal states, which is obtained as the normalized principal eigenvector of the transition matrix $T = \sum_{\{x\}} T^{(x)}$. We will use π to denote the distribution over the causal states as a row vector. Note that a process's statistical complexity can also be

[5] Specifically, each transition matrix $T^{(x)}$ has, at most, one nonzero component in each row.

directly calculated from its ϵ-machine:

$$\begin{aligned} C_\mu &= H[\mathcal{S}] \\ &= -\sum_{\{\mathcal{S}\}} \Pr(\mathcal{S}) \log_2 \Pr(\mathcal{S})\,. \end{aligned} \tag{2.5}$$

Thus, the ϵ-machine directly gives two important properties: a process's rate (h_μ) of producing information and the amount (C_μ) of historical information it stores in doing so.

§2.4 Excess Entropy

Until recently, **E** could not be as directly calculated as the entropy rate and the statistical complexity. This state of affairs was a major roadblock to analyzing the relationships between modeling and predicting and, more concretely, the relationships between (and even the interpretation of) a process's basic properties—h_μ, C_μ, and **E**. Ref. [**?**] announced the solution to this longstanding problem by deriving explicit expressions for **E** in terms of the ϵ-machine, providing a unified information-theoretic analysis of general processes. Here we provide a detailed account of the underlying methods and results.

To get started, we should recall what is already known about the relationships between these various quantities. First, some time ago, an explicit expression was developed from the Hamiltonian for one-dimensional spin chains with range-R interactions [**?**]:

$$\mathbf{E} = C_\mu - R\, h_\mu\,. \tag{2.6}$$

It was demonstrated that **E** is a generalized order parameter: Compared to structure factors, **E** is an assumption-free way to find structure and correlation in spin systems that does not require tuning [**?**].

Second, it has also been known for some time that the statistical complexity is an upper bound on the excess entropy [**?**]:

$$\mathbf{E} \leq C_\mu\,. \tag{2.7}$$

Nonetheless, other than the special, if useful, case of spin systems, until Ref. [**?**] there had been no direct way to calculate **E**. Remedying this limitation required broadening the notion of what a process is.

§2.5 Retrodiction

The original results of computational mechanics concern using the past to predict the future. But we can also retrodict: use the future to predict the past. That is, we scan the measurement variables not in the forward time direction, but in the reverse. The computational mechanics formalism is essentially unchanged, though its meaning and notation need to be augmented [**?**].

With this in mind, the previous mapping from pasts to causal states is now denoted ϵ^+ and it gave, what we will call, the *predictive* causal states $\boldsymbol{\mathcal{S}}^+$. When scanning in the reverse direction, we have a new relation, $\overrightarrow{x} \sim^- \overrightarrow{x}'$, which groups futures that are equivalent for the purpose of retrodicting the past: $\epsilon^-(\overrightarrow{x}) = \{\overrightarrow{x}' : \Pr(\overleftarrow{X} | \overrightarrow{x}) = \Pr(\overleftarrow{X} | \overrightarrow{x}')\}$. It gives the *retrodictive* causal states $\boldsymbol{\mathcal{S}}^- = \Pr(\overleftarrow{X}, \overrightarrow{X}) / \sim^-$. And, not surprisingly, we must also distinguish the forward-scan ϵ-machine M^+ from the reverse-scan ϵ-machine M^-. They assign corresponding entropy rates, h_μ^+ and h_μ^-, and statistical complexities, $C_\mu^+ = H[\mathcal{S}^+]$ and $C_\mu^- = H[\mathcal{S}^-]$, respectively, to the process.

To orient ourselves, a graphical aid, the *hidden process lattice*, is helpful at this point; see Table 2.1.

						Past	Present	Future						
						$\overleftarrow{X}$		$\overrightarrow{X}$						
$\ldots$		X_{-3}		X_{-2}		X_{-1}		X_0		X_1		X_2		$\ldots$
$\ldots$	$\mathcal{S}^+_{-3}$		$\mathcal{S}^+_{-2}$		$\mathcal{S}^+_{-1}$		$\mathcal{S}^+_0$		$\mathcal{S}^+_1$		$\mathcal{S}^+_2$		$\mathcal{S}^+_3$	$\ldots$
$\ldots$	$\mathcal{S}^-_{-3}$		$\mathcal{S}^-_{-2}$		$\mathcal{S}^-_{-1}$		$\mathcal{S}^-_0$		$\mathcal{S}^-_1$		$\mathcal{S}^-_2$		$\mathcal{S}^-_3$	$\ldots$

Table 2.1: Hidden Process Lattice: The X variables denote the observed process; the $\mathcal{S}$ variables, the hidden states. If one scans the observed variables in the positive direction—seeing X_{-3}, X_{-2}, and X_{-1}—then that history takes one to causal state $\mathcal{S}^+_0$. Analogously, if one scans in the reverse direction, then the succession of variables X_2, X_1, and X_0 leads to $\mathcal{S}^-_0$.

Now we are in a position to ask some questions. Perhaps the most obvious is, In which time direction is a process most predictable? The answer is that a process is equally predictable in either:

Proposition 1. *[**?**] For a stationary process, optimally predicting the future and optimally retrodicting the past are equally effective:* $h_\mu^- = h_\mu^+$.

Proof. *A stationary stochastic process satisfies:*

$$H[X_{-L+2}, \ldots, X_0] = H[X_{-L+1}, \ldots, X_{-1}] \, . \tag{2.8}$$

Keeping this in mind, we directly calculate:

$$\begin{aligned}
h_\mu^+ &= H[X_0 | \overleftarrow{X}] \\
&= \lim_{L \to \infty} H[X_0 | X_{-L+1}, \ldots, X_{-1}] \\
&= \lim_{L \to \infty} \big(H[X_{-L+1}, \ldots, X_0] - H[X_{-L+1}, \ldots, X_{-1}] \big) \\
&= \lim_{L \to \infty} \big(H[X_{-L+1}, \ldots, X_0] - H[X_{-L+2}, \ldots, X_0] \big) \\
&= \lim_{L \to \infty} \big(H[X_{-1}, \ldots, X_{L-2}] - H[X_0, \ldots, X_{L-2}] \big) \\
&= \lim_{L \to \infty} H[X_{-1} | X_0, \ldots, X_{L-2}] \\
&= H[X_{-1} | \overrightarrow{X}] \\
&= h_\mu^- \, . \quad \square
\end{aligned}$$

Somewhat surprisingly, the effort involved in optimally predicting and retrodicting is not necessarily the same:

Proposition 2. *[**?**] There exist stationary processes for which* $C_\mu^- \neq C_\mu^+$.

Proof. *The Random Insertion Process, analyzed in a later section, establishes this by example.*

Note that **E** is mute on this score. Since the mutual information I is symmetric in its variables [**?**], **E** is time symmetric. Proposition 2 puts us on notice that **E** necessarily misses many of a process's structural properties.

§2.6 Excess Entropy from Causal States

The relationship between predicting and retrodicting a process, and ultimately **E**'s role, requires teasing out how the states of the forward and reverse ϵ-machines capture information from the past and the future. To do this we analyzed [**?**] a four-variable mutual information: $I[\overleftarrow{X}; \overrightarrow{X}; \mathcal{S}^+; \mathcal{S}^-]$. A large number of expansions of this quantity are possible. A systematic development follows from Ref. [**?**] which showed that Shannon entropy $H[\cdot]$ and mutual information $I[\cdot;\cdot]$ form a signed measure over the space of events. Practically, there is a direct correspondence between set theory and these information measures. Using this, Ref. [**?**] developed an *ϵ-machine information diagram* over four variables, which gives a minimal set of entropies, conditional entropies, mutual informations, and conditional mutual informations necessary to analyze the relationships among h_μ, C_μ, and $\mathbf{E}$ for general stochastic processes.

In a generic four-variable information diagram, there are 15 independent variables. Fortunately, this greatly simplifies in the case of using an ϵ-machine to represent a process; there are only 5 independent variables in the ϵ-machine information diagram [**?**]. (These results are announced in [**?**]; see Fig. 1 there.)

Simplified in this way, we are left with our main results which, due to the preceding effort, are particularly transparent.

Theorem 1. *Excess entropy is the mutual information between the predictive and retrodictive causal states:*

$$\mathbf{E} = I[\mathcal{S}^+;\mathcal{S}^-] \,. \tag{2.9}$$

Proof. *This follows due to the redundancy of pasts and predictive causal states, on the one hand, and of futures and retrodictive causal states, on the other. These redundancies, in turn, are expressed via $\mathcal{S}^+ = \epsilon^+(\overleftarrow{X})$ and $\mathcal{S}^- = \epsilon^-(\overrightarrow{X})$, respectively. That is, we have*

$$\begin{aligned} I[\overleftarrow{X};\overrightarrow{X};\mathcal{S}^+;\mathcal{S}^-] &= I[\overleftarrow{X};\overrightarrow{X}] \\ &= \mathbf{E} \,, \end{aligned} \tag{2.10}$$

on the one hand, and

$$I[\overleftarrow{X};\overrightarrow{X};\mathcal{S}^+;\mathcal{S}^-] = I[\mathcal{S}^+;\mathcal{S}^-] \,, \tag{2.11}$$

on the other. □

That is, the process's channel utilization $\mathbf{E} = I[\overleftarrow{X};\overrightarrow{X}]$ is the same as that of a "channel" between the forward and reverse ϵ-machine states.

Proposition 3. *The predictive and retrodictive statistical complexities are:*

$$C_\mu^+ = \mathbf{E} + H[\mathcal{S}^+|\mathcal{S}^-] \text{ and} \tag{2.12}$$

$$C_\mu^- = \mathbf{E} + H[\mathcal{S}^-|\mathcal{S}^+] \, . \tag{2.13}$$

Proof. $\mathbf{E} = I[\mathcal{S}^+;\mathcal{S}^-] = H[\mathcal{S}^+] - H[\mathcal{S}^+|\mathcal{S}^-]$. *Since the first term is C_μ^+, we have the predictive statistical complexity. Similarly for the retrodictive complexity.* □

Corollary 1. $C_\mu^+ \geq H[\mathcal{S}^+|\mathcal{S}^-]$ *and* $C_\mu^- \geq H[\mathcal{S}^-|\mathcal{S}^+]$.

Proof. $\mathbf{E} \geq 0$.

The Theorem and its companion Proposition give an explicit connection between a process's excess entropy and its causal structure—its ϵ-machines. More generally, the relationships directly tie mutual information measures of observed sequences to a process's internal structure. This is our main result. It allows us to probe the properties that control how closely observed statistics reflect a process's hidden organization. However, this requires that we understand how M^+ and M^- are related. We express this relationship with a unifying model—the bidirectional machine.

§2.7 The Bidirectional Machine

At this point, we have two separate ϵ-machines—one for predicting (M^+) and one for retrodicting (M^-). We will now show that one can do better, by simultaneously utilizing causal information from the past and future.

Definition. *Let $M^\pm$ denote the* bidirectional machine *given by the equivalence relation $\sim^\pm$* [6]*:*

$$\begin{aligned} \epsilon^\pm(\overleftrightarrow{x}) &= \epsilon^\pm(\overleftarrow{x}, \overrightarrow{x}) \\ &= \{(\overleftarrow{x}', \overrightarrow{x}') : \overleftarrow{x}' \in \epsilon^+(\overleftarrow{x}) \text{ and } \overrightarrow{x}' \in \epsilon^-(\overrightarrow{x})\} \end{aligned}$$

with causal states $\boldsymbol{\mathcal{S}}^\pm = \Pr(\overleftrightarrow{X})/\!\sim^\pm$.

That is, the bidirectional causal states are a partition of $\overleftrightarrow{X}$: $\boldsymbol{\mathcal{S}}^\pm \subseteq \boldsymbol{\mathcal{S}}^+ \times \boldsymbol{\mathcal{S}}^-$. This follows from a straightforward adaptation of the analogous result for forward ϵ-machines [**?**].

To illustrate, imagine being given a particular realization $\overleftrightarrow{x}$. In effect, the bidirectional machine $M^\pm$ describes how one can move around on the hidden process lattice of Table 2.1:

[6] Interpret the symbol ± as "plus *and* minus".

1. When scanning in the forward direction, states and transitions associated with M^+ are followed.

2. When scanning in the reverse direction, states and transitions associated with M^- are followed.

3. At any time, one can change to the opposite scan direction, moving to the state of the opposite scan's ϵ-machine. For example, if one moves forward following M^+ and ends in state $\mathcal{S}^+$, having seen $\overleftarrow{x}$ and about to see $\overrightarrow{x}$, then one moves to $\mathcal{S}^- = \epsilon^-(\overrightarrow{x})$.

At time t, the bidirectional causal state is $\mathcal{S}^{\pm}_t = (\epsilon^+(\overleftarrow{x}_t), \epsilon^-(\overrightarrow{x}_t))$. When scanning in the forward direction, the first symbol of $\overrightarrow{x}_t$ is removed and appended to $\overleftarrow{x}_t$. When scanning in the reverse direction, the last symbol in $\overleftarrow{x}_t$ is removed and prefixed to $\overrightarrow{x}_t$. In either situation, the new bidirectional causal state is determined by $\epsilon^{\pm}$ and the updated past and future.

This illustrates the relationship between $\mathcal{S}^+$ and $\mathcal{S}^-$, as specified by $M^{\pm}$, when given a particular realization. Generally, though, one considers an ensemble $\overleftrightarrow{X}$ of realizations. In this case, the bidirectional state transitions are probabilistic and possibly nonunifilar. This relationship can be made more explicit through the use of maps between the forward and reverse causal states. These are the *switching* maps.

The forward map is a linear function from the simplex over $\boldsymbol{\mathcal{S}}^-$ to the simplex over $\boldsymbol{\mathcal{S}}^+$, and analogously for the reverse map. The maps are defined in terms of conditional probability distributions:

1. The *forward map* $f : \Delta^n \to \Delta^m$, where $f(\sigma^-) = \Pr(\mathcal{S}^+|\sigma^-)$; and

2. The *reverse map* $r : \Delta^m \to \Delta^n$, where $r(\sigma^+) = \Pr(\mathcal{S}^-|\sigma^+)$,

where $n = |\boldsymbol{\mathcal{S}}^-|$ and $m = |\boldsymbol{\mathcal{S}}^+|$.

We will sometimes refer to these maps in the Boolean rather than probabilistic sense. The case will be clear from context.

Proposition 4. *r and f are onto.*

Proof. *Consider the reverse map r that takes one from a forward causal state to a reverse causal state. Assume r is not onto. Then there must be a reverse state σ^- that is not in the range of*

$r(\mathcal{S}^+)$. This means that no forward causal state is paired with σ^- and so there is no past $\overleftarrow{x}$ with a possible future $\overrightarrow{x} \in \sigma^-$. That is, $\epsilon^\pm(\overleftarrow{x}, \overrightarrow{x}) = \emptyset$ and, specifically, $\epsilon^-(\overrightarrow{x}) = \emptyset$. Thus, σ^- does not exist.

A similar argument shows that f is onto. □

Definition. *The amount of stored information needed to optimally predict and retrodict a process is $M^\pm$'s statistical complexity:*

$$C_\mu^\pm \equiv H[\mathcal{S}^\pm] = H[\mathcal{S}^+, \mathcal{S}^-] \, . \qquad (2.14)$$

From the immediately preceding results we obtain the following simple, explicit, and useful relationship:

Corollary 2. $\mathbf{E} = C_\mu^+ + C_\mu^- - C_\mu^\pm$.

Thus, we are led to a wholly new interpretation of the excess entropy—in addition to the original three discussed in Ref. [**?**]: **E** is exactly the difference between these structural complexities. Moreover, only when $\mathbf{E} = 0$ does $C_\mu^\pm = C_\mu^+ + C_\mu^-$.

More to the point, thinking of the C_μs as proportional to the size of the corresponding machine, we establish the representational efficiency of the bidirectional machine:

Proposition 5. $C_\mu^\pm \leq C_\mu^+ + C_\mu^-$.

Proof. *This follows directly from the preceding corollary and the non-negativity of mutual information.* □

We can say a bit more, with the following bounds.

Corollary 3. $C_\mu^+ \leq C_\mu^\pm$ *and* $C_\mu^- \leq C_\mu^\pm$.

These results say that taking into account causal information from the past *and* the future is more efficient (i) than ignoring one or the other and (ii) than ignoring their relationship.

§2.7.1 Upper Bounds

Here we give new, tighter bounds for **E** than Eq. (2.7) and greatly simplified proofs than those provided in Refs. [**?**] and [**?**].

Proposition 6. *For a stationary process,* $\mathbf{E} \leq C_\mu^+$ *and* $\mathbf{E} \leq C_\mu^-$.

Proof. *These bounds follow directly from applying basic information inequalities: $I[X,Y] \leq H[X]$ and $I[X,Y] \leq H[Y]$. Thus, $\mathbf{E} = I[\mathcal{S}^-;\mathcal{S}^+] \leq H[\mathcal{S}^-]$, which is C_μ^-. Similarly, since $I[\mathcal{S}^-;\mathcal{S}^+] \leq H[\mathcal{S}^+]$, we have $\mathbf{E} \leq C_\mu^+$.* □

§2.7.2 Causal Irreversibility

We have shown that predicting and retrodicting may require different amounts of information storage ($C_\mu^+ \neq C_\mu^-$). We now examine this asymmetry.

Given a word $w = x_0 x_2 \ldots x_{L-1}$, the word we see when scanning in the reverse direction is $\widetilde{w} = x_{L-1} \ldots x_1 x_0$, where x_{L-1} is encountered first and x_0 is encountered last.

Definition. *A* microscopically reversible process *is one for which* $\Pr(w) = \Pr(\widetilde{w})$*, for all words $w = x^L$ and all L.*

Microscopic reversibility simply means that flipping $t \to -t$ leads to the same process. A microscopically reversible process yields the same word distribution when scanned in either direction; we will denote this $\mathcal{P}^+ = \mathcal{P}^-$.

Proposition 7. *A microscopically reversible process has $M^+ = M^-$.*

Proof. *If $\mathcal{P}^+ = \mathcal{P}^-$, then $M(\mathcal{P}^+) = M(\mathcal{P}^-)$ since M is a function. These are M^+ and M^-, respectively.* □

Now consider a slightly looser, and more helpful, notion of reversibility, expressed quantitatively as a measure of irreversibility.

Definition. *A process's* causal irreversibility *[**?**] is:*

$$\Xi(\mathcal{P}) = C_\mu^+ - C_\mu^- \ . \tag{2.15}$$

Corollary 4. $\Xi(\mathcal{P}) = H[\mathcal{S}^+|\mathcal{S}^-] - H[\mathcal{S}^-|\mathcal{S}^+]$.

Definition. *A* causally reversible *process is one with vanishing causal irreversibility,* $\Xi(\mathcal{P}) = 0$.

Proposition 8. *If a process is microscopically reversible, then the process is causally reversible.*

Proof. *By Prop. 7, a microscopically reversible process has $M^+ = M^-$ and in particular, $\boldsymbol{\mathcal{S}}^+ = \boldsymbol{\mathcal{S}}^-$ and their transition matrices are the same. This means that* $\Pr(\mathcal{S}^+) = \Pr(\mathcal{S}^-)$*. Thus, $C_\mu^+ = C_\mu^-$ and $\Xi = 0$.* □

Thus, the class of causally reversible processes is potentially larger then the class of microscopically reversible processes. That is, there can exist processes with vanishing causal irreversibility ($\Xi = 0$) that are *not* microscopically reversible. For example, the periodic process $\ldots 123123123 \ldots$ is not microscopically reversible, since $\Pr(123) \neq \Pr(321)$. However, as $C_\mu^- = C_\mu^+ = \log_2 3$, this process is causally reversible.

In fact, the class of causally reversible processes includes any process whose left- and right-scan processes are isomorphic under a simultaneous alphabet and state isomorphism. Given that the spirit of symbolic dynamics is to consider processes only up to isomorphism, this measure seems to capture a very natural notion of reversibility. Interestingly, it appears, based on several case studies, that causal reversibility captures *exactly* that notion. That is, it would seem there are no causally reversible processes for which $\mathcal{P}^+ \nsim \mathcal{P}^-$. We leave this as a conjecture.

Finally, note that causal irreversibility is not controlled by **E**, since, as noted above, the latter is scan-symmetric.

§2.7.3 Process Crypticity

Lurking in the preceding development and results is an alternative view of how forecasting and modeling building are related.

We can extend our use of Shannon's communication theory (processes are memoryful channels) to view the activity of an observer building a model of a process as the attempt to decrypt from a measurement sequence the hidden state information [**?**]. The parallel we draw is that the design goal of cryptography is to not reveal internal correlations and structure within an encrypted data stream, even though in fact there is a message—hidden organization and structure—that will be revealed to a recipient with the correct codebook. This is essentially the circumstance a scientist faces when building a model, for the first time, from measurements: What are the states and dynamic (hidden message) in the observed data?

Here, we address only the case of *self-decoding* in which the information used to build a model is only that available in the observed process $\Pr(\overleftrightarrow{X})$. That is, no "side-band" communication, prior knowledge, or disciplinary assumptions are allowed. Note, though, that modeling with such additional knowledge requires solving the self-decoding case, addressed here, first.

The self-decoding approach to building nonlinear models from time series was introduced in Ref. [**?**].

The relationship between excess entropy and statistical complexity established by Thm. 1 indicates that there are fundamental limitations on the amount of a process's stored information directly present in observations, as reflected in the mutual information measure **E**. We now introduce a measure of this accessibility.

Definition. *A process's* crypticity *is:*

$$\chi^{\pm}(M^{+}, M^{-}) = H[\mathcal{S}^{+}|\mathcal{S}^{-}] + H[\mathcal{S}^{-}|\mathcal{S}^{+}] \, . \tag{2.16}$$

Proposition 9. $\chi^{\pm}(M^{+}, M^{-})$ *is a distance between a process's forward and reverse* ϵ*-machines.*

Proof. $\chi^{\pm}(\cdot, \cdot)$ *is non-negative, symmetric, and satisfies a triangle inequality. This follows from the solution of exercise 2.9 of Ref. [**?**]. See also, Ref. [**?**].* □

Theorem 2. $M^{\pm}$*'s statistical complexity is:*

$$C_{\mu}^{\pm} = \mathbf{E} + \chi^{\pm} \, . \tag{2.17}$$

Proof. *This follows directly from the corollary and the predictive and retrodictive statistical complexity relations, Eq. (2.12) and (2.13).* □

Referring to $\chi^{\pm}$ as crypticity comes directly from this result: It is the amount of internal state information ($C_{\mu}^{\pm}$) not locally present in the observed sequence (**E**). That is, a process hides $\chi^{\pm}$ bits of information.

Note that if crypticity is low $\chi^{\pm} \approx 0$, then much of the stored information is present in observed behavior: $\mathbf{E} \approx C_{\mu}^{\pm}$. However, when a process's crypticity is high, $\chi^{\pm} \approx C_{\mu}^{\pm}$, then little of its structural information is directly present in observations. The measurements appear very close to being independent, identically distributed ($\mathbf{E} \approx 0$) despite the fact that the process can be highly structured ($C_{\mu}^{\pm} \gg 0$).

Corollary 5. $M^{\pm}$*'s statistical complexity bounds the process's crypticity:*

$$C_{\mu}^{\pm} \geq \chi^{\pm} \, . \tag{2.18}$$

Proof. $\mathbf{E} \geq 0$. □

Thus, a truly cryptic process has $C_\mu^\pm = \chi^\pm$ or, equivalently, $\mathbf{E} = 0$. In this circumstance, little or nothing can be learned about the process's hidden organization from measurements. This would be perfect encryption.

We will find it useful to discuss the two contributions to $\chi^\pm$ separately. Denote these $\chi^+ = H[\mathcal{S}^+|\mathcal{S}^-]$ and $\chi^- = H[\mathcal{S}^-|\mathcal{S}^+]$.

The preceding results can be compactly summarized in an information diagram that uses the ϵ-machine representation of a process; see Ref. [**?**] and Ref. [**?**]. They also suggest a classification scheme based on crypticty, to complement the Markov-order classification; see Ref. [**?**]. In the following, we phrase the calculation in terms of $\mathbf{E}$, and χ^+, χ^-, $\chi^\pm$, $C_\mu^\pm$, and Ξ follow straightforwardly.

§2.8 Alternative Presentations

The ϵ-machine is a process's unique, minimal unifilar presentation. Now we introduce two alternative presentations, which need not be ϵ-machines, that will be used in the calculation of $\mathbf{E}$. Since the states of these alternative presentations are not causal states, we will use $\mathcal{R}_t$, rather than $\mathcal{S}_t$, to denote the random variable for their state at time t.

§2.8.1 Time-Reversed Presentation

Any machine M transitions from the current state $\mathcal{R}$ to the next state $\mathcal{R}'$ on the current symbol x:

$$T^{(x)}_{\mathcal{R}\mathcal{R}'} \equiv \Pr(X = x, \mathcal{R}'|\mathcal{R})\,. \tag{2.19}$$

Note that $T = \sum_{\{x\}} T^{(x)}$ is a stochastic matrix with principal eigenvalue 1 and left eigenvector π, which gives $\Pr(\mathcal{R})$. Recall that the Perron-Frobenius theorem applied to stochastic matrices guarantees the uniqueness of π.

Using standard probability rules to interchange $\mathcal{R}$ and $\mathcal{R}'$, we can construct a new set of transition matrices which defines a presentation of the process that generates the symbols in reverse order. It is useful to consider a time-reversing operator acting on a machine. Denoting it $\mathcal{T}$, $\widetilde{M} = \mathcal{T}(M)$ is the *time-reversed presentation* of M. It has symbol-labeled transition matrices:

$$\begin{aligned} \widetilde{T}^{(x)}_{\mathcal{R}'\mathcal{R}} &\equiv \Pr(X = x, \mathcal{R}|\mathcal{R}') \\ &= T^{(x)}_{\mathcal{R}\mathcal{R}'} \frac{\Pr(\mathcal{R})}{\Pr(\mathcal{R}')}\,. \end{aligned} \tag{2.20}$$

and stochastic matrix $\widetilde{T} = \sum_{\{x\}} \widetilde{T}^{(x)}$.

Proposition 10. *The stationary distribution $\widetilde{\pi}$ over the time-reversed presentation states is the same as the stationary distribution π of M.*

Proof. *We assume $\widetilde{\pi} = \pi$, the left eigenvector of T, and verify the assumption, recalling the uniqueness of π. We have:*

$$\begin{aligned}
\widetilde{\pi}_\rho &= \sum_{\rho'} \widetilde{\pi}_{\rho'} \widetilde{T}_{\rho'\rho} \\
&= \sum_{\rho'} \widetilde{\pi}_{\rho'} T_{\rho\rho'} \frac{\pi_\rho}{\pi_{\rho'}} \\
&= \sum_{\rho'} T_{\rho\rho'} \pi_\rho \\
&= \pi_\rho \ . \quad \square
\end{aligned}$$

In the second to last line, we recall the assumption $\widetilde{\pi}_{\rho'} = \pi_{\rho'}$. And in the final, we note that T is stochastic. □

Finally, when we consider the product of transition matrices over a given sequence w, it is useful to simplify notation as follows:

$$T^{(w)} \equiv T^{(x_0)} T^{(x_1)} \cdots T^{(x_{L-1})}.$$

§2.8.2 Mixed-State Presentation

The states of machine M can be treated as a standard basis in a vector space. Then, any distribution over these states is a linear combination of those basis vectors. Following Ref. [**?**], these distributions are called *mixed states*.

Now we focus on a special subset of mixed states and define $\mu(w)$ as the distribution over the states of M that is induced after observing w:

$$\begin{aligned}
\mu(w) &\equiv \Pr(\mathcal{R}_L | X_0^L = w) && (2.21) \\
&= \frac{\Pr(X_0^L = w, \mathcal{R}_L)}{\Pr(X_0^L = w)} && (2.22) \\
&= \frac{\pi T^{(w)}}{\pi T^{(w)} \mathbf{1}}, && (2.23)
\end{aligned}$$

where X_0^L is shorthand for an undetermined sequence of L measurements beginning at time $t = 0$ and $\mathbf{1}$ is a column vector of 1s. In the last line, we write the probabilities in terms of the stationary distribution and the transition matrices of M. This expansion is valid for any machine that generates the process in the forward-scan (left-to-right) direction.

If we consider the entire set of such mixed states, then we can construct a presentation of the process by specifying the transition matrices:

$$\Pr(x, \mu(wx)|\mu(w)) \equiv \frac{\Pr(wx)}{\Pr(w)} \tag{2.24}$$

$$= \mu(w)T^{(x)}\mathbf{1} \,. \tag{2.25}$$

Note that many words can induce the same mixed state. As with the time-reversed presentation, it will be useful to define a corresponding operator $\mathcal{U}$ that acts on a machine M, returning its *mixed-state presentation* $\mathcal{U}(M)$.

§2.9 Calculating Excess Entropy

We are now ready to describe how to calculate the excess entropy, using the time-symmetric perspective. Generally, our goal is to obtain a conditional distribution $\Pr(\mathcal{S}^+|\mathcal{S}^-)$ which, when combined with the ϵ-machines, yields a direct calculation of **E** via Thm. 1. This is a two-step procedure which begins with M^+, calculates $\widetilde{M}^+$, and ends with M^-. One could also start with M^- to obtain M^+. These possibilities are captured in the diagram:

$$\begin{tikzcd} M^+ \arrow[d, "\mathcal{T}"'] & \widetilde{M}^- \arrow[l, "\mathcal{U}"'] \\ \widetilde{M}^+ \arrow[r, "\mathcal{U}"'] & M^- \arrow[u, "\mathcal{T}"'] \end{tikzcd} \tag{2.26}$$

In detail, we begin with M^+ and reverse the direction of time by constructing the time-reversed presentation $\widetilde{M}^+ = \mathcal{T}(M^+)$. Then, we construct the mixed-state presentation $\mathcal{U}(\widetilde{M}^+)$ of the time-reversed presentation to obtain M^-.

Note that $\mathcal{T}$ acting on M^+ does not generically yield another ϵ-machine. (This was not the purpose of $\mathcal{T}$.) However, the states will still be useful when we construct the mixed-state presentation of $\widetilde{M}^+$. This is because the states, which serve as basis states in the mixed-state presentation, are in a one-to-one correspondence with the forward causal states of M^+. This correspondence was established by Prop. 10.

Also, note that $\mathcal{U}$ is not guaranteed to construct a minimal presentation of the process. However, this does not appear to be an issue when working with time-reversed presentations of an ϵ-machine. We leave it as a conjecture that $\mathcal{U}(\mathcal{T}(M))$ is always minimal. Even so, the Appendix demonstrates that an appropriate sum can be carried out which always yields the desired conditional distribution.

Returning to the two-step procedure, one must construct the mixed-state presentation of $\widetilde{M}^+$. It is helpful to keep the hidden process lattice of Table 2.1 in mind. Since $\widetilde{M}^+$ generates the process from right-to-left, it encounters symbols of w in reverse order. The consequence of this is that the form of the mixed state changes slightly. However, it *still* represents the distribution over the current state induced by seeing w. We denote this new form by $\nu(w)$:

$$\nu(w) \equiv \Pr(\mathcal{R}_0 | X_0^L = w) \tag{2.27}$$

$$= \frac{\Pr(\mathcal{R}_0, X_0^L = w)}{\Pr(X_0^L = w)} \tag{2.28}$$

$$= \frac{\pi T^{(\widetilde{w})}}{\pi T^{(\widetilde{w})} \mathbf{1}} \,, \tag{2.29}$$

where π and T are the stationary distribution and transition matrices of a machine that generates the process from right-to-left, respectively. In this procedure, we are making use of $\widetilde{M}^+$ and thus, $\widetilde{\pi}$ and $\widetilde{T}$.

Similarly, if we consider the entire set of such mixed states, we can construct a presentation of the process by specifying the transition matrices:

$$\Pr(x, \nu(xw) | \nu(w)) \equiv \frac{\Pr(xw)}{\Pr(w)} \tag{2.30}$$

$$= \nu(w) T^{(x)} \mathbf{1}. \tag{2.31}$$

Focusing again on M^+, we construct $\widetilde{M}^+ = \mathcal{T}(M^+)$. Since $\widetilde{\pi} = \pi$, we can equate $\mathcal{R}_t = \mathcal{S}_t^+$ and the mixed states $\nu(w)$ are actually informing us about the causal states in M^+:

$$\nu(w) = \Pr(\mathcal{R}_0 | X_0^L = w)$$

$$= \Pr(\mathcal{S}_0^+ | X_0^L = w) \,.$$

Whenever the mixed-state presentation is an ϵ-machine, each distribution corresponds to exactly one reverse causal state. Thus, if w induces $\nu(w)$, then $\nu(w)$ is the reverse causal state induced by w. This allows us to reduce the form of $\nu(w)$ even further so that the conditioned variable is a reverse causal state. Continuing,

$$\nu(w) = \Pr(\mathcal{S}_0^+ | X_0^L = w)$$

$$= \Pr\left(\mathcal{S}_0^+ | \mathcal{S}_0^- = \epsilon^-(w)\right) \,.$$

Hence, we can calculate $H[\mathcal{S}^+ | \mathcal{S}^-]$ and obtain $\mathbf{E}$ via (2.9).

§2.10 Calculational Example

To clarify the procedure, we apply it to the Random, Noisy Copy (RnC) Process. The emphasis is on the various process presentations and mixed states that are used to calculate the excess entropy. In the next section, additional examples are provided which skip over these calculational details and, instead, focus on the analysis and interpretation.

The RnC generates a random bit with bias p. If that bit is a 0, it is copied so that the next output is also 0. However, if the bit is a 1, then with probability q, the 1 is not copied and 0 is output instead. The RnC Process is related to the *binary asymmetric channel* of communication theory [**?**].

The forward ϵ-machine has three recurrent causal states $\boldsymbol{\mathcal{S}}^+ = \{A, B, C\}$ and is shown in Fig. 2.1(a). The transition matrices $T^{(x)}$ specify $\Pr(X_0 = x, \mathcal{S}_1^+ | \mathcal{S}_0^+)$ and are given by:

$$T^{(0)} = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} 0 & p & 0 \\ 1 & 0 & 0 \\ q & 0 & 0 \end{pmatrix} \end{array}$$

and

$$T^{(1)} = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} 0 & 0 & 1-p \\ 0 & 0 & 0 \\ 1-q & 0 & 0 \end{pmatrix} \end{array} .$$

(One must explicitly calculate the equivalence classes of histories $\{\overleftarrow{x}\}$ specified in Eq. (A.1) and their associated future conditional distributions $\Pr(\overrightarrow{X}|\overleftarrow{x})$ to obtain the ϵ-machine causal states and transitions.)

These matrices are used calculate the stationary distribution π over the causal states, which is given by the left eigenvector of the stochastic matrix $T \equiv T^{(0)} + T^{(1)}$:

$$\Pr(\mathcal{S}^+) = \frac{1}{2} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} 1 & p & 1-p \end{pmatrix} \end{array} .$$

Using the $T^{(x)}$ and π, we create the time-reversed presentation $\widetilde{M}^+ = \mathcal{T}(M^+)$. This is shown in Fig. 2.1(b). Notice that the machine is not unifilar, and so it is clearly not an ϵ-machine. The transition matrices for the time-reversed presentation are given by:

$$\widetilde{T}^{(0)} = \begin{array}{c|ccc} & A & B & C \\ \hline A & 0 & p & q(1-p) \\ B & 1 & 0 & 0 \\ C & 0 & 0 & 0 \end{array} \text{ and}$$

$$\widetilde{T}^{(1)} = \begin{array}{c|ccc} & A & B & C \\ \hline A & 0 & 0 & (1-q)(1-p) \\ B & 0 & 0 & 0 \\ C & 1 & 0 & 0 \end{array} .$$

As with M^+, we calculate the stationary distribution of $\widetilde{M}^+$, denoted $\widetilde{\pi}$. However, we showed that the stationary distributions for M and $\mathcal{T}(M)$ are identical.

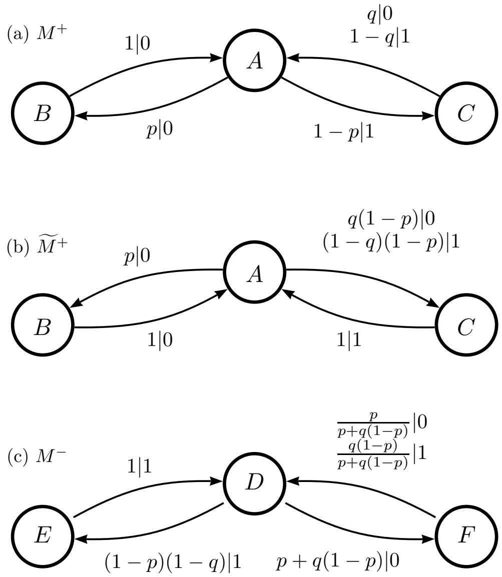

Figure 2.1: The presentations used to calculate the excess entropy for the RnC Process: (a) M^+, (b) $\widetilde{M}^+ = \mathcal{T}(M^+)$, and (c) $M^- = \mathcal{U}(\widetilde{M}^+)$. Edge labels $t|x$ give the probability $t = T^{(x)}_{\mathcal{R}\mathcal{R}'}$ of making a transition and seeing symbol x.

Now we are in a position to calculate the mixed-state presentation, $M^- = \mathcal{U}(\widetilde{M}^+)$, shown in Fig. 2.1(c). Generally, causal states can be categorized into types [**?**]. Of these, the calculation of **E** depends only on the reachable recurrent causal states. The construction of the mixed-state presentation will generate other types of causal states, such as transient causal states, but we eventually remove them.

To begin, we start with the empty word, $w = \lambda$, and append 0 and 1 to consider $\nu(0)$ and $\nu(1)$, respectively, and calculate:

$$\begin{aligned} \nu(0) &= \Pr(\mathcal{S}_0^+ | X_0 = 0) \\ &= \frac{\widetilde{\pi}\widetilde{T}^{(0)}}{\widetilde{\pi}\widetilde{T}^{(0)}\mathbf{1}} \\ &= \frac{\big(p, p, q(1-p)\big)}{2p + q(1-p)} \end{aligned}$$

and

$$\begin{aligned} \nu(1) &= \Pr(\mathcal{S}_0^+ | X_0 = 1) \\ &= \frac{\widetilde{\pi}\widetilde{T}^{(1)}}{\widetilde{\pi}\widetilde{T}^{(1)}\mathbf{1}} \\ &= \frac{\big(1, 0, 1-q\big)}{2-q} ~. \end{aligned}$$

For each mixed state, we append 0s and 1s and calculate again:

$$\begin{aligned} \nu(00) = \Pr(\mathcal{S}_0^+ | X_0^2 = 00) &= \frac{\widetilde{\pi}\widetilde{T}^{(0)}\widetilde{T}^{(0)}}{\widetilde{\pi}\widetilde{T}^{(0)}\widetilde{T}^{(0)}\mathbf{1}} ~, \\ \nu(01) = \Pr(\mathcal{S}_0^+ | X_0^2 = 01) &= \frac{\widetilde{\pi}\widetilde{T}^{(1)}\widetilde{T}^{(0)}}{\widetilde{\pi}\widetilde{T}^{(1)}\widetilde{T}^{(0)}\mathbf{1}} ~, \\ \nu(10) = \Pr(\mathcal{S}_0^+ | X_0^2 = 10) &= \frac{\widetilde{\pi}\widetilde{T}^{(0)}\widetilde{T}^{(1)}}{\widetilde{\pi}\widetilde{T}^{(0)}\widetilde{T}^{(1)}\mathbf{1}} ~, \text{ and} \\ \nu(11) = \Pr(\mathcal{S}_0^+ | X_0^2 = 11) &= \frac{\widetilde{\pi}\widetilde{T}^{(1)}\widetilde{T}^{(1)}}{\widetilde{\pi}\widetilde{T}^{(1)}\widetilde{T}^{(1)}\mathbf{1}} ~. \end{aligned}$$

Note that

$$\nu(10) = \frac{\nu(0)\widetilde{T}^{(1)}}{\nu(0)\widetilde{T}^{(1)}\mathbf{1}} ~. \tag{2.32}$$

This latter form is important in that it allows us to build mixed states from prior mixed states by prepending a symbol.

One continues constructing mixed states of longer and longer words until no more new mixed states appear. As an example, $\nu(1001) = \nu(111001)$ for the right-scanned RnC Process.

To illustrate calculating the transition probabilities, consider the transition from $\nu(00)$ to $\nu(100)$ [7]. By Eq. (2.31), we have

$$\begin{aligned} \Pr\big(1, \nu(100) | \nu(00)\big) &= \Pr(1|00) \\ &= \nu(00)\widetilde{T}^{(1)}\mathbf{1} \\ &= \frac{1-p}{1+p+q-pq} ~. \end{aligned}$$

[7] This calculation gives the probability of transitioning from a transient causal state to a recurrent causal state on seeing 1.

After constructing the mixed-state presentation, one calculates the stationary state distribution. The causal states which have $\Pr(\mathcal{S}^-) > 0$ are the recurrent causal states. These are $\boldsymbol{\mathcal{S}}^- = \{D, E, F\}$:

$$D = \nu(1001) = \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} 0 & 0 & 1 \end{pmatrix} \end{array}$$

$$E = \nu(100) = \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \end{array}$$

$$F = \nu(10) = \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} 0 & \frac{p}{p+q(1-p)} & \frac{q(1-p)}{p+q(1-p)} \end{pmatrix} \end{array} .$$

These mixed states give $\Pr(\mathcal{S}^+|\mathcal{S}^-)$ which, when combined with $\Pr(\mathcal{S}^+)$, allows us to calculate:

$$\mathbf{E} = I[\mathcal{S}^+;\mathcal{S}^-] = H[\mathcal{S}^+] - H[\mathcal{S}^+|\mathcal{S}^-] = C_\mu^+ - \chi^+$$

with

$$C_\mu^+ = 1 + \frac{H(p)}{2}$$

and

$$\chi^+ = \frac{p + q(1-p)}{2} H\left(\frac{p}{p + q(1-p)}\right) ,$$

where $H(\cdot)$ is the binary entropy function.

§2.11 Examples

With the calculational procedure laid out, we now analyze the information processing properties of several examples—two of which are familiar from symbolic dynamics.

§2.11.1 Even Process

The Even Process is a stochastic generalization of the Even System: the canonical example of a *strictly sofic* subshift—a symbolic dynamical system that cannot be expressed as a subshift of finite type [**?**, **?**]. In terms of measure, this means that the Even Process cannot be represented as a finite Markov chain; however, it has a two-state ϵ-machine representation. See Figure A.1(a). Its behavior is characterized by consecutive 1s always appearing in even blocks. With probability p, each block of 1s can be followed by a 0, which can repeat until the next even block of 1s.

Somewhat surprisingly, the Even Process turns out to be quite simple in terms of the properties we are addressing. As we will now show, the mapping between forward and reverse causal states is one-to-one and so $\chi^{\pm} = 0$. All of its internal state information is present in measurements; we call it an *explicit,* or *non-cryptic* process.

Its forward ϵ-machine has two recurrent causal states $\boldsymbol{\mathcal{S}}^{+} = \{A, B\}$ and transition matrices [**?**]:

$$T^{(0)} = \begin{array}{c} \\ A \\ B \end{array} \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{pmatrix} p & 0 \\ 0 & 0 \end{pmatrix} \end{array} \text{ and}$$

$$T^{(1)} = \begin{array}{c} \\ A \\ B \end{array} \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{pmatrix} 0 & 1-p \\ 1 & 0 \end{pmatrix} \end{array} .$$

Figure A.1(a) gives M^{+}, while A.1(b) gives M^{-}. We see that the ϵ-machines are the same and so the Even Process is causally reversible ($\Xi = 0$). Note that $\widetilde{M}^{+}$ is unifilar.

We can give general expressions for the information processing properties as a function of the probability $p = \Pr(0|A)$ of the self-loop. A simple calculation shows that

$$\Pr(\mathcal{S}^{+}) = \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{pmatrix} \frac{1}{2-p} & \frac{1-p}{2-p} \end{pmatrix} \end{array} \text{ and}$$

$$\Pr(\mathcal{S}^{-}) = \begin{array}{c} \begin{array}{cc} C & D \end{array} \\ \begin{pmatrix} \frac{1}{2-p} & \frac{1-p}{2-p} \end{pmatrix} \end{array} .$$

And so, $C_{\mu}^{+} = H\left(1/(2-p)\right)$ and $h_{\mu} = H(p)/(2-p)$. Also, since $\chi^{\pm} = 0$ for all p, we will have $\mathbf{E} = C_{\mu}^{\pm}$.

Now, let's analyze its bidirectional machine, which is shown in Fig. A.1(c). The reverse and

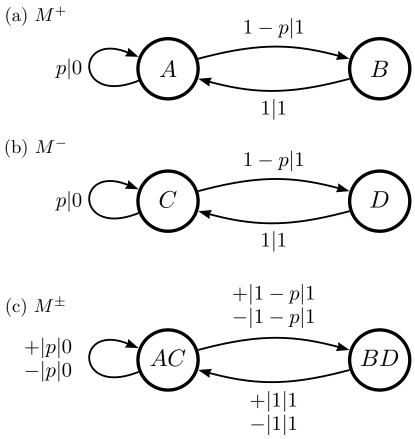

Figure 2.2: Forward and reverse ϵ-machines for the Even Process: (a) M^+ and (b) M^-. (c) The bidirectional machine $M^{\pm}$. Edge labels are prefixed by the scan direction $\{-,+\}$.

forward maps are given by:

$$\Pr(\mathcal{S}^+|\mathcal{S}^-) = \begin{array}{c c} & \begin{array}{cc} A & B \end{array} \\ \begin{array}{c} C \\ D \end{array} & \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \end{array} \text{ and}$$

$$\Pr(\mathcal{S}^-|\mathcal{S}^+) = \begin{array}{c c} & \begin{array}{cc} C & D \end{array} \\ \begin{array}{c} A \\ B \end{array} & \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \end{array} .$$

From which one calculates that $\Pr(\mathcal{S}^{\pm}) = \Pr(AC, BD) = (2/3, 1/3)$ for $p = 1/2$. This and the switching maps above give $C_\mu^{\pm} = H[\mathcal{S}^{\pm}] = H(2/3) \approx 0.9183$ bits and $\mathbf{E} = I[\mathcal{S}^+;\mathcal{S}^-] \approx 0.9183$ bits.

Without going into details to be reported elsewhere, the Even Process is also notable since it is difficult to empirically estimate its **E**. (The convergence as a function of the number of measurements is extremely slow.) Viewed in terms of the quantities C_μ^+, C_μ^-, χ^+, χ^-, and Ξ, though, it is quite simple. This illustrates one strength of the time-symmetric analysis. The latter's new and independent set of informational measures lead one to explore new regions of process space (see Fig. 2.3) and to ask structural questions not previously capable of being asked (or answered, for that matter). To see exactly why the Even Process is so simple, let's look at its causal states.

Its histories can be divided into two classes: those that end with an even number of 1s and those that end with an odd number of 1s. Similarly, its futures divide into two classes: those that

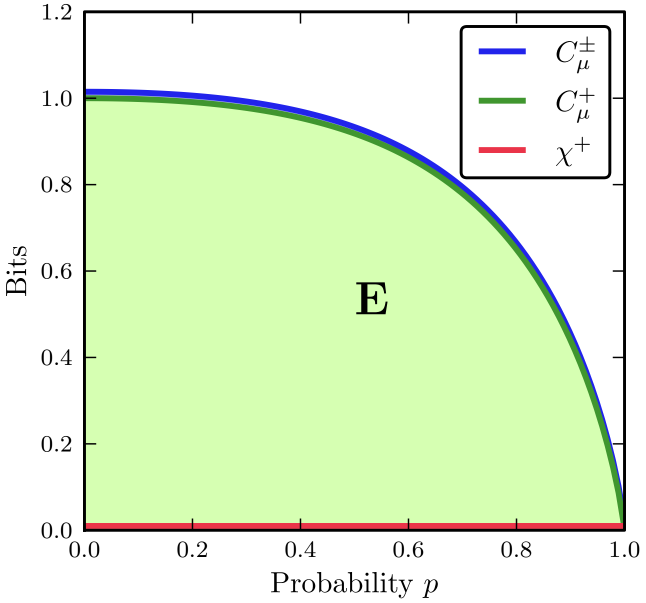

Figure 2.3: The Even Process's information processing properties—$C_\mu^\pm$, C_μ^+, and χ^+—as its self-loop probability p varies. The colored area bounded by the curves show the magnitude of **E**.

begin with an even number of 1s and those that begin with an odd number of 1s. The analysis here shows that these classes are causal states A, B, C, and D, respectively; see Fig. A.1.

Beginning with a bi-infinite string, wherever we choose to split it into $(\overleftarrow{X}, \overrightarrow{X})$, we can be in one of only two situations: either (A, C) or (B, D), where A (C) ends (begins) with an even number of 1s, and B (D) ends (begins) with an odd number of 1s. This one-to-one correspondence simultaneously implies causal reversibility $(\Xi = 0)$ and explicitness $(\chi^\pm = 0)$. Thinking in terms of the bidirectional machine, we can predict and retrodict, changing direction as often as we like and forever maintain optimal predictability and retrodictability. Since we can switch directions with no loss of information, there is no asymmetry in the loss; this reflects the process's causal reversibility.

Plotting C_μ^+, $C_\mu^\pm$, and χ^+, Fig. 2.3 rather directly illustrates these properties and shows that they are maintained across the entire process family as the self-loop probability p is varied.

§2.11.2 Golden Mean Process

The Golden Mean Process generates all binary sequences except for those with two contiguous 0s. Its name derives from the Golden Mean subshift whose topological entropy is $\log_2(\varphi)$, where

φ is the golden mean ratio. Like the Even Process, it has two recurrent causal states, but unlike the Even Process, its support is a subshift of finite type. It is describable by a chain over three Markov states that correspond to the length-2 words 01, 10, and 11.

Nominally, it is considered to be a very simple process. However, it reveals several surprising subtleties. M^+ and M^- are the same ϵ-machine—it is causally reversible ($\Xi = 0$). However, $M^\pm$ has three states and the forward and reverse state maps are no longer the identity. Thus, $\chi^\pm > 0$ and the Golden Mean Process is cryptic and so hides much of its state information from an observer.

Its forward ϵ-machine has two recurrent causal states $\boldsymbol{\mathcal{S}}^+ = \{A, B\}$ and transition matrices [**?**]:

$$T^{(0)} = \begin{array}{c} \\ A \\ B \end{array} \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{pmatrix} 0 & 1-p \\ 0 & 0 \end{pmatrix} \end{array}$$

and

$$T^{(1)} = \begin{array}{c} \\ A \\ B \end{array} \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{pmatrix} p & 0 \\ 1 & 0 \end{pmatrix} \end{array} .$$

Figure 2.4(a) gives M^+, while (b) gives M^-. We see that the ϵ-machines are the same and so the Golden Mean Process is causally reversible ($\Xi = 0$).

Again, we can give general expressions for the information processing measures as a function of the probability $p = \Pr(1|A)$ of the self-loop. The state-to-state transition matrix is the same as that for the Even Process and we also have the same causal state probabilities. Thus, we have $C_\mu = H\left(1/(2-p)\right)$ and $h_\mu = H(p)/(2-p)$ again, just as for the Even Process above. Indeed, a quick comparison of the state-transition diagrams does not reveal any overt difference with the Even Process's ϵ-machines.

However, since $\chi^\pm \neq 0$ for $p \in (0, 1)$ and since the process is also a one-dimensional spin chain, we have $\mathbf{E} = C_\mu - Rh_\mu$ with $R = 1$. (Recall Eq. (2.6).) Thus,

$$\mathbf{E} = H\left(\frac{1}{2-p}\right) - \frac{H(p)}{2-p} . \qquad (2.33)$$

Putting these closed-form expressions together gives us a graphical view of how the various information measures change as the process's parameter is varied. This is shown in Fig. 2.5.

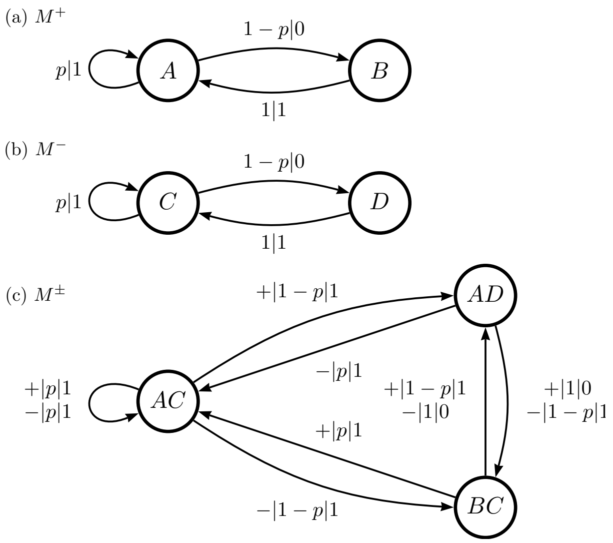

Figure 2.4: Forward and reverse ϵ-machines for the Golden Mean Process: (a) M^+ and (b) M^-. (c) The bidirectional machine $M^\pm$.

In contrast to the Even Process, the excess entropy is substantially less than the statistical complexities, the signature of a cryptic process: $\chi^\pm = H(p)/(2-p)$.

The origin of its crypticity is found by analyzing the bidirectional machine, which is shown in Fig. 2.4(c). The reverse and forward maps are given by:

$$\Pr(\mathcal{S}^+|\mathcal{S}^-) = \begin{array}{c c} & \begin{array}{cc} A & B \end{array} \\ \begin{array}{c} C \\ D \end{array} & \begin{pmatrix} p & 1-p \\ 1 & 0 \end{pmatrix} \end{array} \text{ and}$$

$$\Pr(\mathcal{S}^-|\mathcal{S}^+) = \begin{array}{c c} & \begin{array}{cc} C & D \end{array} \\ \begin{array}{c} A \\ B \end{array} & \begin{pmatrix} p & 1-p \\ 1 & 0 \end{pmatrix} \end{array} .$$

From $M^\pm$, one can calculate the stationary distribution over the bidirectional causal states: $\Pr(\mathcal{S}^\pm) = \Pr(AC, AD, BC) = \left(p, 1-p, 1-p\right)/(2-p)$. For $p = 1/2$, we obtain $C_\mu^\pm = H[\mathcal{S}^\pm] = \log_2 3 \approx 1.5850$ bits, but an $\mathbf{E} = I[\mathcal{S}^+;\mathcal{S}^-] \approx 0.2516$ bits. Thus, $\mathbf{E}$ is substantially less than the C_μs, a cryptic process: $\chi^\pm \approx 1.3334$ bits.

The Golden Mean Process is a perfect complement to the Even Process. Previously, it was viewed as a simple process for many reasons: It is based on a subshift of finite type and order-1 Markov, the causal-state process is *itself* a Golden Mean Process, it is microscopically reversible,

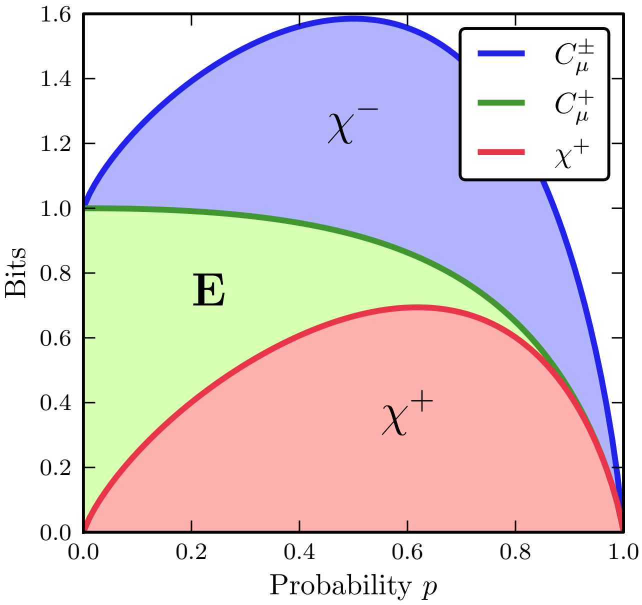

Figure 2.5: The Golden Mean Process's information processing measures—$C_\mu^\pm$, C_μ^+, and χ^+—as its self-loop probability p varies. Colored areas bounded by the curves give the magnitude at each p of χ^-, $\mathbf{E}$, and χ^+.

and $\mathbf{E}$ was exactly calculable (even before the introduction of the methods here). However, the preceding analysis shows that the Golden Mean Process displays a new feature that the Even Process does not—crypticity.

We can gain an intuitive understanding of this by thinking about classes of histories and futures. In this case, a bi-infinite string can be split in three ways $(\overleftarrow{X}, \overrightarrow{X})$: (A, C), (A, D), or (B, C), where A (C) is any past (future) that ends (begins) with a 0 and B (D) is any past (future) that ends (begins) with a 1. In terms of the bidirectional machine, there is a cost associated with changing direction. It is the *mixing* among the causal states above that is responsible for this cost. Further, this cost is symmetric because of the microscopic reversibility. Switching from prediction to retrodiction causes a loss of χ^+ bits of memory and a generation of χ^- bits of uncertainty.

Each complete round-trip state switch (e.g., forward-backward-forward) leads to a geometric reduction in state knowledge of $\mathbf{E}^2/(C_\mu^+ C_\mu^-)$. One can characterize this information loss with a half-life—the number of complete switches required to reduce state knowledge to half of its initial value.

Figure 2.5 shows that these properties are maintained across the entire Golden Mean Process family, except at extremes. When $p = 0$, it degenerates to a simple period-2 process, with $\mathbf{E} = C_\mu^+ = C_\mu^- = C_\mu^\pm = 1$ bit of memory. When $p = 1$, it is even simpler, the period-1 process, with no memory. As it approaches this extreme, $\mathbf{E}$ vanishes rapidly, leaving processes with internal state memory dominated by crypticity: $C_\mu^\pm \approx \chi^+ + \chi^-$.

§2.11.3 Random Insertion Process

Our final example is chosen to illustrate what appears to be the typical case—a cryptic, causally irreversible process. This is the random insertion process (RIP) which generates a random bit with bias p. If that bit is a 1, then it outputs another 1. If the random bit is a 0, however, it inserts another random bit with bias q, followed by a 1.

Its forward ϵ-machine has three recurrent causal states $\boldsymbol{\mathcal{S}}^+ = \{A, B, C\}$ and transition matrices:

$$T^{(0)} = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} 0 & p & 0 \\ 0 & 0 & q \\ 0 & 0 & 0 \end{pmatrix} \end{array} \text{ and}$$

$$T^{(1)} = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} 0 & 0 & 1-p \\ 0 & 0 & 1-q \\ 1 & 0 & 0 \end{pmatrix} \end{array} .$$

Figure 2.6(b) shows M^- which has four recurrent causal states $\boldsymbol{\mathcal{S}}^- = \{D, E, F, G\}$. We see that the ϵ-machines are not the same and so the RIP is causally irreversible. A direct calculation gives:

$$\Pr(\mathcal{S}^+) = \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} \frac{1}{p+2} & \frac{p}{p+2} & \frac{1}{p+2} \end{pmatrix} \end{array} \text{ and}$$

$$\Pr(\mathcal{S}^-) = \begin{array}{c} \begin{array}{cccc} D & E & F & G \end{array} \\ \begin{pmatrix} \frac{1}{p+2} & \frac{1-pq}{p+2} & \frac{pq}{p+2} & \frac{p}{p+2} \end{pmatrix} \end{array} .$$

If $p = q = 1/2$, for example, these give us $C_\mu^+ \approx 1.5219$ bits, $C_\mu^- \approx 1.8464$ bits, and $h_\mu = 3/5$ bits per measurement. The causal irreversibility is $\Xi \approx 0.3245$ bits.

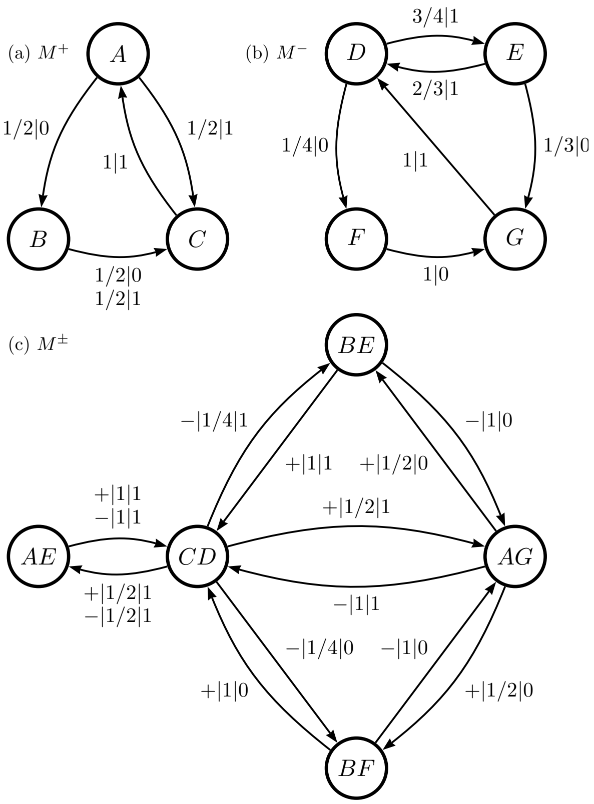

Figure 2.6: Forward and reverse ϵ-machines for the RIP with $p = q = 1/2$: (a) M^+ and (b) M^-. (c) The bidirectional machine $M^\pm$ also for $p = q = 1/2$. (Reprinted with permission from Refs. [**?**].)

Let's analyze the RIP bidirectional machine, which is shown in Fig. 2.6(c) for $p = q = 1/2$. The reverse and forward maps are given by:

$$\Pr(\mathcal{S}^+|\mathcal{S}^-) = \begin{array}{c c} & \begin{array}{ccc} A & B & C \end{array} \\ \begin{array}{c} D \\ E \\ F \\ G \end{array} & \left(\begin{array}{ccc} 0 & 0 & 1 \\ 2/3 & 1/3 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{array}\right) \end{array} \text{ and}$$

$$\Pr(\mathcal{S}^-|\mathcal{S}^+) = \begin{array}{c c} & \begin{array}{cccc} D & E & F & G \end{array} \\ \begin{array}{c} A \\ B \\ C \end{array} & \left(\begin{array}{cccc} 0 & 1/2 & 0 & 1/2 \\ 0 & 1/2 & 1/2 & 0 \\ 1 & 0 & 0 & 0 \end{array}\right) \end{array} .$$

Or, for general p and q, we have

$$\Pr(\mathcal{S}^+, \mathcal{S}^-) = \frac{1}{(p+2)} \begin{array}{c c} & \begin{array}{cccc} D & E & F & G \end{array} \\ \begin{array}{c} A \\ B \\ C \end{array} & \left(\begin{array}{cccc} 0 & 1-p & 0 & p \\ 0 & p(1-q) & pq & 0 \\ 1 & 0 & 0 & 0 \end{array}\right) \end{array} .$$

By way of demonstrating the exact analysis now possible, **E**'s closed-form expression for the RIP family is

$$\mathbf{E} = \log_2(p+2) - \frac{p \log_2 p}{p+2} - \frac{1-pq}{p+2} H\left(\frac{1-p}{1-pq}\right) .$$

The first two terms on the RHS are C_μ^+ and the last is χ^+.

Setting $p = q = 1/2$, one calculates that $\Pr(\mathcal{S}^\pm) = \Pr(AE, AG, BE, BF, CD) = (1/5, 1/5, 1/10, 1/10, 2/5)$. This and the joint distribution give $C_\mu^\pm = H[\mathcal{S}^\pm] \approx 2.1219$ bits, but an $\mathbf{E} = I[\mathcal{S}^+; \mathcal{S}^-] \approx 1.2464$ bits. That is, the excess entropy (the apparent information) is substantially less than the statistical complexities (stored information)—a moderately cryptic process: $\chi^\pm \approx 0.8755$ bits.

Figure 2.7 shows how the RIP's informational character varies along one-dimensional paths in its parameter space: $(p, q) \in [0, 1]^2$. The four extreme-p and -q paths illustrate that the RIP borders on (i) non-cryptic, reversible processes (solid line), (ii) semi-cryptic, irreversible processes (long dash), (iii) cryptic, reversible processes (short dash), and (iv) cryptic, irreversible processes (very short dash). The horizontal path ($q = 0.5$) and two diagonal paths ($p = q$ and $p = 1 - q$) show the typical cases within the parameter space of cryptic, irreversible processes.

§2.12 Conclusions

Casting stochastic dynamical systems in a time-agnostic framework revealed a landscape that quickly led one away from familiar entrances, along new and unfamiliar pathways. Old informational quantities were put in a new light, new relationships among them appeared, and explicit calculation methods became available. The most unexpected appearances, though, were the new information measures that captured novel properties of general processes.

Excess entropy, a familiar quantity in a long-applied family of mutual informations, is often estimated [**?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**] and is broadly considered an important information measure for organization in complex systems. The exact analysis afforded by our time-agnostic framework gave an important calibration in our studies. Specifically, it showed how difficult accurate estimates of the excess entropy can be. While we intend to report on this in some detail elsewhere, suffice it to say that the convergence of empirical estimates of **E**, in even very benign (and low statistical complexity) cases, can be so slow as to make estimation computationally intractable. This problem would never have been clear without the closed-form expressions. It, with nothing else said, calls into doubt many of the reported uses and estimations of excess entropy and related mutual information measures.

Fortunately, we now have access to the analytic calculation of the excess entropy from the ϵ-machine. Note that the latter is no more difficult to estimate than, say, estimating the entropy rate of an information source. (Both are dominated by obtaining accurate estimates of a process's sequence distribution.) Notably, the calculation relied on connecting prediction and retrodiction, which we accomplished via the composition of the time-reversal operation on ϵ-machines and the mixed-state-presentation algorithm. As the analyses of the various example processes illustrated, the technique yields closed-form expressions for **E**. More generally, though, the explicit relationship between a process's ϵ-machine and its excess entropy clearly demonstrates why the statistical complexity, and not the excess entropy, is the information stored in the present.

In addition to the analytical advantage of having **E** in hand, we learned a pointed lesson about the difference between prediction (reflected in **E**) and modeling (reflected in C_μ). In particular, a system's causal representation yields more direct access to fundamental properties than others—such as, histograms of word counts or general hidden Markov models. The differences between prediction and modeling unearthed new information measures—crypticity and causal irreversibility.

Crypticity describes the amount of stored state information that is not shared in the measurement sequence. One might think of this as "wasted" information, although the minimality of the ϵ-machine suggests that this waste is necessary—that is, an intrinsic property of the process. Possibly we could better think of this as modeling overhead.

When analyzing time symmetry, one can use notions such as microscopic reversibility or, more broadly, reversible support. We introduced the yet-broader notion of causal irreversibility Ξ. It has the advantage of being scalar rather than Boolean and so has something to say quantitatively about all processes. Also, it derives naturally from its simple relationship to **E** and $\chi^{\pm}$. In this light, microscopic reversibility appears to be too strong a criterion, missing important structural properties.

First, we described parallel predictive and retrodictive causal models joined by the switching maps. Then, the time-agnostic perspective required expanding the space of representations. This expansion allowed us to define a bidirectional machine that compressed C_μ^+ and C_μ^- into $C_\mu^\pm$, an object that can be somewhat non-intuitive.

For example, the three-state bidirectional machine for the Golden Mean Process might seem overcomplicated given that the forward and reverse ϵ-machines each require just two states. Surprisingly, three states are indeed required if one wishes to predict *and* retrodict; whereas just two states are required if one wants only to predict or only to retrodict. Alternatively, one might also wonder why the bidirectional machine does not have four states, if it truly can predict and retrodict. This is because the bidirectional machine compresses the two processes, providing a new conception of the amount of information stored in the present.

The operational meaning of the bidirectional machine certainly warrants further attention. In particular, it seems likely that its nonunifilarity has not yet been fully appreciated. One might wish to consider, for example, a unifilar representation of it. Somewhat hopefully, we end by noting that the bidirectional machine suggests an extension of ϵ-machine analysis beyond one-dimensional processes.

Acknowledgments

Chris Ellison was partially supported on a GAANN fellowship. The Network Dynamics Program funded by Intel Corporation also partially supported this work.

§.1 Appendix: The Mixed-State Presentation is Sufficient to Calculate the Switching Maps

While we conjecture that the mixed-state operation $\mathcal{U}(\widetilde{M}^+)$ yields an ϵ-machine, this remains an open problem. Our conjecture, however, is based on a rather large number of test cases in which it is an ϵ-machine. Fortunately for our present needs, we can show that $\mathcal{U}(\widetilde{M}^+)$ is sufficient for calculating the conditional probability distribution $\Pr(\mathcal{S}^+|\mathcal{S}^-)$.

For a moment, ignore the details of forward and reverse machines and simply consider machines A and B such that $\mathcal{U}(A) = B$ where neither A nor B is necessarily an ϵ-machine. We would like to learn the conditional probability distribution $\Pr(\mathcal{R}_A|\mathcal{R}_B)$, where $\mathcal{R}_A$ and $\mathcal{R}_B$ are A's and B's states, respectively.

Proposition 11. *B's states are mixed states of A.*

Proof. *We use the mixed-state presentation algorithm to form states based on the transition matrices of A. If a state $\mathcal{R}_B$ is induced by a word w, then:*

$$\mathcal{R}_B = \frac{\pi_A T_A^{\omega}}{\pi_A T_A^{w} \mathbf{1}} \ . \quad \square$$

We now show that B is deterministic.

Proposition 12. $H[\mathcal{R}'|\mathcal{R}, X] = 0$ *for machine B.*

Proof. *Although any given state in B will generally be a distribution over states in A, each of these distributions* defines *a state of B. The particular state of B (or distribution over states in A), $\mathcal{R}'$, that follows $\mathcal{R}$ and X can be written:*

$$\mathcal{R}'_B = \frac{\pi_A T_A^{\omega} T^X}{\pi_A T_A^{\omega} T^X \eta} \ .$$

So, by construction, B is deterministic. □

Moreover, $\mathcal{R}_B$ is a refinement of $\boldsymbol{\mathcal{S}}_B$.

Proposition 13. *Two pasts that induce the same state in B must be pasts in the same causal state of B's ϵ-machine.*

Proof. *The future probability distribution given a word is exactly the future probability distribution given the mixed state induced by that word:*

$$\Pr(\overrightarrow{X}|\omega) = \frac{\pi T^{\omega} T^{\overrightarrow{X}}}{\pi T^{\omega} T^{\overrightarrow{X}} \eta}$$

$$\Pr(\overrightarrow{X}|\mu(\omega)) = \frac{\frac{\pi T^{\omega}}{\pi T^{\omega}\eta} T^{\overrightarrow{X}}}{\frac{\pi T^{\omega} T^{\overrightarrow{X}} \eta}{\pi T^{\omega}\eta}} = \frac{\pi T^{\omega} T^{\overrightarrow{X}}}{\pi T^{\omega} T^{\overrightarrow{X}} \eta}$$

Therefore, if two words induce the same mixed state, the future probability distribution conditioned on those words are the same. This means that those words are causally equivalent and thus in the same causal state. □

Now we show how, even in this very generic case, we can calculate the relevant conditional probability distribution.

The mixed-state construction of B implicitly has given us $\Pr(\mathcal{R}_A|\mathcal{R}_B)$, which we can use to find $\Pr(\mathcal{R}_A|\mathcal{S}_B)$, our goal:

$$\begin{aligned}
\Pr(\mathcal{R}_A|\mathcal{S}_B) &= \sum_{\mathcal{R}_B} \Pr(\mathcal{R}_A|\mathcal{S}_B, \mathcal{R}_B)\Pr(\mathcal{R}_B|\mathcal{S}_B) \\
&= \sum_{\mathcal{R}_B} \Pr(\mathcal{R}_A|\mathcal{R}_B)\Pr(\mathcal{R}_B|\mathcal{S}_B) \\
&= \sum_{\mathcal{R}_B} \Pr(\mathcal{R}_A|\mathcal{R}_B)\Pr(\mathcal{S}_B|\mathcal{R}_B)\frac{\Pr(\mathcal{R}_B)}{\Pr(\mathcal{S}_B)} \\
&= \sum_{\mathcal{R}_B} \Pr(\mathcal{R}_A|\mathcal{R}_B)\delta_{\mathcal{R}_B\in\mathcal{S}_B}\frac{\Pr(\mathcal{R}_B)}{\Pr(\mathcal{S}_B)} \\
&= \sum_{\mathcal{R}_B} \Pr(\mathcal{R}_A|\mathcal{R}_B)\frac{\Pr(\mathcal{R}_B)}{\Pr(\mathcal{S}_{\mathcal{R}_B})} \, .
\end{aligned}$$

The second line follows since $\boldsymbol{\mathcal{R}}_B$ is a refinement of $\boldsymbol{\mathcal{S}}_B$. The third line is an application of Bayes Rule. The fourth line follows again from the refinement. The final form reminds us that $\boldsymbol{\mathcal{S}}_B$ is not a free variable.

To sum up, we calculate the conditional distribution using this final form as follows. The first factor is found by applying $\mathcal{U}$ to A. Granting ourselves the ability to ascertain predictive equality among a finite set of states $\boldsymbol{\mathcal{R}}_B$, we determine if $\mathcal{R}_B \in \mathcal{S}_B$ for each $\mathcal{R}_B$. Lastly, we compute the stationary distribution over the states of B and divide by the stationary probability of the corresponding causal state.

In effect, this establishes a general method for computing the conditional probability of states from the "input" machine given a state of the "resultant" machine. We can now recall the specific context of forward and reverse ϵ-machines and apply this technique to calculate $\mathbf{E}$ in the case where the resultant machine $\mathcal{T}(M^+)$ is not an ϵ-machine.

The input machine is the reversed ϵ-machine $\mathcal{T}(M^+)$, whose states $\widetilde{\boldsymbol{\mathcal{S}}}^+$ are in one-to-one correspondence with $\boldsymbol{\mathcal{S}}^+$. Thus, the previous result:

$$\Pr(\mathcal{R}_A|\mathcal{S}_B) = \sum_{\mathcal{R}_B} \Pr(\mathcal{R}_A|\mathcal{R}_B)\frac{\Pr(\mathcal{R}_B)}{\Pr(\mathcal{S}_{\mathcal{R}_B})}$$

now becomes:

$$\Pr(\mathcal{S}_A|\mathcal{S}_B) = \sum_{\mathcal{R}_B} \Pr(\mathcal{S}_A|\mathcal{R}_B)\frac{\Pr(\mathcal{R}_B)}{\Pr(\mathcal{S}_{\mathcal{R}_B})}$$

or, more specifically,

$$\Pr(\mathcal{S}^+|\mathcal{S}^-) = \sum_{\mathcal{R}_B} \Pr(\mathcal{S}^+|\mathcal{R}_B)\frac{\Pr(\mathcal{R}_B)}{\Pr(\mathcal{S}^-_{\mathcal{R}_B})} \, .$$

From which we readily calculate **E** using:

$$\begin{aligned}\mathbf{E} &= I[\mathcal{S}^+;\mathcal{S}^-] \\ &= H[\mathcal{S}^+] - H[\mathcal{S}^+|\mathcal{S}^-] \ .\end{aligned}$$

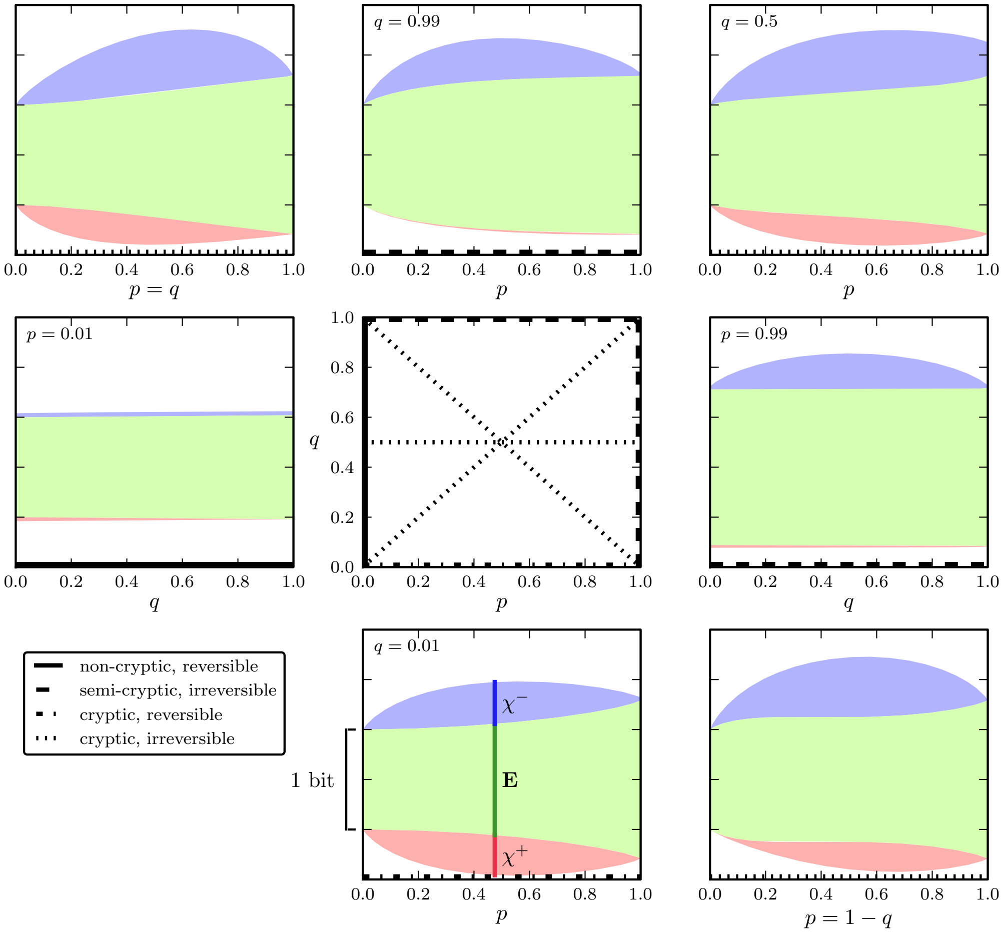

Figure 2.7: The Random Insertion Process's information processing measures as its two probability parameters p and q vary. The central square shows the (p, q) parameter space, with solid and dashed lines indicating the paths in parameter space for each of the other information versus parameter plots. The latter's vertical axes are scaled so that two tick marks measure 1 bit of information. The inset legend indicates the class of process illustrated by the paths. Colored areas give the magnitude of χ^-, $\mathbf{E}$, and χ^+.

APPENDIX A
IACP

§A.1 Introduction

The data of phenomena come to us through observation. A large fraction of the theoretical activity of model building, though, focuses on internal mechanism. How are observation and modeling related? A first step is to frame the problem in terms of hidden processes—internal mechanisms probed via instruments that, in particular, need not accurately report a process's internal state. A practical second step is to measure the difference between internal structure and the information in observations.

We recently established that the amount of observed information a process communicates from the past to the future—the *excess entropy*—is the mutual information between its forward- and reverse-time minimal causal representations [**?**, **?**]. This closed-form expression gives a concrete connection between the observed information and a process's internal structure.

Excess entropy, and related mutual information quantities, are widely used diagnostics for complex systems. They have been applied to detect the presence of organization in dynamical systems [**?**, **?**, **?**, **?**], in spin systems [**?**, **?**, **?**], in neurobiological systems [**?**, **?**], and even in language [**?**, **?**], to mention only a very few uses. Thus, understanding how much internal state structure is reflected in the excess entropy is critical to whether or not these and other studies of complex systems can draw structural inferences about the internal mechanisms that produce observed behavior.

Unfortunately, there is a fundamental problem. The excess entropy is *not* the internal state information the process stores—rather, the latter is the process's *statistical complexity* [**?**, **?**]. On the positive side, there is a diagnostic. The difference between, if you will, experiment and theory (between observed information and internal structure) is controlled by the difference between a process's excess entropy and its statistical complexity. This difference is called the *crypticity*—how much internal state information is inaccessible [**?**, **?**]. Here we introduce a classification of processes using a systematic expansion of crypticity.

The starting point is *computational mechanics*'s minimal causal representation of a stochastic process $\mathcal{P}$—the ϵ*-machine* [**?**, **?**]. There, a process is viewed as a channel that communicates information from the past, $\overleftarrow{X} = \ldots X_{-3}X_{-2}X_{-1}$, to the future, $\overrightarrow{X} = X_0X_1X_2\ldots$. ($X_t$ takes values in a finite measurement alphabet $\mathcal{A}$.) The excess entropy is the shared (or mutual) information between the past and the future: $\mathbf{E} = I[\overleftarrow{X};\overrightarrow{X}]$. The amount of historical information that a process stores in the present is different. It is given by the Shannon information $C_\mu = H[\mathcal{S}]$ of the distribution over the ϵ-machine's *causal states* $\boldsymbol{\mathcal{S}}$. C_μ is called the *statistical complexity* and the causal states are sets of pasts $\overleftarrow{x}$ that are equivalent for prediction [**?**]:

$$\epsilon(\overleftarrow{x}) = \{\overleftarrow{x}' : \Pr(\overrightarrow{X}|\overleftarrow{x}) = \Pr(\overrightarrow{X}|\overleftarrow{x}')\} . \tag{A.1}$$

Causal states have a Markovian property that they render the past and future statistically independent; they *shield* the future from the past [**?**]:

$$\Pr(\overleftarrow{X},\overrightarrow{X}|\mathcal{S}) = \Pr(\overleftarrow{X}|\mathcal{S})\Pr(\overrightarrow{X}|\mathcal{S}) . \tag{A.2}$$

ϵ-Machines are also *unifilar* [**?**, **?**]: From the start state, each observed sequence $\ldots x_{-3}x_{-2}x_{-1}\ldots$ corresponds to one and only one sequence of causal states. The signature of unifilarity is that on knowing the current state and measurement, the uncertainty in the next state vanishes: $H[\mathcal{S}_{t+1}|\mathcal{S}_t,X_t] = 0$.

Although they are not the same, the basic relationship between these quantities is clear: $\mathbf{E}$ is the process's channel utilization and C_μ is the sophistication of that channel. Their difference, one of our main concerns in the following, indicates how a process stores, manipulates, and hides internal state information.

Until recently, $\mathbf{E}$ could not be as directly calculated from the ϵ-machine as the process's entropy rate h_μ and its statistical complexity. References [**?**] and [**?**] solved this problem, giving a closed-form expression for the excess entropy:

$$\mathbf{E} = I[\mathcal{S}^+;\mathcal{S}^-] , \tag{A.3}$$

where $\mathcal{S}^+$ are the causal states of the process scanned in the "forward" direction and $\mathcal{S}^-$ are the causal states of the process scanned in the "reverse" time direction.

This result comes in a historical context. Some time ago, an explicit expression for the excess entropy had been developed from the Hamiltonian for one-dimensional spin chains with range-R interactions [**?**]:

$$\mathbf{E} = C_\mu - R\, h_\mu \, . \tag{A.4}$$

One-dimensional spin chains are special cases of order-R Markov processes. For this more general class of processes, a similar, but slightly less compact form is known:

$$\mathbf{E} = H[X_0^R] - R\, h_\mu \, , \tag{A.5}$$

where $X_0^R = X_0, \ldots, X_{R-1}$. It has also been known for some time that the statistical complexity is an upper bound on the excess entropy [**?**]:

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

which follows from the equality derived there:

$$\mathbf{E} = C_\mu - H[\mathcal{S}^+ | \overrightarrow{X}] \, .$$

Using forward and reverse ϵ-machines, Ref. [**?**] extended this, deriving the closed-form expression for $\mathbf{E}$ in Eqn. A.3 and two new bounds on $\mathbf{E}$: $\mathbf{E} \leq C_\mu^-$ and $\mathbf{E} \leq C_\mu^+$. It also showed that:

$$H[\mathcal{S}^+ | \overrightarrow{X}] = H[\mathcal{S}^+ | \mathcal{S}^-] \tag{A.6}$$

and identified this quantity as controlling how a process hides its internal state information. For this reason, it is called the process's *crypticity*:

$$\chi^+ = H[\mathcal{S}^+ | \overrightarrow{X}] \, . \tag{A.7}$$

In the context of forward and reverse ϵ-machines, one must distinguish two crypticities; depending on the scan direction one has:

$$\chi^+ = H[\mathcal{S}^+ | \mathcal{S}^-] \text{ or}$$

$$\chi^- = H[\mathcal{S}^- | \mathcal{S}^+] \, .$$

In the following we will not concern ourselves with reverse representations and so can simplify the notation, using C_μ for C_μ^+ and χ for χ^+.

Here we show that, for a restricted class of processes, the crypticity in Eqn. A.6 can be systematically expanded to give an alternative closed-form to the excess entropy in Eqn. A.3. One ancillary benefit is a new and, we argue, natural hierarchy of processes in terms of information accessibility.

§A.2 k-Crypticity

The process classifications based on spin-block length and order-R Markov are useful. They give some insight into the nature of the kinds of process we can encounter and, concretely, they allow for closed-form expressions for the excess entropy (and other system properties). In a similar vein, we wish to carve the space of processes with a new blade. We define the class of *k-cryptic* processes and develop their properties and closed-form expressions for their excess entropies.

For convenience, we need to introduce several shorthands. First, to denote a symbol sequence that begins at time t and is L symbols long, we write X_t^L. Note that X_t^L includes X_{t+L-1}, but not X_{t+L}. Second, to denote a symbol sequence that begins at time t and continues on to infinity, we write $\overrightarrow{X}_t$. Analogously, the causal state at time t is denoted $\mathcal{S}_t$, and a sequence of states beginning at time t that is L states long is denoted $\mathcal{S}_t^L$.

Definition. *The* k-crypticity criterion *is satisfied when*

$$H[\mathcal{S}_k | \overrightarrow{X}_0] = 0\,. \tag{A.8}$$

Definition. *A* k-cryptic *process is one for which the process's ϵ-machine satisfies the k-crypticity criterion.*

Definition. *An* ∞-cryptic *process is one for which the process's ϵ-machine does not satisfy the k-crypticity criterion for any finite k.*

Lemma 1. $H[\mathcal{S}_k | \overrightarrow{X}_0]$ *is a nonincreasing function of k.*

Proof. This follows directly from stationarity and the fact that conditioning on more random variables cannot increase entropy:

$$H[\mathcal{S}_{k+1} | \overrightarrow{X}_0] = H[\mathcal{S}_k | \overrightarrow{X}_{-1}] \leq H[\mathcal{S}_k | \overrightarrow{X}_0]\,.$$

□

Lemma 2. *If $\mathcal{P}$ is k-cryptic, then $\mathcal{P}$ is also j-cryptic for all $j > k$.*

Proof. Being k-cryptic implies $H[\mathcal{S}_k | \overrightarrow{X}_0] = 0$. Applying Lem. 1, $H[\mathcal{S}_j | \overrightarrow{X}_0] \leq H[\mathcal{S}_k | \overrightarrow{X}_0] = 0$. By positivity of entropy, we conclude that $\mathcal{P}$ is also j-cryptic. □

This provides us with a new way of partitioning the space of processes. We create a parametrized class of sets $\{\chi_k : k = 0, 1, 2, \ldots\}$, where $\chi_k = \{\mathcal{P} : \mathcal{P} \text{ is } k\text{-cryptic and not } (k-1)\text{-cryptic}\}$.

The following result provides a connection to a very familiar class of processes.

Proposition 14. *If a process $\mathcal{P}$ is order-k Markov, then it is k-cryptic.*

Proof. If $\mathcal{P}$ is order-k Markov, then $H[\mathcal{S}_k|X_0^k] = 0$. Conditioning on more variables does not increase uncertainty, so:

$$H[\mathcal{S}_k|X_0^k, \overrightarrow{X}_k] = 0 \, .$$

But the lefthand side is $H[\mathcal{S}_k|\overrightarrow{X}_0]$. Therefore, $\mathcal{P}$ is k-cryptic. □

Note that the converse of Prop. 14 is not true. For example, the Even Process (EP), the Random Noisy Copy Process (RnC), and the Random Insertion Process (RIP) (see Ref. [**?**] and Ref. [**?**]), are all 1-cryptic, but are not order-R Markov for any finite R.

Note also that Prop. 14 does not preclude an order-k Markov process from being j-cryptic, where $j < k$. Later we will show an example demonstrating this.

Given a process, in general one will not know its cryptic order. One way to investigate this is to study the sequence of estimates of χ at different orders. To this end, we define the k-cryptic approximation.

Definition. *The* k-cryptic approximation *is defined as*

$$\chi(k) = H[\mathcal{S}_0|X_0^k, \mathcal{S}_k] \, .$$

§A.2.1 The k-Cryptic Expansion

We will now develop a systematic expansion of χ to order k in which $\chi(k)$ appears directly and the k-crypticity criterion plays the role of an error term.

Theorem 3. *The process crypticity is given by*

$$\chi = \chi(k) + H[\mathcal{S}_k|\overrightarrow{X}_0] \, . \tag{A.9}$$

Proof. We calculate directly, starting from the definition, adding and subtracting the k-crypticity criterion term from χ's definition, Eqn. A.7:

$$\chi = H[\mathcal{S}_0|\overrightarrow{X}_0] - H[\mathcal{S}_k|\overrightarrow{X}_0] + H[\mathcal{S}_k|\overrightarrow{X}_0] \, .$$

We claim that the first two terms are $\chi(k)$. Expanding the conditionals in the purported $\chi(k)$ terms and then canceling, we get joint distributions:

$$H[\mathcal{S}_0|\overrightarrow{X}_0] - H[\mathcal{S}_k|\overrightarrow{X}_0] = H[\mathcal{S}_0, \overrightarrow{X}_0] - H[\mathcal{S}_k, \overrightarrow{X}_0] \, .$$

Now, splitting the future into two pieces and using this to write conditionals, the righthand side becomes:

$$H[\overrightarrow{X}_k|\mathcal{S}_0, X_0^k] + H[\mathcal{S}_0, X_0^k] - H[\overrightarrow{X}_k|\mathcal{S}_k, X_0^k] - H[\mathcal{S}_k, X_0^k] .$$

Appealing to the ϵ-machine's unifilarity, we then have:

$$H[\overrightarrow{X}_k|\mathcal{S}_k] + H[\mathcal{S}_0, X_0^k] - H[\overrightarrow{X}_k|\mathcal{S}_k, X_0^k] - H[\mathcal{S}_k, X_0^k] .$$

Now, applying causal shielding gives:

$$H[\overrightarrow{X}_k|\mathcal{S}_k] + H[\mathcal{S}_0, X_0^k] - H[\overrightarrow{X}_k|\mathcal{S}_k] - H[\mathcal{S}_k, X_0^k] .$$

Canceling terms, this simplifies to:

$$H[\mathcal{S}_0, X_0^k] - H[\mathcal{S}_k, X_0^k] .$$

We now re-expand, using unifilarity to give:

$$H[\mathcal{S}_0, X_0^k, \mathcal{S}_k] - H[\mathcal{S}_k, X_0^k] .$$

Finally, we combine these, using the definition of conditional entropy, to simplify again:

$$H[\mathcal{S}_0|X_0^k, \mathcal{S}_k] .$$

Note that this is our definition of $\chi(k)$.

This establishes our original claim:

$$\chi = \chi(k) + H[\mathcal{S}_k|\overrightarrow{X}_0] ,$$

with the k-crypticity criterion playing the role of an approximation error.

□

Corollary 6. *A process $\mathcal{P}$ is k-cryptic if and only if*

$$\chi = \chi(k) .$$

Proof. Given the order-k expansion of χ just developed, we now assume the k-crypticity criterion is satisfied; viz., $H[\mathcal{S}_k|\overrightarrow{X}_0] = 0$. Thus, we have from Eqn. A.9:

$$\chi = \chi(k) .$$

Likewise, assuming $\chi = \chi(k)$ requires, by Eqn. A.9 that $H[\mathcal{S}_k|\overrightarrow{X}_0] = 0$ and thus the process is k-cryptic. □

Corollary 7. *For any process, $\chi(0) = 0$.*

Proof.

$$\begin{aligned}\chi(0) &= H[\mathcal{S}_0 | X_0^0, \mathcal{S}_0] \\ &= H[\mathcal{S}_0 | \mathcal{S}_0] \\ &= 0\,.\end{aligned}$$

□

§A.2.2 Convergence

Proposition 15. *The approximation $\chi(k)$ is a nondecreasing function of k.*

Proof. Lem. 1 showed that $H[\mathcal{S}_k | \overrightarrow{X}_0]$ is a nonincreasing function of k. By Thm. 3, $\chi(k)$ must be a nondecreasing function of k. □

Corollary 8. *Once $\chi(k)$ reaches the value χ, $\chi(j) = \chi$ for all $j > k$.*

Proof. If there exists such a k, then by Thm. 3 the process is k-cryptic. By Lem. 2, the process is j-cryptic for all $j > k$. Again, by Thm. 3, $\chi(j) = \chi$. □

Corollary 9. *If there is a $k \geq 1$ for which $\chi(k) = 0$, then $\chi(1) = 0$.*

Proof. By positivity of the conditional entropy $H[\mathcal{S}_0 | X_0, \mathcal{S}_1]$, $\chi(1) \geq 0$. By the nondecreasing property of $\chi(k)$ from Prop. 15, $\chi(1) \leq \chi(k) = 0$. Therefore, $\chi(1) = 0$. □

Corollary 10. *If $\chi(1) = 0$, then $\chi(k) = 0$ for all k.*

Proof. Applying stationarity, $\chi(1) = H[\mathcal{S}_0 | X_0, \mathcal{S}_1] = H[\mathcal{S}_k | X_k, \mathcal{S}_{k+1}]$. We are given $\chi(1) = 0$ and so $H[\mathcal{S}_k | X_k, \mathcal{S}_{k+1}] = 0$. We use this below. Expanding $\chi(k+1)$,

$$\begin{aligned}\chi(k+1) &= H[\mathcal{S}_0 | X_0^{k+1}, \mathcal{S}_{k+1}] \\ &= H[\mathcal{S}_0 | X_0^k, X_k, \mathcal{S}_{k+1}] \\ &= H[\mathcal{S}_0 | X_0^k, \mathcal{S}_k, X_k, \mathcal{S}_{k+1}] \\ &\leq H[\mathcal{S}_0 | X_0^k, \mathcal{S}_k] \\ &= \chi(k)\,.\end{aligned}$$

The third line follows from $\chi(1) = 0$. By Prop. 15, $\chi(k+1) \geq \chi(k)$. Therefore, $\chi(k+1) = \chi(k)$. Finally, using $\chi(1) = 0$, we have by induction that $\chi(k) = 0$ for all k. □

Corollary 11. *If there is a $k \geq 1$ for which $\chi(k) = 0$, then $\chi(j) = 0$ for all $j \geq 1$.*

Proof. This follows by composing Cor. 9 with Cor. 10. □

Together, the proposition and its corollaries show that $\chi(k)$ is a nondecreasing function of k which, if it reaches χ at a finite k, remains at that value for all larger k.

Proposition 16. *The cryptic approximation $\chi(k)$ converges to χ as $k \to \infty$.*

Proof. Note that $\chi = \lim_{k\to\infty} H[\mathcal{S}_0|X_0^k]$ and recall that $\chi(k) = H[\mathcal{S}_0|X_0^k, \mathcal{S}_k]$. We show that the difference approaches zero:

$$\begin{aligned} H[\mathcal{S}_0|X_0^k] - H[\mathcal{S}_0|X_0^k, \mathcal{S}_k] & \\ &= H[\mathcal{S}_0, X_0^k] - H[X_0^k] \\ &\quad - H[\mathcal{S}_0, X_0^k, \mathcal{S}_k] + H[X_0^k, \mathcal{S}_k] \\ &= H[\mathcal{S}_0, X_0^k] - H[X_0^k] \\ &\quad - H[\mathcal{S}_0, X_0^k] + H[X_0^k, \mathcal{S}_k] \\ &= H[X_0^k, \mathcal{S}_k] - H[X_0^k] \\ &= H[\mathcal{S}_k|X_0^k] . \end{aligned}$$

Moreover, $\lim_{k\to\infty} H[\mathcal{S}_k|X_0^k] = 0$ by the ϵ map from pasts to causal states of Eqn. A.1. Therefore, as $k \to \infty$, $\chi(k) \to \chi$. □

§A.2.3 Excess Entropy for k-Cryptic Processes

Given a k-cryptic process, we can calculate its excess entropy in a form that involves a sum of $\propto |\mathcal{A}^k|$ terms, where each term involves products of k matrices. Specifically, we have the following.

Corollary 12. *A process $\mathcal{P}$ is k-cryptic if and only if* $\mathbf{E} = C_\mu - \chi(k)$.

Proof. From Ref. [**?**], we have $\mathbf{E} = C_\mu - \chi$, and by Cor. 6, $\chi = \chi(k)$. Together, these complete the proof. □

The following proposition is a simple and useful consequence of the class of k-cryptic processes.

Corollary 13. *A process $\mathcal{P}$ is 0-cryptic* if *and* only if $\mathbf{E} = C_\mu$.

Proof. If $\mathcal{P}$ is 0-cryptic, then $\mathbf{E} = C_\mu - \chi(0)$ and Cor. 7 says that $\chi(0) = 0$. To establish the opposite direction, note that $\mathbf{E} = C_\mu$ implies $\chi = 0$. Applying Cor. 7 shows $\chi = \chi(0)$, and so the process is 0-cryptic by Cor. 6. □

§A.2.4 Crypticity of Spin Chains

Now, we provide results on the crypticity of one-dimensional spin chains to complement prior results on Markovity and excess entropy. First recall Eqn. A.5, which gives the excess entropy for order-R Markov processes:

$$\mathbf{E} = H[X_0^R] - R\, h_\mu\,.$$

By Prop. 14, such processes are also R-cryptic and so:

$$\mathbf{E} = C_\mu - \chi(R)\,.$$

One-dimensional spin chains are precisely those order-R Markov processes for which the statistical complexity, $C_\mu \equiv H[\mathcal{S}_R]$, equals the entropy over R-blocks, $H[X_0^R]$. Reference [**?**] stated a condition under which equality held in terms of transfer matrices. Here, we state a simpler condition by equating two chain-rule expansions of $H[X_0^R, \mathcal{S}_R]$:

$$H[X_0^R|\mathcal{S}_R] + H[\mathcal{S}_R] = H[\mathcal{S}_R|X_0^R] + H[X_0^R]\,.$$

Since the process is Markov, $H[\mathcal{S}_R|X_0^R] = 0$ and thus:

$$H[X_0^R] = H[\mathcal{S}_R] \quad\iff\quad H[X_0^R|\mathcal{S}_R] = 0\,.$$

In words, spin chains are processes for which there exists a one-to-one correspondence between the R-blocks and the causal states, confirming the interpretation specified in Ref. [**?**].

The above equations also show that spin chains have $\chi(R) = Rh_\mu$. Here we provide another proof:

Proposition 17.

$$H[X_0^R|\mathcal{S}_R] = 0 \quad\iff\quad \chi(R) = R\, h_\mu\,, \tag{A.10}$$

where h_μ is the process's entropy rate.

Proof. The proof is a direct calculation:

$$\begin{aligned}\chi(R) &= H[\mathcal{S}_0|X_0^R, \mathcal{S}_R]\\ &= H[\mathcal{S}_0, X_0^R] - H[X_0^R, \mathcal{S}_R]\\ &= H[\mathcal{S}_0, X_0^R] - H[X_0^R|\mathcal{S}_R] - H[\mathcal{S}_R]\\ &= H[\mathcal{S}_0, X_0^R] - H[X_0^R|\mathcal{S}_R] - H[\mathcal{S}_0]\\ &= H[X_0^R|\mathcal{S}_0] - H[X_0^R|\mathcal{S}_R]\\ &= Rh_\mu - H[X_0^R|\mathcal{S}_R]\ .\end{aligned}$$

□

Proposition 18. *Periodic processes are* 0*-cryptic.*

Proof. Periodic processes are order-R Markov spin chains, so $\mathbf{E} = C_\mu - Rh_\mu$. Since $h_\mu = 0$, $\mathbf{E} = C_\mu$. By Cor. 13 the process is 0-cryptic. □

Proposition 19. *An order-R spin chain with positive entropy rate is not* $(R-1)$*-cryptic.*

Proof. Assume that the order-R Markov spin chain *is* $(R-1)$-cryptic.

For $R \geq 1$, if the process is $(R-1)$-cryptic, then by Cor. 6 $\chi(R-1) = \chi$. Combining this with the above Prop. 17, we have $\chi(R-1) = (R-1)h_\mu - H[X_0^{R-1}|\mathcal{S}_{R-1}]$. If it is an order-$R$ Markov spin chain, then we also have from Eqn. A.4 that $\chi = Rh_\mu$. Combining this with the previous equation, we find that $H[X_0^{R-1}|\mathcal{S}_{R-1}] = -h_\mu$. By positivity of conditional entropies, we have reached a contradiction. Therefore an order-R Markov spin chain must not be $(R-1)$-cryptic.

For $R = 0$, the proof also holds since negative cryptic orders are not defined. □

Proposition 20. *An order-R spin chain with positive entropy rate is not k-cryptic for any* $0 \leq k < R$.

Proof. By Lem. 2, if the process where k-cryptic for some $0 \leq k < R$, then it would also be $(R-1)$-cryptic. By Prop. 19, this is not true. Therefore, the primitive orders of Markovity and crypticity are the same. □

§A.3 Examples

It is helpful to see crypticity in action. We now turn to a number of examples to illustrate how various orders of crypticity manifest themselves in ϵ-machine structure and what kinds of processes are cryptic and so hide internal state information from an observer. For details (transition matrices, notation, and the like) not included in the following and for complementary discussions and analyses of them, see Refs. [**?**, **?**, **?**].

We start at the bottom of the crypticity hierarchy with a 0-cryptic process and then show examples of 1-cryptic and 2-cryptic processes. Continuing up the hierarchy, we generalize and give a parametrized family of processes that are k-cryptic. Finally, we demonstrate an example that is ∞-cryptic.

It should be pointed out, though, that these examples were hand-chosen to illustrate some of the range of possible processes in terms of cryptic and Markov orders. If one were to encounter a process in the wild, its cryptic order would not be known and the calculation of crypticity would require that one determines the cryptic order. One can estimate the cryptic order by calculating the cryptic approximation until it appears to have converged or computational power has run out. Alternatively, one might deduce the order exactly via some other technique, as we do in the upcoming examples. Of course, we wish to note that Ref. [**?**] demonstrates how to calculate χ without any knowledge of the cryptic order.

§A.3.1 Even Process: 0-Cryptic

Figure A.1 gives the ϵ-machine for the Even Process. The Even Process produces binary sequences in which all blocks of uninterrupted 1s are even in length, bounded by 0s. Further, after each even length is reached, there is a probability p of breaking the block of 1s by inserting one or more 0s.

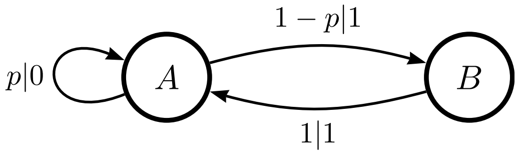

Figure A.1: A 0-cryptic process: Even Process. The transitions denote the probability p of generating symbol x as $p|x$.

Reference [**?**] showed that the Even Process is 0-cryptic with a statistical complexity of $C_\mu = H\left(1/(2-p)\right)$, an entropy rate of $h_\mu = H(p)/(2-p)$, and crypticity of $\chi = 0$. Note that $H(p)$ is the binary entropy function. If $p = \frac{1}{2}$, then $\mathbf{E} = C_\mu = \log_2(3) - \frac{2}{3}$ bits. (As Ref. [**?**] notes, these closed-form expressions for C_μ and $\mathbf{E}$ have been known for some time.)

To see why the Even Process is 0-cryptic, first note that the semi-infinite string $\overrightarrow{X}_0 = 1, 1, 1\ldots$ occurs with probability zero. So with probability one, a given future will have only a finite number of 1s before a 0 is seen. Once the 0 is seen, it is straightforward to count the number of 1s preceding it. If the number of 1s is even, then $\mathcal{S}_0$, the causal state that preceded this future, is A. Otherwise, it is B. In either case, we know the causal state with certainty, and so, $H[\mathcal{S}_0|\overrightarrow{X}_0] = 0$.

It is important to note that this process is *not* order-R Markov for any finite R [**?**]. Nonetheless, our new expression for $\mathbf{E}$ is valid. This shows the broadening of our ability to calculate $\mathbf{E}$ even for low complexity processes that are, in effect, infinite-order Markov.

§A.3.2 Golden Mean Process: 1-Cryptic

Figure A.2 shows the ϵ-machine for the Golden Mean Process [**?**]. The Golden Mean Process is one in which no two 0s occur consecutively. After each 1, there is a probability p of generating a 0. As sequence length grows, the ratio of the number of allowed words of length L to the number of allowed words at length $L-1$ approaches the golden ratio; hence, its name. The Golden Mean Process ϵ-machine looks remarkably similar to that for the Even Process. The informational analysis, however, shows that they have markedly different properties.

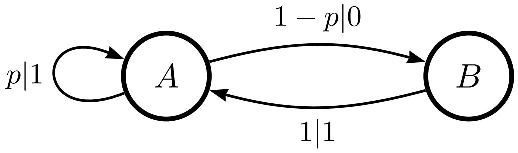

Figure A.2: A 1-cryptic process: Golden Mean Process.

Reference [**?**] showed that the Golden Mean Process has the same statistical complexity and entropy rate as the Even Process: $C_\mu = H\left(1/(2-p)\right)$ and $h_\mu = H(p)/(2-p)$. However, the crypticity is not zero (for $0 < p < 1$). From Cor. 6 we calculate:

$$\begin{aligned} \chi &= \chi(1) \\ &= H[\mathcal{S}_0 | X_0^1, \mathcal{S}_1] \\ &= H[\mathcal{S}_0 | X_0^1] \\ &= \Pr(0) H[\mathcal{S}_0 | X_0 = 0] + \Pr(1) H[\mathcal{S}_0 | X_0 = 1] \\ &= H(p)/(2-p) \, . \end{aligned}$$

If $p = \frac{1}{2}$, $C_\mu = \log_2(3) - \frac{2}{3}$ bits, excess entropy $\mathbf{E} = \log_2(3) - \frac{4}{3}$ bits, and crypticity $\chi = \frac{2}{3}$ bits. Thus, the excess entropy differs from that of the Even Process. (As with the Even Process, these closed-form expressions for C_μ and $\mathbf{E}$ have been known for some time.)

The Golden Mean Process is 1-cryptic. To see why, it is enough to note that it is order-1 Markov. By Prop. 14, it is 1-cryptic. We know it is not 0-cryptic since any future beginning with 1 could have originated in either state A or B. In addition, the spin-block expression for excess entropy of Ref. [**?**], Eqn. A.4 here, applies for an $R = 1$ Markov chain.

§A.3.3 Butterfly Process: 2-Cryptic

The next example, the Butterfly Process of Fig. B.1, illustrates, in a more explicit way than possible with the previous processes, the role that crypticity plays and how it can be understood in terms of an ϵ-machine's structure. Most of the explanation does not require calculating much, if anything.

It is first instructive to see why the Butterfly Process is *not* 1-cryptic.

If we can find a family $\{\overrightarrow{x}_0\}$ such that $H[\mathcal{S}_1 | \overrightarrow{X}_0 = \overrightarrow{x}_0] \neq 0$, then the total conditional entropy will be positive and, thus, the machine will not be 1-cryptic. To show that this can happen, consider the future $\overrightarrow{x}_0 = (0, 1, 2, 4, 4, 4, \ldots)$. It is clear that the state following 1 must be A. Thus, in order to generate 0 or 1 before arriving at A, the state pair $(\mathcal{S}_0, \mathcal{S}_1)$ can be either (B, C) or (D, E). This uncertainty in $\mathcal{S}_1$ is enough to break the criterion, and this occurs for the family of futures beginning with 01.

To see that the process is 2-cryptic, notice that the two paths (B, C) and (D, E) converge on A. Therefore, there is no uncertainty in $\mathcal{S}_2$ given this future. It is reasonably straightforward to see that indeed *any* two-symbol word (X_0, X_1) will lead to a unique causal state. This is because the Butterfly Process is a very limited version of an 8-symbol, order-2 Markov process.

Note that the transition matrix is doubly-stochastic and so the stationary distribution is uniform. The statistical complexity is rather direct in this case: $C_\mu = \log_2 5$. We now can calculate χ using Cor. 6:

$$\begin{aligned}
\chi &= \chi(2) \\
&= H[\mathcal{S}_0 | X_0^2, \mathcal{S}_2] \\
&= H[\mathcal{S}_0 | X_0^2] \\
&= \Pr(01) \cdot H[\mathcal{S}_0 | X_0^2 = 01] \\
&\quad + \Pr(12) \cdot H[\mathcal{S}_0 | X_0^2 = 12] \\
&\quad + \Pr(13) \cdot H[\mathcal{S}_0 | X_0^2 = 13] \\
&= \frac{1}{10} \cdot 1 + \frac{1}{10} \cdot 1 + \frac{1}{10} \cdot 1 \\
&= \frac{3}{10} \text{ bits.}
\end{aligned}$$

From Cor. 12, we get an excess entropy of

$$\begin{aligned}
\mathbf{E} &= C_\mu - \chi(2) \\
&= \log_2 5 - \frac{3}{10} \\
&\approx 2.0219 \text{ bits.}
\end{aligned}$$

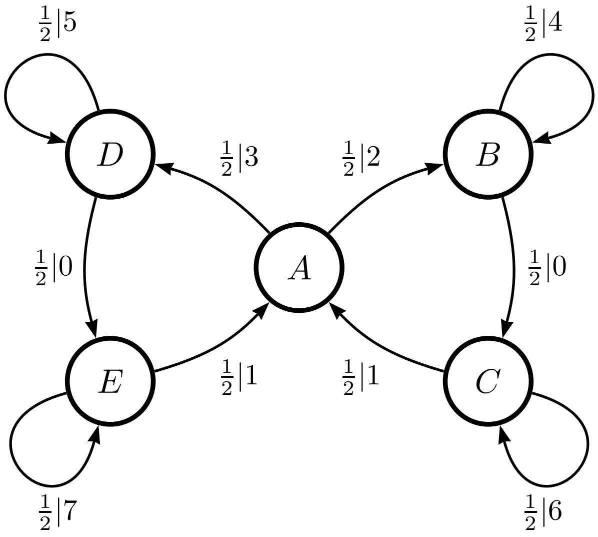

Figure A.3: A 2-cryptic process: Butterfly Process over a 6-symbol alphabet.

For comparison, if we had assumed the Butterfly Process was 1-cryptic, then we would have:

$$\begin{aligned}\mathbf{E} &= C_\mu - \chi(1)\\ &= C_\mu - (H[\mathcal{S}_0, X_0] - H[\mathcal{S}_1, X_0])\\ &\approx \log 2(5) - (3.3219 - 2.5062)\\ &= \log 2(5) - 0.8156 \approx 1.5063 \text{ bits.}\end{aligned}$$

We can see that this is substantially below the true value: a 25% error.

§A.3.4 Restricted Golden Mean: k-Cryptic

Now, we turn to illustrate a crypticity-parametrized family of processes, giving examples of k-cryptic processes for any k. We call this family the Restricted Golden Mean as its support is a restriction of the Golden Mean support. (See Fig. B.4 for its ϵ-machines.) The $k = 1$ member of the family is exactly the Golden Mean.

It is straightforward to see that this process is order-k Markov since each word of length k induces just one causal state. Proposition 14 then implies it is (at most) k-cryptic. In order to show that it is not $(k-1)$-cryptic, consider the case $\overrightarrow{x}_0 = 1^k 0^\infty$. The first $(k-1)$ 1s will induce a mixture over states k and 0. The following future $\overrightarrow{x}_k = 10^\infty$ is consistent with both states k and 0. Therefore, the $(k-1)$-crypticity criterion is not satisfied. Therefore, it is k-cryptic.

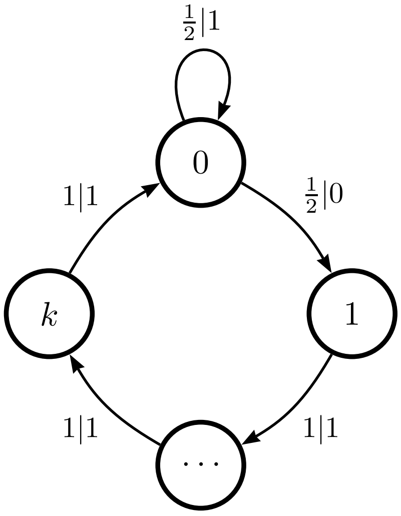

Figure A.4: k-cryptic processes: Restricted Golden Mean Family.

For arbitrary k, there are $k+1$ causal states and the stationary distribution is:

$$\pi = \left(\frac{2}{k+2}, \frac{1}{k+2}, \frac{1}{k+2}, \ldots, \frac{1}{k+2}\right) .$$

The statistical complexity is

$$C_\mu = \log_2(k+2) - \frac{2}{k+2} \, .$$

For the k-th member of the family, we have for the crypticity:

$$\chi = \chi(k) = \frac{2k}{k+2} \, .$$

And the excess entropy follows directly from Cor. 12:

$$\mathbf{E} = C_\mu - \chi = \log_2(k+2) - \frac{2(k+1)}{k+2} \, ,$$

which diverges with k. (Calculational details are found in Ref. [**?**].)

§A.3.5 Stretched Golden Mean

The Stretched Golden Mean is a family of processes that does not occupy the same support as the Golden Mean. Instead of requiring that blocks of 0s are of length 1, we require that they are of length k. The ϵ-machine for this process is shown in Fig. A.5.

Again, it is straightforward to see that this process is order-k Markov. To see that it is not 0-cryptic, note that:

$$\begin{aligned} H[\mathcal{S}_0|\overrightarrow{X}_0] &= H[\mathcal{S}_0|X_0 = 0, \overrightarrow{X}_1] + H[\mathcal{S}_0|X_0 = 1, \overrightarrow{X}_1] \\ &\geq H[\mathcal{S}_0|X_0 = 1, \overrightarrow{X}_1] \\ &= \frac{2}{k+2} \sum_{\overrightarrow{x}_1} H[\mathcal{S}_0|X_0 = 1, \overrightarrow{X}_1 = \overrightarrow{x}_1] \\ &\geq \frac{2}{k+2} H[\mathcal{S}_0|\overrightarrow{X}_1 = 1^\infty] \\ &= \frac{2}{k+2} \\ &> 0 \, . \end{aligned}$$

To see that this family is 1-cryptic, first note that if $X_0 = 1$, then $\mathcal{S}_1 = 0$. Next, consider the case when $X_0 = 0$. If the future $\overrightarrow{x}_1 = 1^\infty$, then $\mathcal{S}_1 = k$. Similarly, if the future $\overrightarrow{x}_1 = 0^n 1^\infty$, then $\mathcal{S}_1 = k - n$.

This family provides an example for which the cryptic order is strictly less than the Markov order. In this case, the cryptic order is fixed at 1 for all k, while the Markov order is k. Note that the separation between the Markov and cryptic order can grow arbitrarily large and, thus, the two properties are clearly not redundant.

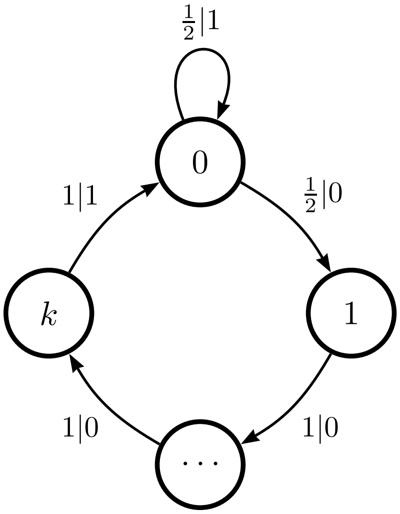

Figure A.5: k-cryptic processes: Stretched Golden Mean Family.

The stationary distribution is the same as for the Restricted Golden Mean and so, then, is the statistical complexity. In addition, we have:

$$\begin{aligned}\chi &= \chi(1)\\ &= H[\mathcal{S}_0|X_0, \mathcal{S}_1]\\ &= h_\mu \,.\end{aligned}$$

Consequently,

$$\mathbf{E} = C_\mu - \chi = C_\mu - h_\mu \,.$$

§A.3.6 Nemo Process: ∞-Cryptic

We close our cryptic process bestiary with a (very) finite-state process that has infinite cryptic order: The three-state Nemo Process. Over no finite-length sequence will all of the internal state information be present in the observations. The Nemo Process ϵ-machine is shown in Fig. B.7.

Its stationary state distribution is

$$\Pr(\mathcal{S}) \equiv \pi = \frac{1}{3-2p} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} 1 & 1-p & 1-p \end{pmatrix} \end{array},$$

from which one calculates the statistical complexity:

$$C_\mu = \log_2(3-2p) - \frac{2(1-p)}{3-2p}\log_2(1-p) \,.$$

The Nemo Process is not a finite-cryptic process. That is, there exists no finite k for which $H[\mathcal{S}_k|\overrightarrow{X}_0] = 0$. To show this, we must demonstrate that there exists a family of futures such

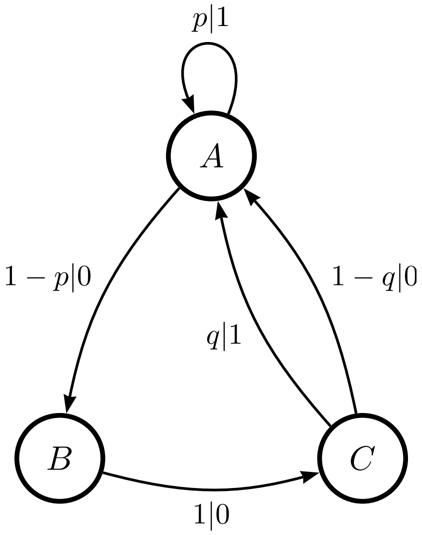

Figure A.6: The ∞-cryptic Nemo Process.

that for each future $H[\mathcal{S}_k|\overrightarrow{X}_0 = \overrightarrow{x}] > 0$. The family of futures we use begins with all 0s and then has a 1. Intuitively, the 1 is chosen because it is a synchronizing word for the process—after observing a 1, the ϵ-machine is always in state A. Then, causal shielding will decouple the infinite future from the first few symbols, thereby allowing us to compute the conditional entropies for the entire family of futures.

First, recall the shorthand:

$$\Pr(\mathcal{S}_k|\overrightarrow{X}_0) = \lim_{L\to\infty} \Pr(\mathcal{S}_k|X_0^L) .$$

Without loss of generality, assume $k < L$. Then,

$$\begin{aligned}\Pr(\mathcal{S}_k|X_0^L) &= \frac{\Pr(X_0^k, \mathcal{S}_k, X_k^L)}{\Pr(X_0^L)} \\ &= \frac{\Pr(X_k^L|X_0^k, \mathcal{S}_k)\Pr(X_0^k, \mathcal{S}_k)}{\Pr(X_0^L)} \\ &= \frac{\Pr(X_k^L|\mathcal{S}_k)\Pr(X_0^k, \mathcal{S}_k)}{\Pr(X_0^L)} ,\end{aligned}$$

where the last step is possible since the causal states are Markovian [**?**], shielding the past from the future. Each of these quantities is given by:

$$\Pr(X_k^L = w|\mathcal{S}_k = \sigma) = [T^{(w)}\mathbf{1}]_\sigma$$

$$\Pr(X_0^k = w, \mathcal{S}_k = \sigma) = [\pi T^{(w)}]_\sigma$$

$$\Pr(X_0^L = w) = \pi T^{(w)}\mathbf{1} .$$

where $T^{(w)} \equiv T^{(x_0)}T^{(x_1)}\cdots T^{(x_{L-1})}$, $\mathbf{1}$ is a column vector of 1s, and $T^{(x)}_{\sigma\sigma'} = \Pr(\mathcal{S}' = \sigma', X = x|\mathcal{S} =$

σ). To establish $H[\mathcal{S}_k \,|\, \overrightarrow{X}_0] > 0$ for any k, we rely on using values of k that are multiples of three. So, we concentrate on the following for $n = 0, 1, 2, \ldots$:

$$H[\mathcal{S}_{3n} | X_0^{3n+1} = 0^{3n}1, \overrightarrow{X}_{3n+1}] > 0 \,.$$

Since 1 is a synchronizing word, we can greatly simplify the conditional probability distribution. First, we freely include the synchronized causal state A and rewrite the conditional distribution as a fraction:

$$\begin{aligned}
&\Pr(\mathcal{S}_{3n} | X_0^{3n+1} = 0^{3n}1, \overrightarrow{X}_{3n+1}) \\
&= \Pr(\mathcal{S}_{3n} | X_0^{3n+1} = 0^{3n}1, \mathcal{S}_{3n+1} = A, \overrightarrow{X}_{3n+1}) \\
&= \frac{\Pr(\mathcal{S}_{3n}, X_0^{3n+1} = 0^{3n}1, \mathcal{S}_{3n+1} = A, \overrightarrow{X}_{3n+1})}{\Pr(X_0^{3n+1} = 0^{3n}1, \mathcal{S}_{3n+1} = A, \overrightarrow{X}_{3n+1})} \,.
\end{aligned}$$

Then, we factor everything except $\overrightarrow{X}_{3n+1}$ out of the numerator and make use of causal shielding to simplify the conditional. For example, the numerator becomes:

$$\begin{aligned}
&\Pr(\mathcal{S}_{3n}, X_0^{3n+1} = 0^{3n}1, \mathcal{S}_{3n+1} = A, \overrightarrow{X}_{3n+1}) \\
&= \Pr(\overrightarrow{X}_{3n+1} | \mathcal{S}_{3n}, X_0^{3n+1} = 0^{3n}1, \mathcal{S}_{3n+1} = A) \\
&\qquad \times \Pr(\mathcal{S}_{3n}, X_0^{3n+1} = 0^{3n}1, \mathcal{S}_{3n+1} = A) \\
&= \Pr(\overrightarrow{X}_{3n+1} | \mathcal{S}_{3n+1} = A) \\
&\qquad \times \Pr(\mathcal{S}_{3n}, X_0^{3n+1} = 0^{3n}1, \mathcal{S}_{3n+1} = A) \\
&= \Pr(\overrightarrow{X}_{3n+1} | \mathcal{S}_{3n+1} = A) \Pr(\mathcal{S}_{3n}, X_0^{3n+1} = 0^{3n}1) \,.
\end{aligned}$$

Similarly, the denominator becomes:

$$\begin{aligned}
&\Pr(X_0^{3n+1} = 0^{3n}1, \mathcal{S}_{3n+1} = A, \overrightarrow{X}_{3n+1}) \\
&= \Pr(\overrightarrow{X}_{3n+1} | \mathcal{S}_{3n+1} = A) \Pr(X_0^{3n+1} = 0^{3n}1) \,.
\end{aligned}$$

Combining these results, we obtain a finite form for the entropy of $\mathcal{S}_{3n}$ conditioned on a family of infinite futures, first noting:

$$\Pr(\mathcal{S}_{3n} | X_0^{3n+1} = 0^{3n}1, \overrightarrow{X}_{3n+1}) = \Pr(\mathcal{S}_{3n} | X_0^{3n+1} = 0^{3n}1) \,.$$

Thus, for all $\overrightarrow{x}_{3n+1}$, we have:

$$\begin{aligned}
&H[\mathcal{S}_{3n} | X_0^{3n+1} = 0^{3n}1, \overrightarrow{X}_{3n+1} = \overrightarrow{x}_{3n+1}] \\
&= H[\mathcal{S}_{3n} | X_0^{3n+1} = 0^{3n}1] \,.
\end{aligned}$$

Now, we are ready to compute the conditional entropy for the entire family. First, note that $T^{(0)}$ raised to the third power is a diagonal matrix with each element equal to $(1-p)(1-q)$.

Thus, for $j = 1, 2, 3\ldots$:

$$\left[T^{(0)}\right]^{3j}_{\sigma\sigma} = (1-p)^j(1-q)^j \ .$$

Using all of the above relations, we can easily calculate:

$$\Pr(\mathcal{S}_{3n}|X_0^{3n+1} = 0^{3n}1) = \frac{1}{3-2p} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} p & 0 & q(1-p) \end{pmatrix} \end{array} .$$

Thus, for $p, q \in (0,1)$, we have:

$$\begin{aligned}
& H[\mathcal{S}_{3n}|\overrightarrow{X}_0] \\
& \geq H[\mathcal{S}_{3n}|X_0^{3n+1} = 0^{3n}1, \overrightarrow{X}_{3n+1}] \\
& = \sum_{\overrightarrow{x}_{3n+1}} \Pr\left(X_0^{3n+1} = 0^{3n}1, \overrightarrow{X}_{3n+1} = \overrightarrow{x}_{3n+1}\right) \\
& \quad \times H[\mathcal{S}_{3n}|X_0^{3n+1} = 0^{3n}1, \overrightarrow{X}_{3n+1} = \overrightarrow{x}_{3n+1}] \\
& = H[\mathcal{S}_{3n}|X_0^{3n+1} = 0^{3n}1] \\
& \quad \times \sum_{\overrightarrow{x}_{3n+1}} \Pr\left(X_0^{3n+1} = 0^{3n}1, \overrightarrow{X}_{3n+1} = \overrightarrow{x}_{3n+1}\right) \\
& = H[\mathcal{S}_{3n}|X_0^{3n+1} = 0^{3n}1] \Pr(X_0^{3n+1} = 0^{3n}1) \\
& = \left(\frac{p}{3-2p}\log_2\frac{3-2p}{p} + \frac{q(1-p)}{3-2p}\log_2\frac{3-2p}{q(1-p)}\right) \\
& \quad \times [(1-p)(1-q)]^{3n} \\
& > 0 \, .
\end{aligned}$$

So, any time k is a multiple of three, $H[S_k|\overrightarrow{X}_0] > 0$. Finally, suppose $(k \mod 3) = i$, where $i \neq 0$. That is, suppose k is not a multiple of three. By Lem. 1, $H[\mathcal{S}_k|\overrightarrow{X}_0] \geq H[\mathcal{S}_{k+i}|\overrightarrow{X}_0]$ and, since we just showed that the latter quantity is always strictly greater than zero, we conclude that $H[\mathcal{S}_k|\overrightarrow{X}_0] > 0$ for every value of k.

The above establishes that the Nemo Process does not satisfy the k-crypticity criterion for any finite k. Thus, the Nemo process is ∞-cryptic. This means that we cannot make use of the k-cryptic approximation to calculate χ or $\mathbf{E}$.

Fortunately, the techniques introduced in Refs. [**?**] and [**?**] do not rely on an approximation method. To avoid ambiguity, denote the statistical complexity we just computed as C_μ^+. When those techniques are applied to the Nemo Process, we find that the process is causally reversible

($C_\mu^+ = C_\mu^-$) and has the following forward-reverse causal-state conditional distribution:

$$\Pr(\mathcal{S}^+|\mathcal{S}^-) = \frac{1}{p+q-pq} \begin{array}{c} \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} p & 0 & q(1-p) \\ 0 & q & p(1-q) \\ q & p(1-q) & 0 \end{pmatrix} \end{array} .$$

With this, one can calculate **E**, in closed-form, via:

$$\mathbf{E} = C_\mu^+ - H[\mathcal{S}^+|\mathcal{S}^-] .$$

(Again, calculational details are provided in Ref. [**?**].)

§A.4 Conclusion

Calculating the excess entropy $I[\overleftarrow{X}; \overrightarrow{X}]$ is, at first blush, a daunting task. We are asking for a mutual information between two infinite sets of random variables. Appealing to $\mathbf{E} = I[\mathcal{S}; \overrightarrow{X}]$, we use the compact representation of the ϵ-machine to reduce one infinite set (the past) to a (usually) finite set. A process's k-crypticity captures something similar about the infinite set of future variables and allows us to further compact our form for excess entropy, reducing an infinite variable set to a finite one. The resulting stratification of process space is a novel way of thinking about its *structure* and, as long as we know in which stratum we lie, we can rapidly calculate many quantities of interest.

Unfortunately, in the general case, one will not know a priori a process's cryptic order. Worse, as far as we are aware, there is no known finite method for calculating the cryptic order. This strikes us as an interesting open problem and challenge.

If, by construction or by some other means, one does know it, then, as we showed, crypticity and **E** can be calculated using the crypticity expansion. Failing this, though, one might consider using the expansion to search for the order. There is no known stopping criterion, so this search may not find k in finite time. Moreover, the expansion is a calculation that grows exponentially in computational complexity with cryptic order, as we noted. Devising a stopping criterion would be very useful to such a search.

Even without knowing the k-crypticity, the expansion is often still useful. For use in estimating **E**, it provides us with a bound from above. This is complementary to the lower bound one finds using the typical expansion $\mathbf{E}(L) = H[X_0^L] - h_\mu L$ [**?**]. Using these upper and lower bounds, one may determine that for a given purpose, the estimate of χ or $\mathbf{E}$ is within an acceptable tolerance.

The crypticity hierarchy is a revealing way to carve the space of processes in that it concerns how they hide internal state information from an observer. The examples were chosen to illustrate several features of this new view. The Even Process, a canonical example of order-∞ Markov, resides instead at the very bottom of this ladder. The two example families show us how k-cryptic is neither a parallel nor independent concept to order-R Markov. Finally, we see in the last example an apparently simple process with ∞-crypticity.

The general lesson is that internal state information need not be immediately available in measurement values, but instead may be spread over long measurement sequences. If a process is k-cryptic and k is finite, then internal state information is accessible over sequences of length k. The existence, as we demonstrated, of processes that are ∞-cryptic is rather sobering. Interpreted as a statement of the impossibility of extracting state information, it reminds us of earlier work on hidden spatial dynamical systems that exhibit a similar encrypting of internal structure in observed spacetime patterns [**?**].

Due to the exponentially growing computational effort to search for the cryptic order and, concretely, the existence of ∞-cryptic processes, the general theory introduced in Ref. [**?**] and Ref. [**?**] is seen to be necessary. It allows one to directly calculate $\mathbf{E}$ and crypticity and to do so efficiently.

APPENDIX B

IACPLCOCS

§B.1 Introduction

We introduced a new system “invariant”—the *crypticity* χ—for stationary hidden stochastic processes to capture how much internal state information is directly accessible (or not) from observations [**?**, **?**, **?**]. Two approaches to calculate χ were given. The first, reported in Ref. [**?**] and Ref. [**?**], used the so-called *mixed-state* method, which employs linear combinations of a process's forward-time ϵ-machine. The second, appearing in Ref. [**?**], developed a systematic expansion $\chi(k)$ as a function of the length k of observed sequences over which internal state information can be extracted. The mixed-state method is the most efficient way to calculate crypticity and other important system properties, such as the excess entropy $\mathbf{E}$, since it avoids having to write out all of the terms required for calculating $\chi(k)$. It also does not rely on knowing in advance a process's cryptic order.

As such, we reported results in Ref. [**?**] that use the mixed-state method to, in a sense, calibrate the $\chi(k)$ expansion and to understand its convergence.

Here we provide the calculational details behind those results. Generally, though, the goal is to find out what a stochastic process looks like when scanned in the “opposite” time direction. Specifically, starting with a given ϵ-machine M of a process, calculate its reverse-time representation M^-. (The latter is not always minimal and so not, in that case, an ϵ-machine.) This is done in two steps: (i) time-reverse M, producing $\widehat{M} = \mathcal{T}(M)$, and (ii) convert $\widehat{M}$ to a unifilar presentation $\mathcal{U}(\widehat{M})$ using mixed states, which are linear combinations of the states of $\widehat{M}$.

In the following, we show how to implement these steps for the various example processes presented in Ref. [**?**]: the Butterfly, Restricted Golden Mean, and Nemo Processes. We jump

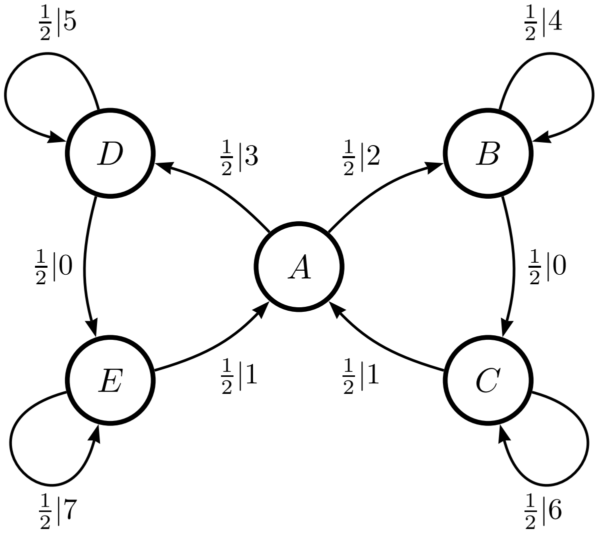

Figure B.1: A 2-cryptic process: The ϵ-machine representation of the Butterfly Process. Edge labels $t|x$ give the probability $t = T^{(x)}_{\sigma\sigma'}$ of making a transition and from causal state σ to causal state σ' and seeing symbol x.

directly into the calculations, assuming the reader is familiar with Refs. [**?**], [**?**], and [**?**]. Those references provide, in addition, more discussion and motivation and reasonable list of citations.

§B.2 Butterfly Process

Figure B.1 shows the ϵ-machine for Ref. [**?**]'s Butterfly process—an output process over eight symbols $\mathcal{A} = \{0, 1, \ldots, 7\}$.

Since its transition matrices are doubly stochastic, the stationary state distribution is uniform. This immediately gives its stored information: the statistical complexity is $C_\mu = \log_2(5)$ bits. It also makes the construction of the time-reverse machine straightforward: We simply reverse the directions of all the arrows. (See Fig. B.2.) Note that the time-reverse presentation is no longer unifilar and, therefore, it is not the reversed process's ϵ-machine.

Due to this we must calculate the mixed-state presentation to find a unifilar presentation. The calculated mixed states and the words which induce them are given in Table B.1.

The result is the reverse ϵ-machine shown in Fig. B.3. Note that it has two more states than the original (forward) ϵ-machine of Fig. B.1.

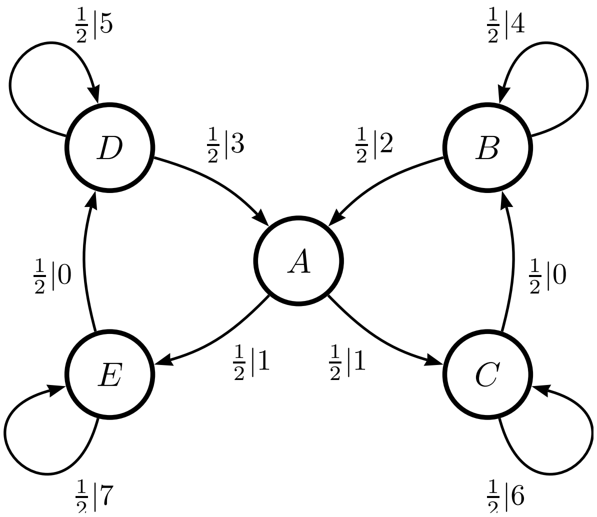

Figure B.2: Time-reversed Butterfly Process.

The stationary distribution of this reversed machine is $\pi = (0.1, 0.2, 0.2, 0.15, 0.15, 0.1, 0.1)$. Now we are in position to calculate **E** using the result of Ref. [**?**]:

$$\mathbf{E} = C_\mu - \chi \quad \text{(B.1)}$$

$$\mathbf{E} = C_\mu - H[\mathcal{S}^+|\overrightarrow{X}] \quad \text{(B.2)}$$

$$= C_\mu - H[\mathcal{S}^+|\mathcal{S}^- = \epsilon^+(\overrightarrow{X})] \,. \quad \text{(B.3)}$$

In this case, we find a crypticity of:

$$\begin{aligned}\chi &= H[\mathcal{S}^+|\mathcal{S}^-] \\ &= 0.1H[(0, \frac{1}{2}, 0, \frac{1}{2}, 0)] + 0.2H[(0, 0, \frac{1}{2}, 0, \frac{1}{2})] \\ &\quad + 0.2H[(1, 0, 0, 0, 0)] + 0.15H[(0, 1, 0, 0, 0)] \\ &\quad + 0.15H[(0, 0, 0, 1, 0)] + 0.1H[(0, 0, 1, 0, 0)] \\ &\quad + 0.1H[(0, 0, 0, 0, 1)] \\ &= 0.1 + 0.2 \\ &= 0.3 \text{ bits.}\end{aligned}$$

So, $\mathbf{E} = \log_2(5) - 0.3 \approx 2.0219$ bits, in accord with the result calculated via Thm. 1 of Ref. [**?**].

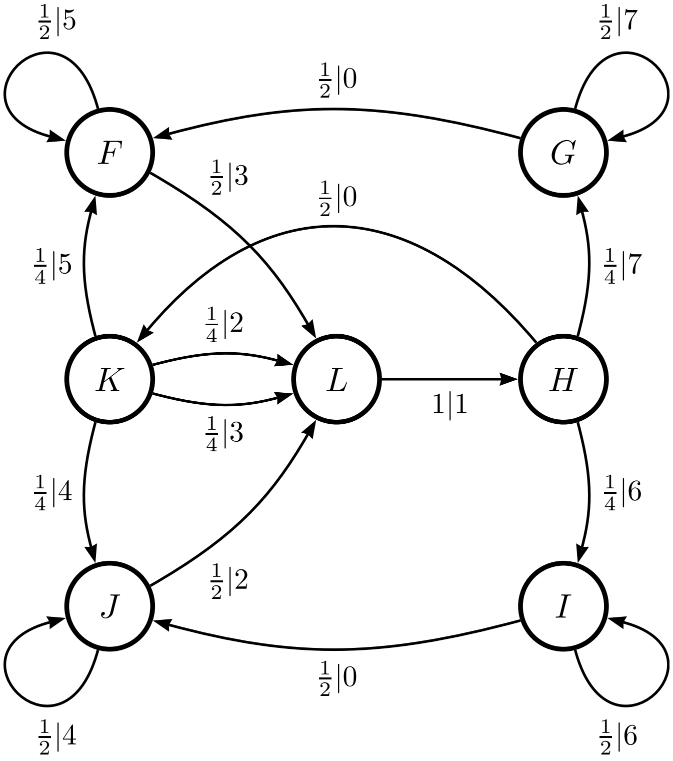

Figure B.3: Reverse Butterfly Process.

§B.3 Restricted Golden Mean Process

For reference, we give the family of labeled transition matrices for the binary Restricted Golden Mean Process (RGMP):

$$T^{(0)} = \begin{pmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \\ 0 & 0 & 0 & 0 & 0 & \cdots \end{pmatrix}$$

and

$$T^{(1)} = \begin{pmatrix} \frac{1}{2} & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 1 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \\ 1 & 0 & 0 & 0 & 0 & \cdots \end{pmatrix} .$$

Its ϵ-machine is given in Fig. B.4 and its stationary distribution is:

$$\pi = \left(\frac{2}{k+2}, \frac{1}{k+2}, \frac{1}{k+2}, \ldots, \frac{1}{k+2} \right) .$$

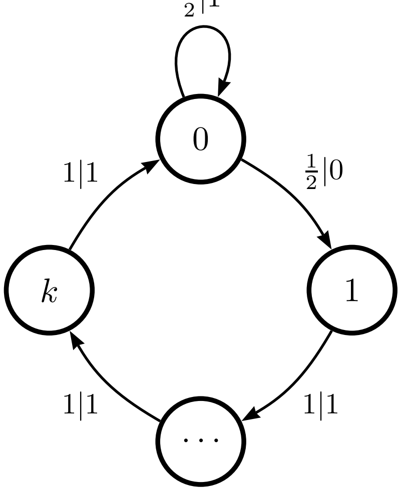

Figure B.4: The ϵ-machine for the Restricted Golden Mean Process.

Through other methods, we can show that the RGMP is reversible. We "push" RGMP to an edge machine presentation and "pull" $\mathcal{T}$(RGMP) also the same type of presentation. (An edge machine presentation of a machine M has states that are the edges of M.) These machines are the same. Therefore, the forward and reverse ϵ-machines are the same and, moreover, we can use the same mixed-state inducing word list. It is easy to see that one such list is $(0, 01, 011, \ldots, 01^k)$. Table B.2 gives the mixed states for these allowed words. It is also reason-

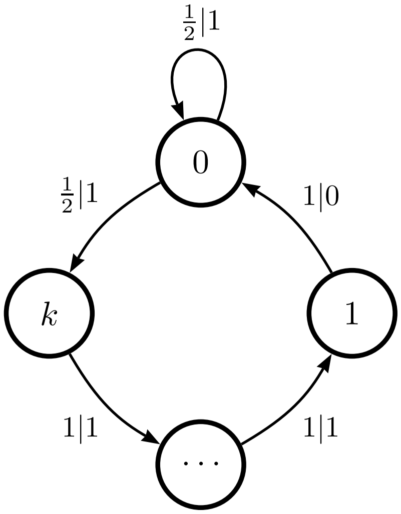

Figure B.5: Time-reversed presentation of the Restricted Golden Mean Process.

ably clear from the above mixed-state presentation that these correspond to the recurrent causal states for the time-reversed process's ϵ-machine.

With this, we can now compute χ using $H[\mathcal{S}^+|\mathcal{S}^-]$, as follows:

$$H[\mathcal{S}^+|\mathcal{S}^- = 0] = H[(1, 0^k)] = 0 \text{ and}$$
$$H[\mathcal{S}^+|\mathcal{S}^- = 0(1)^n] = H[(\frac{1}{2^n}, 0^{k-n}, \frac{1}{2^1}\frac{1}{2^2}\frac{1}{2^3}, \ldots, \frac{1}{2^n})] \, .$$

So that, in general, we have:

$$\begin{aligned} H[\mathcal{S}^+|\mathcal{S}^-] = \sum_{n=1}^{k-1} \frac{1}{k+2} H[(\frac{1}{2^n}, 0^{k-n}, \frac{1}{2^1}\frac{1}{2^2}\frac{1}{2^3}, \ldots, \frac{1}{2^n})] \\ + \frac{2}{2+k} H[(\frac{1}{2^k}, \frac{1}{2^1}\frac{1}{2^2}\frac{1}{2^3}, \ldots, \frac{1}{2^k})] \, . \end{aligned}$$

It can then be shown that:

$$\begin{aligned} & H[(\frac{1}{2^n}, 0^{k-n}, \frac{1}{2^1}\frac{1}{2^2}\frac{1}{2^3}, \ldots, \frac{1}{2^n})] \\ & \quad = H[(\frac{1}{2^n}, \frac{1}{2^1}\frac{1}{2^2}\frac{1}{2^3}, \ldots, \frac{1}{2^n})] \\ & \quad = 2 - 2^{(1-n)} \, . \end{aligned}$$

Therefore, returning to the causal-state-conditional entropy of interest, we have:

$$\begin{aligned} H[\mathcal{S}^+|\mathcal{S}^-] &= \frac{1}{k+2} \sum_{n=1}^{k-1} (2 - 2^{(1-n)}) + \frac{2}{2+k}(2 - 2^{(1-k)}) \\ &= \frac{1}{k+2}(2(k-1) + 2(2 - 2^{1-k}) - (2 - 2^{2-k})) \\ &= \frac{2k}{k+2} \, . \end{aligned}$$

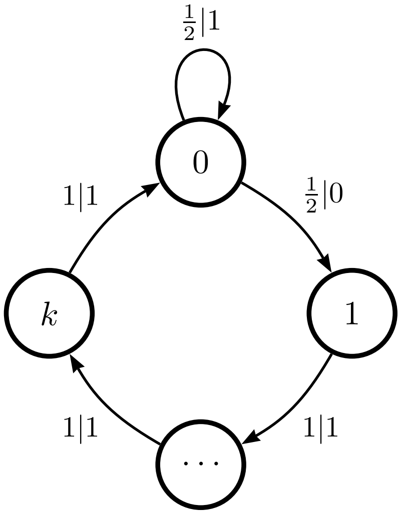

Figure B.6: Reverse Restricted Golden Mean Process.

With a few more steps, we arrive at our destination—the RGMP's informational quantities:

$$
\begin{aligned}
C_\mu &= \log 2(k+2) - \frac{2}{k+2} \, , \\
\chi &= \frac{2k}{k+2}, \text{ and} \\
\mathbf{E} &= \log 2(k+2) - \frac{2(k+1)}{k+2} \, .
\end{aligned}
$$

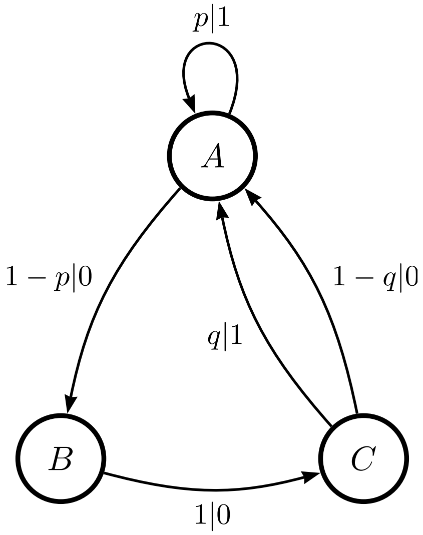

Figure B.7: The ϵ-machine for the ∞-cryptic Nemo Process.

§B.4 Nemo Process

We now demonstrate how to calculate χ and **E** for Ref. [**?**]'s ∞-cryptic process—the Nemo Process—using mixed-state methods. As emphasized in Ref. [**?**], the k-cryptic expansion there cannot be applied in this case. Thus, the Nemo Process demonstrates that Refs. [**?**] and [**?**]'s mixed-state method is essential.

Figure B.7 shows M^+, the ϵ-machine for the forward-scanned Nemo Process. Its transition matrices are:

$$T^{(0)} = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} 0 & 1-p & 0 \\ 0 & 0 & 1 \\ 1-q & 0 & 0 \end{pmatrix} \end{array} \text{ and}$$

$$T^{(1)} = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} p & 0 & 0 \\ 0 & 0 & 0 \\ q & 0 & 0 \end{pmatrix} \end{array} .$$

The stationary state distribution is the normalized left-eigenvector of $T \equiv T^{(0)} + T^{(1)}$ and is given by:

$$\Pr(\mathcal{S}^+) \equiv \pi^+ = \frac{1}{3-2p} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} 1 & 1-p & 1-p \end{pmatrix} \end{array} .$$

Then, the statistical complexity is the Shannon entropy over these states:

$$\begin{aligned} C_\mu &= H[\mathcal{S}^+] \\ &= \log_2(3-2p) - \frac{2(1-p)}{3-2p}\log_2(1-p) \, . \end{aligned}$$

The next step is to construct the time-reversed presentation $\widetilde{M}^+ = \mathcal{T}(M^+)$, shown in Fig. B.8.

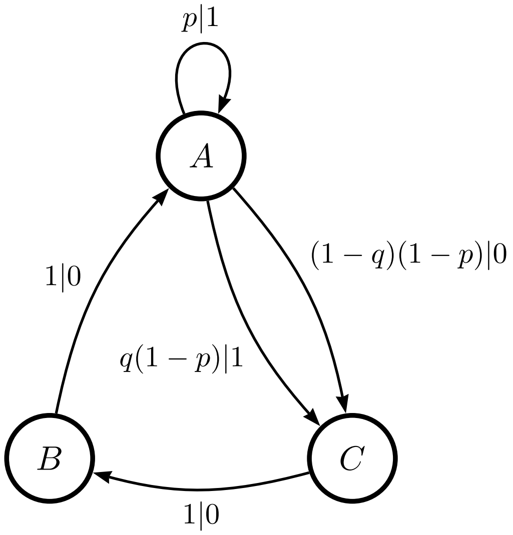

Figure B.8: The time-reversed presentation, $\widetilde{M}^{+} = \mathcal{T}(M^{+})$, of the Nemo Process.

The transition matrices of this machine are:

$$\widetilde{T}^{(0)} = \begin{array}{c c} & \begin{array}{ccc} A & B & C \end{array} \\ \begin{array}{c} A \\ B \\ C \end{array} & \begin{pmatrix} 0 & 0 & (1-q)(1-p) \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \end{array} \text{ and}$$

$$\widetilde{T}^{(1)} = \begin{array}{c c} & \begin{array}{ccc} A & B & C \end{array} \\ \begin{array}{c} A \\ B \\ C \end{array} & \begin{pmatrix} p & 0 & q(1-p) \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \end{array} .$$

Finally, we construct the mixed-state presentation of the time-reversed presentation, $\mathcal{U}(\widetilde{M}^{+})$,

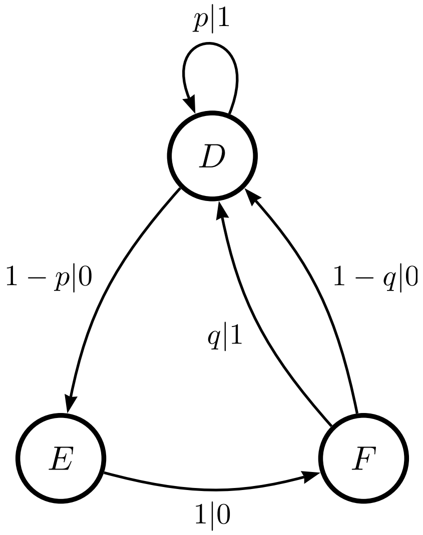

Figure B.9: The reverse ϵ-machine for the Nemo Process.

which is shown in Fig. B.9. On doing so, we obtain the following mixed states:

$$D \equiv \nu(1) = \frac{1}{p+q-pq} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} p & 0 & q(1-p) \end{pmatrix} \end{array} ,$$

$$E \equiv \nu(01) = \frac{1}{p+q-pq} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} 0 & q & p(1-q) \end{pmatrix} \end{array} , \text{ and}$$

$$F \equiv \nu(001) = \frac{1}{p+q-pq} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} q & p(1-q) & 0 \end{pmatrix} \end{array} .$$

These mixed states form the reverse ϵ-machine causal states, which are exactly the same as the forward ϵ-machine. Thus, the Nemo Process is causally reversible. The mixed states are distributions giving the probabilities of the forward causal states conditioned on a reverse causal state:

$$\Pr(\mathcal{S}^+|\mathcal{S}^-) = \frac{1}{p+q-pq} \begin{array}{c|ccc} & A & B & C \\ \hline D & p & 0 & q(1-p) \\ E & 0 & q & p(1-q) \\ F & q & p(1-q) & 0 \end{array} .$$

We use this to directly compute:

$$\begin{aligned} H[\mathcal{S}^+|\mathcal{S}^-] = \frac{1}{3-2p} & \left[\frac{p}{p+q-pq} \log_2\left(\frac{p+q-pq}{p}\right) \right. \\ & \left. + \frac{q(1-p)}{p+q-pq} \log_2\left(\frac{p+q-pq}{q(1-p)}\right) \right] \\ & + \frac{2(1-p)}{3-2p} \left[\frac{q}{p+q-pq} \log_2\left(\frac{p+q-pq}{q}\right) \right. \\ & \left. + \frac{p(1-q)}{p+q-pq} \log_2\left(\frac{p+q-pq}{p(1-q)}\right) \right] . \end{aligned}$$

Finally, we have:

$$\begin{aligned} \mathbf{E} & = C_\mu - H[\mathcal{S}^+|\mathcal{S}^-] \\ & = \log_2(3-2p) - \frac{2(1-p)}{3-2p} \log_2(1-p) \\ & \quad - \frac{1}{3-2p} \left[\frac{p}{p+q-pq} \log_2\left(\frac{p+q-pq}{p}\right) \right. \\ & \quad \left. + \frac{q(1-p)}{p+q-pq} \log_2\left(\frac{p+q-pq}{q(1-p)}\right) \right] \\ & \quad + \frac{2(1-p)}{3-2p} \left[\frac{q}{p+q-pq} \log_2\left(\frac{p+q-pq}{q}\right) \right. \\ & \quad \left. + \frac{p(1-q)}{p+q-pq} \log_2\left(\frac{p+q-pq}{p(1-q)}\right) \right] . \end{aligned}$$

§B.5 Conclusion

The detailed calculations make evident that Refs. [**?**] and [**?**]'s mixed-state method gives a new level of direct analysis for the informational properties of stationary stochastic processes, such as the crypticity and the excess entropy. The complementary approach given by the crypticity expansion $\chi(k)$ is useful in understanding information accessibility—how internal state information is spread over time in measurement sequences [**?**]. Nonetheless, while $\chi(k)$ can be calculated in particular finite cases, the mixed-state method is the most general and efficient method.

Allowed Words	μ or Previous Word
0	$(0,\frac{1}{2},0,\frac{1}{2},0)$
1	$(0,0,\frac{1}{2},0,\frac{1}{2})$
2	(1,0,0,0,0)
3	2
4	(0,1,0,0,0)
5	(0,0,0,1,0)
6	(0,0,1,0,0)
7	(0,0,0,0,1)
02	2
03	2
04	4
05	5
10	0
16	6
17	7
21	1
42	2
44	4
53	2
55	5
60	4
66	6
70	5
77	7

Table B.1: Calculating the time-reversed Butterfly Process's ϵ-machine via the forward ϵ-machine's mixed states. The 5-vector denotes the mixed-state distribution $\mu(w)$ reached after having seen the corresponding allowed word w. If the word leads to a unique state with probability one, we give instead the state's name.

Allowed Words	μ or Previous Word
0	$(1, 0^k)$
1	$(\frac{1}{k+1}, \frac{1}{k+1}, \ldots, \frac{1}{k+1})$
01	$(\frac{1}{2}, 0^{k-1}, \frac{1}{2})$
10	0
11	$\frac{1}{k}(\frac{1}{2}, 1, 1, \ldots, 1, \frac{1}{2})$
$\vdots$	$\vdots$
$0(1)^n$ for $1 \leq n \leq k$	$(\frac{1}{2^n}, 0^{k-n}, \frac{1}{2^1} \frac{1}{2^2} \frac{1}{2^3}, \ldots, \frac{1}{2^n})$
$1(1)^n$ for $1 \leq n \leq k$	$\frac{1}{k-n+1}(\frac{1}{2^n}, 1^{k-n}, \frac{1}{2^1} \frac{1}{2^2} \frac{1}{2^3}, \ldots, \frac{1}{2^n})$
$0(1)^k$	$(\frac{1}{2^k}, \frac{1}{2^1} \frac{1}{2^2} \frac{1}{2^3}, \ldots, \frac{1}{2^k})$
$1(1)^k$	$0(1)^k$
$0(1)^k0$	0
$0(1)^k1$	$0(1)^k$

Table B.2: Calculating the reversed RGMP using mixed states over the ϵ-machine states.