# Extensions of the Theory of Computational Mechanics

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

Computational mechanics is a theory that attempts to describe dynamical systems in a universal language. The central coordinating role in this theory is played by the $\epsilon$-machine—the minimal unifilar representation that is an optimal predictor of the dynamical system.

Properties of the dynamical system are then simply properties of the $\epsilon$-machine. The oldest of these is the entropy rate, or information generation rate, $h_\mu$. A second measure, the first real mark of computational mechanics, is the statistical complexity, or information storage—$C_\mu$. A third is the amount of information transmitted from the infinite past to the infinite future, the excess entropy—**E**.

The fundamental quantities, $h_\mu$ and $C_\mu$, are calculated in closed form directly from the $\epsilon$-machine. Despite being straightforward to define, and having simple geometric relationships to common quantities, excess entropy has resisted a closed form calculation. This has made estimation of **E**, in some cases, quite problematic.

This thesis describes a novel technique for deriving **E** directly from an $\epsilon$-machine. In addition to providing an exact quantification of **E**, it allows for the calculation of **E** as a function of parameterized classes of $\epsilon$-machines. As a theoretical by-product of this technique, several natural quantifiers of stochastic processes have emerged, along with their calculation method. These include: new quantifiers of information storage—reverse and bidirectional statistical complexity; information overhead—directional and bidirectional crypticity; and causal irreversibility.

Of these quantities, the crypticity is treated in most detail. In particular, a new correlation length scale—the cryptic order—is introduced as a natural analog to Markov order. The cryptic order describes the range over which the crypticity exists just as the Markov order describes the time extent of state or synchronization information.

The approaches contained in this thesis are part of a unified effort to delineate a comprehensive theoretical framework for the analysis and understanding of nonlinear dynamical

systems. The novel constructs and calculation techniques described herein represent concrete steps in this direction. We anticipate that this work will have impact on any field concerning stochastic processes, both in algorithms and in conceptual framework.

# Acknowledgments

There are, of course, too many people to thank. If you are reading this, thank *you*.

Thank you, Jim, for sharing with me a whole new field of research, for reminding me to think about the *big ideas*, for demonstrating that physicists can also go to Burning Man, and for incredible freedom in my research.

Thank you, Chris, for your fresh and surprising perspective, your keen eye for fonts, and the endless hours spent helping me compile bleeding matplotlib versions.

Thank you, Sean, for reminding me that sound is important, for introducing me to Mexican Coke, and for giving me an 'in' with the new overlords.

Thank you, Benny, for teaching me about everything from assembly to soldering, for helping me keep my diodes straight, and for spending those many hours on bicycle experiments.

Thank you, Ami, for deep thoughts about deep space, and chocolate cake.

Thank you, Mark, for the canonical and grand canonical ensembles.

Thank you, Marcus, for not teaching me LQG and Farsi at the same time.

Thank you to my one and only and highly singular family. Mom and Dad, I may be one in $2^{23}$, but together you are two in $2^{46}$, which is a lot more than twice as amazing. Max and Molly, thanks for letting me be the oldest, so that we could all be in grad school together.

Thank you, Diane and Keith, for being curious about what I do.

Thank you to Chance, for here we are on a hot planet in a cold Universe and things are so very interesting.

And thank you, Nora, te juro que no vivo un día más sin ti..

# How To Read This Thesis

This thesis draws largely from our series of published papers **cite tba, pratisp, iacp**. Some additional material is drawn from **emim, iacplocs, ruro2, iacp2**.

The first chapter, Ch. 1, provides an introduction to the topic of computational mechanics. It begins with some philosophy and motivation, followed by definitions and some examples of the main constructs. Ultimately it sets the stage for the topics that the remaining chapters address. The reader already familiar with standard computational mechanics might just read the short philosophy section Sec. 1.1.

The next chapter, Ch. 2, contains the main results of **tba,pratisp**. The topics include forward, reverse and bidirectional representations, closed form calculation of the excess entropy and the introduction of several new stochastic process quantifiers.

Following, Ch. 3 presents the work published in **iacp**. This is a detailed exploration of crypticity, in particular the cryptic order—the here defined analog of Markov order.

Several appendices are provided which contain some calculational details, and additionally, a short tutorial on information diagrams. As information diagrams are used throughout, this is a good place to start for those not familiar. For students of information theory, this can provide a nice visual reference for several basic concepts.

# Table of Contents

# List of Figures

# List of Tables

CHAPTER 1

# Computational Mechanics

"Perplexity is the beginning of knowledge. [1]" - Khalil Gibran, trans. Ferris.

## §1.1 Philosophy

The first goal of computational mechanics[2] is to provide a common language with which to describe arbitrary dynamical systems. Ptolemy conceived of the Celestial Sphere in terms of planets, epicycles, deferants and equants in a very geometric theory; Maxwell described charges and currents making use of fields and vector calculus; Modern physicists represent particles using constructs ranging from group representations to topological invariants. How can such theories be compared when their components are so diverse? Computational mechanics seeks to describe these various facets of physical phenomena so that we might begin to ask such questions as, 'Which is least predictable, the sun rising, the ticking of a clock, or the exponential decay of $^{87}$Rb', and, 'Can we claim that to orchestrate the motion of the planets requires more effort than to know the state the stock market? '.

Even limiting ourselves to the ontology of partial differential equations, we find that there are only a few tools available to us for making comparisons. These include: Lyapunov exponents, fractal dimension of the attractor, etc. These tools have a long and important history that continues to impact many fields. However here we aspire to a more extensive and *principled* accounting of the system, and further, we would like for this accounting to serve as more than a method of comparing, ie. testing likeness, and serve as a satisfactory avatar of the system-in-itself. This incarnation is known as the $\epsilon$-machine.

[1] "$[2^{-\sum_p p \log p} > 1] \Rightarrow [H > 0]$" - Khalil Gibran, trans. anonymous.

[2] For a complementary introduction to computational mechanics, there exist several excellent resources including, but not limited to Refs. **DF thesis, CRS thesis, Hanson?, Upper?**.

The second, and more operative, goal of computational mechanics is to make good predictions. Good predictions can be seen as a consequence of having described a system correctly. Stated as an independent goal, it draws attention to two things. First, prediction will be the practical measure of this work; Funding and interest will continue to fuel this research primarily as it pertains to prediction. Conversely, that prediction is viewed here as a consequence of correct description, as much as it is itself a goal, we hope will help to persuade scientists to consider a different perspective.

### §1.1.1 $\epsilon$-machine Prelude

The $\epsilon$-machine is a construct that depends only on two simple ideas. The first is that any description of a dynamical system should be one that is constructed in the language of our interactions with the system. If the way in which we know the system is by sight, then our $\epsilon$-machine ought to be one that is somehow composed of 'sight events'. We might instead, as we will throughout this work, assume to know of our system through digital (wlog binary) measurements. We may wish to think of these as sight events as well since we are likely observing a digital readout display. This is as deeply as we wish to discuss sensory epistemology; we simply wish to contrast our use of sensory data in the construction of an $\epsilon$-machine with a construction that makes use of any preexisting ontology, say, marbles or matrices.

We will use matrices and tensors, but really just as an organizational tool; they will house the manifold probabilities that we must concern ourselves with. We take probability to be something outside of the set of constructive assumptions—fair game for a 'blind' theory. Of course one may argue that this is an assumption with important consequences. For instance, **KW** discusses the implications of an underlying quantum theory. There are surely many interesting things to be said on this topic, but at the risk of becoming overly entangled, we will assume that whatever quantum mechanical operations are at work, are so in a way that is only classically correlated with the observer's data file on their hard drive.

The second idea is an interesting nugget about information—part common sense, part tautology, part kōan. To paraphrase Bateson, *information* is $\{\Delta : \Delta \Rightarrow \Delta'\}$, or 'the differences that make a difference'. What we take from this lesson is that, just as in communication theory, infor-

mation is about deviation from expectation. **say this better** We formalize these concepts in the next section.

## §1.2 Our Domain: Processes

The language in the preceding section is purposefully somewhat vague. The idea being that these principles of dynamical system description might be quite broadly applied. Indeed they have been - to: communication channels, cellular automata **CRS, JPC**, spin systems **DF**, continuous space dynamical systems, continuous time dynamical systems **KW**, quantum dynamical systems **KW**, and dynamical networks **Olaf, other?**.

In this work, we focus on discrete-time, finite-alphabet stationary stochastic processes. These will be referred to as just *[stochastic] processes* or *process languages*. Although the results contained are described and proven in this context, we maintain that the spirit of what we do ought to survive translation into many of the other contexts described above, most likely with some degree of reinterpretation or generalization. We feel that the value of this work is as much in the solving of problems for processes as it is in the generic concepts defined and explored.

While generically the event space for each random variable in a set may be different, we will study those sets of random variables for which the event space is identical for each random variable. For this reason, we allow the following definition.

**Definition.** *An* alphabet, $\mathcal{A}$, *is a set of (possibly continuum) events appropriate to each of a set of random variables.*

**Definition.** *A* word *is a concatenation of symbols from the alphabet.*

$$w = x_0 x_1 \dots x_k \quad , \quad x_i \in \mathcal{A}$$

We will focus on alphabets that are both discrete and finite. These definitions are as you would expect. In fact, for many purposes it will be sufficient to consider only the alphabet, $\mathcal{A} = \{0, 1\}$.

**Definition.** *A* discrete-time, finite-alphabet stationary stochastic process *, or just* process, $\mathcal{P}$, *is a bi-infinite string of random variables,*

$$\dots X_{-3}, X_{-2}, X_{-1}, X_0, X_1, X_2, X_3, \dots$$

*where the random variables have a finite alphabet, $\mathcal{A}$, and,*

$$\Pr(X_t, X_{t+1}, \ldots, X_{t+j}) = \Pr(X_{t+k}, X_{t+1+k}, \ldots, X_{t+j+k})$$

*for any $t, j, k \in \mathbb{Z}$.*

Since the interpretation of the 'time' index in one of these processes is often *time*, and we, being stuck somewhere in the middle of time, have a notion of *past* and *future*, we will use these words to conveniently describe particular sets of random variables. Since we are interested in stationary processes, we may choose to insert ourselves at $t = 0$ and declare one side the past, and the other the future [3].

**Definition.** *A* [finite] future, *denoted $X_0^k$, refers to the $k$ random variables $X_0, X_1, \ldots, X_{k-1}$.*

The short-hand for the finite past is similarly defined.

**Definition.** *A* [finite] past, *denoted $X_{-k}^k$, refers to the $k$ random variables $X_{-k}, X_{-k+1}, \ldots, X_{-1}$.*

Naturally, we are interested in the infinite limits of finite futures and pasts.

**Definition.** *An* infinite future, *or* future, *denoted $\overrightarrow{X}_0$, refers to the infinite set of random variables $X_0, X_1, \ldots$.*

**Definition.** *An* infinite past, *or* past, *denoted $\overleftarrow{X}_{-1}$, refers to the infinite set of random variables $\ldots, X_{-2}, X_{-1}$.*

When we wish to discuss an instance of a random variable, or future, past, finite or infinite, we will use the lowercase, $x_t, x_t^{t+k}, \overrightarrow{x}_t, \overleftarrow{x}_t$.

Note that there are slight asymmetries in the notation. This is just because when states are introduced'zero' is forced to choose a side. We choose zero to fall on the side of the future because of programmers' predjudice. The more complete view of the indexing scheme will be seen when states are involved (see Fig. 1.6).

To make some use of our shorthand, we can now compactly refer to a process and its variables by the appealing form, $\mathcal{P} = \Pr(\overleftarrow{X}, \overrightarrow{X})$.

[3] Note that the 'time' index may also be used to describe a spatial dimension. Occasionally it seems useful to play with the interpretation of this index as it provides useful alternate perspectives.

## §1.3 $\epsilon$-machines

Now that we have laid out the goals and principles of the construction as well as the domain to which it will be applied, we describe the $\epsilon$-machine itself.

Recalling the two simple ideas underlying the $\epsilon$-machine, we see that the first is satisfied by assuming that the process at hand was obtained through digital measurement of some dynamical system. The second is satisfied by utilizing the appropriate equivalence relation.

**Definition.** *Given a process, $\mathcal{P}$, define an equivalence relation $\sim_\epsilon$ where*

$$\overleftarrow{x} \sim_\epsilon \overleftarrow{x}' \Leftrightarrow \Pr(\overrightarrow{X} \,|\, \overleftarrow{X} = \overleftarrow{x}) = \Pr(\overrightarrow{X} \,|\, \overleftarrow{X} = \overleftarrow{x}')$$

That this relation is reflexive, symmetric and transitive follows straightforwardly from the dependence on the inner equivalence relation, $=$. Throughout this work, we will only use the subscripted symbol, $\sim_\epsilon$, when contrasting with another relation.

Having defined an equivalence relation, several things are immediately forthcoming. First, the equivalence relation defines equivalence classes.

$$[\overleftarrow{x}] = \{\overleftarrow{x}' : \overleftarrow{x}' \sim \overleftarrow{x}\}$$

The set of these equivalence classes is a quotient set.

$$\{\overleftarrow{x}\}/\sim = \{[\overleftarrow{x}'] : \overleftarrow{x}' \in \{\overleftarrow{x}\}\}$$

The relation also induces a surjective map from the original set to the equivalence classes.

$$\epsilon : \{\overleftarrow{x}\} \to \{\overleftarrow{x}\}/\sim \quad , \quad \epsilon(\overleftarrow{x}) = [\overleftarrow{x}]$$

As these equivalence classes are the building blocks of the $\epsilon$-machine, and are at the core of nearly all calculations, we allow ourselves to dub these particular classes, *causal states*. This definition is intended to be suggestive of the $\epsilon$-machine as a probabilistic automaton, and also for notational convenience.

**Definition.** *The set of* causal states, $\{\mathcal{S}\} = \boldsymbol{\mathcal{S}}$, *is in one-to-one correspondence with the set of equivalence classes,* $\{\overleftarrow{x}\}/\sim$.

The set of causal states [4] can be discrete, fractal, or continuous; see, e.g., Figs. 7, 8, 10, and 17 in Ref. [**?**].

[4] 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.

**Definition.** *We say that a causal state, $\mathcal{S}$, is* induced *by a past, $\overleftarrow{x}'$, if $\mathcal{S}$ corresponds to $[\overleftarrow{x}]$ where $\overleftarrow{x}' \in [\overleftarrow{x}]$.*

Let us note that it makes sense to think of causal states are being induced, not only by particular pasts, but also by equivalence classes of pasts.

If two pasts, $\overleftarrow{x}$ and $\overleftarrow{x}'$, are members of the same equivalence class, $[\overleftarrow{x}]$, then by definition, $\Pr(\overrightarrow{X}|\overleftarrow{X} = \overleftarrow{x}) = \Pr(\overrightarrow{X}|\overleftarrow{X} = \overleftarrow{x}')$. We might wish to make a less particular statement such as, $\Pr(\overrightarrow{X}|\overleftarrow{X} = \overleftarrow{x}) = \Pr(\overrightarrow{X}|\overleftarrow{X} = \overleftarrow{x}') \simeq \Pr(\overrightarrow{X}|\overleftarrow{X} = [\overleftarrow{x}])$. This is intuitively correct, but slightly awkward since the random variable $\overleftarrow{X}$ does not have events in the space of equivalence classes. Instead we say $\Pr(\overrightarrow{X}|\overleftarrow{X} = \overleftarrow{x}) = \Pr(\overrightarrow{X}|\overleftarrow{X} = \overleftarrow{x}') = \Pr(\overrightarrow{X}|\mathcal{S} = \sigma)$ where $\sigma$ is the particular causal state induced by any member of the class $[\overleftarrow{x}]$. This leads to the primary utility of causal states, which is as a short-hand, or stand-in for particular pasts.

So far the causal states have been constructed as sufficient replacement variables for the infinite past. Given that the set of pasts is uncountably infinite and the set of causal states is, for the present, finite, this is a tremendous compactification of knowledge. However, it is not yet the useful tool we desire. What we have provided so far is a 'routing' variable [5]. The real power of the causal state will be its dynamic function in the $\epsilon$-machine—doling out the appropriate future bit by bit. In this way, the causal state is not only a short-hand for the infinite past, but it also shields us from the infinite variety of the future, allowing us to ratchet forward in time one symbol at a time.

In addition to states, we evidently need some notion of dynamic, as it is a dynamical system we are representing. We capture this dynamic with a set of [symbol-] labeled transition matrices. Specifically, the value of an element of the "$x$'th" matrix is defined,

$$T^{x}_{\sigma,\sigma'} = \Pr(\mathcal{S}_{t+1} = \sigma', X_{t+1} = x|\mathcal{S}_t = \sigma)$$

This is the conditional probability that, given a particular causal state, $\sigma$, or equivalently a past that induces $\sigma$, the following measurement symbol will be $x$; this will consequently induce causal state $\sigma'$. Since the process is stationary, the value of the variable $t$ is unimportant.

Causal states have a Markovian property that they render the past and future statistically

[5] Imagine a www tool that accepts an infinite past and then provides you with a URL. This URL then leads to a page with an infinite set of infinite futures. This *is* in some sense what we want, but we'd rather not drown in the data just yet. We would hope to control the data flow.

independent; they *shield* the future from the past [**?**]:

$$\Pr(\overleftarrow{X}, \overrightarrow{X} | \mathcal{S}) = \Pr(\overleftarrow{X} | \mathcal{S}) \Pr(\overrightarrow{X} | \mathcal{S}) . \tag{1.1}$$

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: $I[\mathcal{S}; \overrightarrow{X}] = \mathbf{E}$.

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

$$\Pr(\overleftarrow{X}, \overrightarrow{X} | \mathcal{S}) = \Pr(\overleftarrow{X} | \mathcal{S}) \Pr(\overrightarrow{X} | \mathcal{S}) . \tag{1.2}$$

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: $I[\mathcal{S}; \overrightarrow{X}] = \mathbf{E}$.

$\epsilon$-Machines have an important structural property called *unifilarity* [**?**, **?**]: From the start state, each symbol sequence corresponds to exactly one sequence of causal states [6]. $\epsilon$-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 measurement symbol appears on at most one outgoing transition [7]. 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 $\epsilon$-machine is its unique, minimal unifilar model.

To summarize, a causal state is a set of pasts that each have the same correlation with, or prediction for, the future, $\Pr(\overrightarrow{X} | \overleftarrow{x})$. The $\epsilon$-machine is obtained by linking neighboring causal states together with the appropriate interstitial observed symbol. As promised, its component states are equivalent to sets of past observations. The probabilistic dynamic induced on the causal states is exactly the one which the data demands as sequential observations induce sequential

[6] 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 [**?**].

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

causal states.

## §1.4 First Examples

All of this may appear a bit more abstract than necessary, and in some sense this is true. The example $\epsilon$-machines that appear throughout this work are straightforward to draw with a pen and paper in only a minute or two. The beauty is that even these elementary examples will provide us with a means for motivating and uncovering plenty of interesting science.

### §1.4.1 IID

We would be remiss if we failed to begin with the an independent, identically distributed (IID) process. By definition, each measurement is independent of the past. Therefore, there is only one conditional distribution: $\Pr(0|\cdot) = p, \Pr(1|\cdot) = 1 - p$. In turn, there is only one equivalence class of histories, and thus only one causal state 1.1.

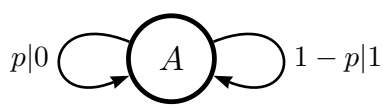


Figure 1.1: IID binary process.

The labeled transition matrices are one-dimensional,

$$T^0_{A,A} = \begin{bmatrix} p \end{bmatrix}, T^1_{A,A} = \begin{bmatrix} 1-p \end{bmatrix}$$

This is a rather trivial, but important process to keep in mind because many kinds of data are taken without consideration for temporal correlation. We will see in how many *independent* ways this process class actually is trivial. Let us use this to motivate our discussion of measures of processes.

### §1.4.2 Entropy and Entropy Rate

The first aspect of a process we would like to characterize is its variability; more precisely, we would like to know the degree of uncertainty in particular groups of random variables. Following Shannon **cite shannon**, we characterize the uncertainty by the Shannon information.

**Definition.** *The* Shannon information, *or here the* entropy [8], *of a random variable is given by,*

$$H[X] = -\sum_{x \in X} \Pr(X = x) \log \Pr(X = x)$$

*Analogously, the entropy of a set of random variables, or their joint probability distribution, is defined,*

$$H[X, Y] = -\sum_{x \in X, y \in Y} \Pr(X = x, Y = y) \log \Pr(X = x, Y = y)$$

In the trivial case of our first example (Fig. 1.1), the uncertainty in the bi-infinite string of random variables is infinite. Each random variable has some finite uncertainty associated with it, and by definition of being IID, the uncertainty in any given variable is independent of any other variable. Therefore, the uncertainty in the entire string is infinite. Since a categorically infinite entropy leaves little to discuss [9], we will be primarily interested in the functional relation between entropy and length scale as the length grows. Furthermore, we are interested in not only the asymptotic behavior (some kind of exponential envelope), but also the finite length behavior. In fact we are more interested in the finite length behavior as long as we are guaranteed *some* kind of convergence. The parent entropy function that many important process features derive from is the block entropy.

**Definition.** *The* block entropy function *(or curve [10]) is is the entropy of the distribution of words at length L.*

$$H[X_0^L] = -\sum_{x_0^L \in \mathcal{A}^L} \Pr(X_0^L = x_0^L) \log \Pr(X_0^L = x_0^L)$$

**properties: nondecreasing, concave**

**Definition.** *The* entropy rate, $h_\mu$, *of a process is the limit of the conditional entropy,*

$$h_\mu = \lim_{L \to \infty} H[X_0 | X_{-L}, X_{-L+1}, \ldots, X_{-1}]$$

[8] As in most information theory, computer science and symbolic dynamics, the log function in this work will always mean base two.

[9] Actually there is much to discuss. For continuous time or continous valued output, measures such as the differential entropy will generically be infinite. We hope that a talented mathematician will generalize the ideas here to the continuous case.

[10] This is a discrete function and so we use the word curve to be suggestive of the fact that these functions are highly restricted by monotonicity and such and so are, as far as discrete functions go, relatively curve-like.

*In our shorthand,* $h_\mu = H[X_0|\overleftarrow{X}]$. *We are also interested in finite length approximations to the entropy rate,*

$$\begin{aligned} h_\mu(L) &= H[X_0|X_{-L}, X_{-L+1}, \ldots, X_{-1}] \\ &= H[X_{-L}, X_{-L+1}, \ldots, X_{-1}, X_0] - H[X_{-L}, X_{-L+1}, \ldots, X_{-1}] \end{aligned}$$

**properties: nonincreasing, convex** The entropy rate estimate at $L = 0$ is defined to be $h_\mu(0) = \log(|\mathcal{A}|)$. This is just stating that if you know nothing about the process other than the size of the alphabet, your uncertainty is maximal,that is, the entropy of a uniform distribution over the alphabet.

$$\begin{aligned} H[\frac{1}{|\mathcal{A}|} \times (1, 1, \ldots, 1)] &= -\sum_{i=1}^{N} \frac{1}{|\mathcal{A}|} \log \frac{1}{|\mathcal{A}|} \\ &= \log |\mathcal{A}| \end{aligned}$$

If we think of this geometrically, the entropy rate is the limit of the discrete slope of the block entropy function. Simple geometric features of these key functions will play a primary role in both motivating classifications of processes and also in understanding features otherwise defined. We can write the probability distribution for an IID process as a product of distributions

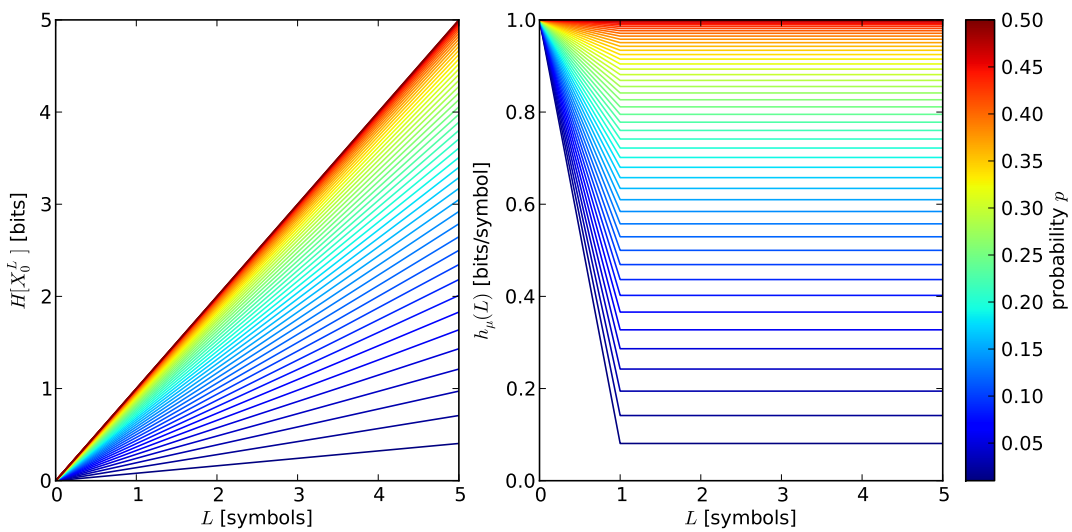


Figure 1.2: (Left) Block entropy curve for all binary IID processes ($\Pr(X = 0) = p$). Each is simply a line through the origin with slope determined by $p$. Since there is a symmetry in the class about $p = 0.5$, only on half of the $p$ values are illustrated. (Right) Finite length entropy rate approximations. Colorbar indicates the probability, $p$, that is varied to obtain different members of a process family. The particular probability, $p$, refers to the variable in Fig. 1.1.

for each random variable.

$$\Pr(\overleftarrow{X}, \overrightarrow{X}) = \ldots \times \Pr(X_{-1}) \times \Pr(X_0) \times \Pr(X_1) \ldots$$

As a consequence, the block entropy is linear in the length, $L$.

$$\begin{aligned}
H[X_0^L] &= -\sum_{x_0^L \in \mathcal{A}^L} \Pr(X_0^L = x_0^L) \log \Pr(X_0^L = x_0^L) \\
&= -\sum_{x_0, x_1, \ldots, x_{L-1} \in \mathcal{A}} \Pr(X_0 = x_0) \times \ldots \times \Pr(X_{L-1} = x_{L-1}) \log \Pr(X_0 = x_0) \times \ldots \times \Pr(X_{L-1} = x_{L-1}) \\
&= -\sum_{x_0 \in \mathcal{A}} \Pr(X_0 = x_0 \log \Pr(X_0 = x_0)) - \ldots - \sum_{x_{L-1} \in \mathcal{A}} \Pr(X_{L-1} = x_{L-1} \log \Pr(X_{L-1} = x_{L-1})) \\
&= -L \times \sum_{x_0 \in \mathcal{A}} \Pr(X_0 = x_0) \log \Pr(X_0 = x_0) \\
&= -LH[X_0]
\end{aligned}$$

An immediate consequence of this is that the entropy rate is equal to the length-one approximation.

$$\begin{aligned}
h_\mu &= \lim_{L \to \infty} H[X_0^L] - H[X_0^{L-1}] \\
&= \lim_{L \to \infty} LH[X_0] - (L-1)H[X_0] \\
&= H[X_0]
\end{aligned}$$

That is, beyond considering the most trival statistic, there is nothing more to learn about this system. Since this entropy, or entropy rate, will arise rather frequently, it is often referred to simply as the *binary entropy* and denoted $H(p)$.

A more interesting process is one for which the finite length approximations to $h_\mu$ are something other than constant. A very simple example that illustrates this [11] is the Golden Mean Process (Fig. 1.3). Notice that in Fig. 1.4, the entropy rate estimates reach the entropy rate at $L = 2$.

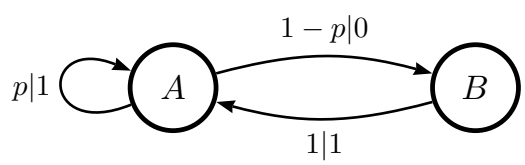


Figure 1.3: The Golden Mean Process is one that disallows consecutive zeros. After a zero [one] is seen, the process is in causal state $B$ [$A$].

To restate, for the Golden Mean process, the conditional uncertainty in every symbol, including

[11] In fact, *any* process other than IID will illustrate this. I have just chosen a simple one that is convenient because it is simple in other ways.

and beyond the second, is $h_\mu$.

$$H[X_0|X_{-k},\ldots,X_{-1}] = H[X_0|X_{-1}] = h_\mu$$

This leads one to naturally speculate that there is a conditional *probabilistic* independence as well as this conditional *entropic* independence. This is true, but because it is an intuitive result with a somewhat inelegant proof, it can be found in App. C We should be sure to point out

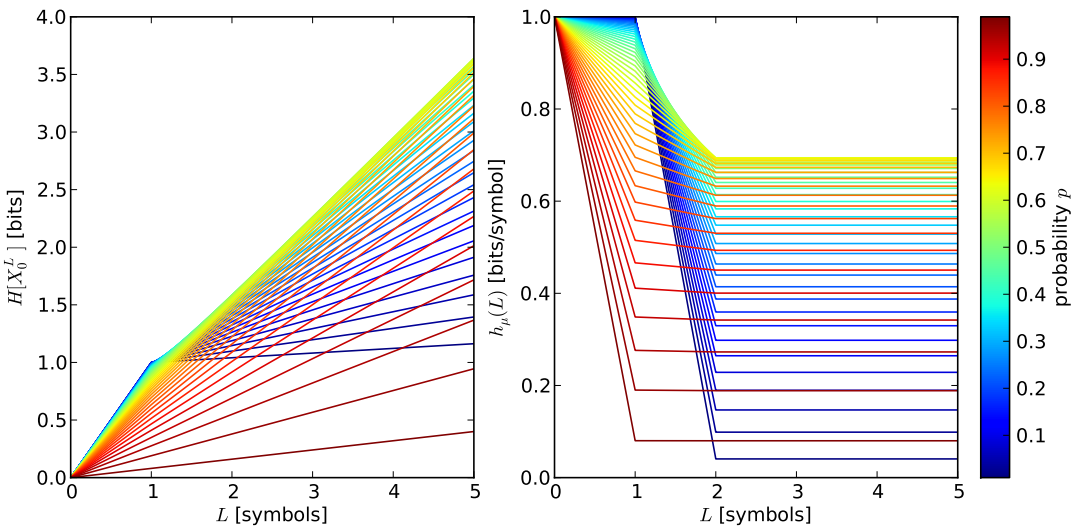


Figure 1.4: (Left) Block entropy curves for all Golden Mean processes. Each is linear for $L \geq 1$. (Right) Each finite length entropy rate estimate reaches its asymptotic value $h_\mu$, at $L = 2$. This indicates that the additional uncertainty in the $L = 2$ blocks, beyond the $L = 1$ blocks, is already $h_\mu$. This implies that the minimum correlation length required for maximal prediction ability is $L = 1$. That is, the Golden Mean is an order-1 Markov process.

an important feature of the $\epsilon$-machine. In contrast with generic hidden Markov models, the $\epsilon$-machine has the property that the entropy rate can be calculated directly from it. We argue that this is a consequence of the $\epsilon$-machine being the natural representation of the process.

Now that the entropy rate has been defined as a limit and is certainly straightforward enough to estimate, we should look for a closed form. We might imagine that, for any hidden Markov model (composed of states $\mathcal{R}$), that the time average surprise is the same as the state average surprise (when weighted by state visitation probabilities). Specifically, that

$$h_\mu \stackrel{?}{=} \sum_\rho \Pr(\mathcal{R}_0 = \rho) H[X_0|\mathcal{R}_0 = \rho])$$

It is easy to see that this is not true. Consider a nonunifilar presentation of the Golden Mean Process. The conditional entropies are $H[X_0|\mathcal{R}_0 = A] = 0$ and $H[X_0|\mathcal{R}_0 = B] = 0$. No weighted

sum of these conditional entropies will yield what we know to be a non-zero entropy rate. The

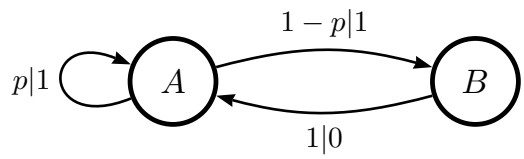


Figure 1.5: Nonunifilar presentation of the Golden Mean Process. The entropy rate of the process is *not* simply the weighted average of the entropies of symbols emitted after visiting each state.

correct form for the entropy rate is in fact this weighted sum of individual state entropies, but *only* when the states in question are *causal* states. So for $\epsilon$-machines, we have the following closed form for the entropy rate in terms of causal state asymptotic probabilities and transition probabilities.

$$\begin{aligned} h_\mu &= H[X_0|\mathcal{S}_0] \\ &= \sum_{\sigma\in\mathcal{A}} \Pr(\mathcal{S}_0=\sigma) H[X_0|\mathcal{S}_0=\sigma]) \\ &= -\sum_{\sigma\in\boldsymbol{\mathcal{S}}} \Pr(\mathcal{S}) \sum_{x\in\mathcal{A}\sigma'\in\boldsymbol{\mathcal{S}}} T^{(x)}_{\mathcal{S}\mathcal{S}'} \log_2 \sum_{\sigma'\in\boldsymbol{\mathcal{S}}} T^{(x)}_{\mathcal{S}\mathcal{S}'} \end{aligned}$$

## §1.5 Statistical Complexity

Above, $\Pr(\mathcal{S})$ 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 $\pi$ to denote the distribution over the causal states as a row vector.[12] This distribution over states leads to a second fundamental characterization of processes—the statistical complexity.

**Definition.** *A process's* statistical complexity, $C_\mu$, *can be directly calculated from its $\epsilon$-machine as it is a property of the dynamic over the causal states:*

$$\begin{aligned} C_\mu &= H[\mathcal{S}] \\ &= -\sum_{\{\mathcal{S}\}} \Pr(\mathcal{S}) \log_2 \Pr(\mathcal{S}) \,. \end{aligned} \tag{1.3}$$

The statistical complexity is a *statistical* complexity as opposed to a deterministic one, such as Kolmogorov complexity,[13] meaning that the measure is intended to capture the complexity

[12] The matrix algebra here follows the 'state on the left, transition matrix on the right' convention.

[13] Kolmogorov complexity is also known as: descriptive complexity, Kolmogorov-Chaitin complexity, stochastic complexity, algorithmic complexity, algorithmic entropy, and program-size complexity.

of a class of data rather than a particular instance. To illustrate, a coin may be flipped to generate a variety of sequences. One such sequence is an alternating sequence of heads and tails, $(HT)^N$. This sequence may be generated by the compact[14] program: i=0,for(i<N){write H, write T, i=i+1}. The increasing majority of sequences will not have this compactness, yet all are possible realizations of the output of this simple dynamical system. The goal of a statistical complexity is to provide a characterization of all of these possiblilities.

Why should we aim to describe the broad class behavior rather than the detailed behavior indicated by a particular data string? If our aim is to describe the physical dynamical system, and if we believe that this system has inherent unpredictability, then describing a particular instance, as the Kolmogorov complexity would do, might actually *overspecify* the physical system. For instance, a flipped coin could certainly produce a binary representation of $\pi$; another might code for the name of the next president. Neither one of these things captures the essence of the physical system—that it is IID and uniformly[15] random heads and tails. If we insist on characterizing the coin's behavior by the two instances above, then I argue that we ought to consider *all* possible instances. This is clearly not a productive use of time.

What exactly is the statistical complexity telling us? As $C_\mu$ is defined as the entropy of a probability distribution over some event space, it can immediately be understood in the context of communication theory. If Alice wishes Bob to synchronize his ensemble of identical dynamical systems to hers, she must communicate $C_\mu$ bits per member of the ensemble.[16] If it is $C_\mu$ bits that is passed to Bob to describe the state of the system, then it could be said that each system *carries* that amount of information. This is why the statistical complexity is interpreted as *stored information*.

Why is this a good measure of complexity? As we live in a world where complexity measure abound, it is important to pause and reflect on the particular contribution of a particular measure. We first claim that the above description of $C_\mu$ as stored information is strong evidence for its naturalness. Additionally, it has some properties that although somewhat trivial, are not shared by all. In thinking about the range of possible processes, it is hard to argue that IID pro-

[14]We call this compact because as the size of the program goes as $\log(N) + C$. Thus the limit of the ratio of output sequence size to program size is zero. This indicates that this subprocess has and entropy rate of zero.

[15]Some might argue that we have to toss the coin with more vigor (see **diaconis coin**).

[16]Of course she must communicate $C_\mu$ bits *on average*, but that is the standard assumption made in information theory. In fact, without it, information does not have the same meaning.

cesses are not on *some* particular extreme. Correspondingly, the statistical complexity of any IID process is zero. This seems to satisfy our intuition about what a complexity measure ought to say about IID processes. In another corner of process space lie completely predictable. These processes, certainly for finite cases, are just the periodic ones. The statistical complexity of these will be the log of the period length. Another good reason is that $C_\mu$ has a kind of extensivity. If two uncorrelated processes are 'placed side-by-side', which is the usual thing to do when testing for extensivity, the joint process characterized by the process language over the appropriate tuples of symbols has a statistical complexity which is simply the sum of the individual complexities.[17]

Thus, the $\epsilon$-machine directly gives two important properties: a process's rate ($h_\mu$) of producing information and the amount ($C_\mu$) of historical information it stores in doing so.

## §1.6 Excess Entropy

The entropy rate is a property that comes straight out of communication theory. In that context, it is the minimum capacity of an error-free channel; equivalantly, it is the amount of supplementary information required for maintaining perfect decoding (which we think of as prediction). Another concept that arises very naturally in the communication context is the transference of information from input to ouput. Different channels, depending on their capacity, noise present, etc., will have varying abilities to transfer information from one side of the channel to the other. We can cast a dynamical system or time-series as a channel in the following way; The past is considered the input, the future is the output, and the channel itself is the $\epsilon$-machine. Given this picture, the excess entropy is the amount of information about the past that is transmitted via the $\epsilon$-machine channel to the future (See Fig. 1.6). We express this mathematically in terms of a mutual information.

$$I[\overleftarrow{X}; \overrightarrow{X}] = \mathbf{E}$$

Excess entropy has gone by several different names and has been reinvented several times **cite early JPC, Grassberger, etc**. There are several equivalent forms for **E**, see Ref. [**?**, and references

[17] I might argue that this way of thinking about extensivity is a little mundane. The side-by-side test for extensivity is born of thinking about equilibrium systems. As $\epsilon$-machines are definitely not equilibrium systems, it would be most interesting to test for extensivity in different ways. One might sample from the two systems in an alternating manner. We can imagine something more drastic, and maybe harder to motivate, like the graph-join of the two $\epsilon$-machines. Presently, only little is known about about the consequences of these types of actions.

therein]. Here we quote the definition of excess entropy from Ref. **RURO**, where the name is somewhat more intuitive.

**Definition.** *The* excess entropy *is the sum over word lengths of the degree to which the entropy rate estimate is in excess of the true entropy rate.*

$$\mathbf{E} = \lim_{L'\to\infty} \sum_{L=1}^{L'} (h_\mu(L) - h_\mu)$$

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 ($\mathbf{E}$) 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 [**?**, **?**].

It can be demonstrated that this definition is, at least for the types of processes studied in this thesis,[18] equivalent to the mutual information concept.

$$\begin{aligned}
I[\overleftarrow{X}; \overrightarrow{X}] &= \lim_{L\to\infty} I[X_{-L}^L; X_0^L] \\
&= \lim_{L\to\infty} H[X_{-L}^L] + H[X_0^L] - H[X_{-L}^L, X_0^L] \\
&= \lim_{L\to\infty} H[X_0^L] + H[X_0^L] - H[X_0^{2L}] \\
&= \lim_{L\to\infty} 2H[X_0^L] - H[X_0^{2L}] \\
&= \lim_{L\to\infty} 2(H[X_0^L] - Lh_\mu) - H[X_0^{2L}] + 2Lh_\mu \\
&= \lim_{L\to\infty} 2\sum_{L'=1}^{L} (H[X_0^{L'}] - H[X_0^{L'-1}] - h_\mu) - \sum_{L'=1}^{2L} (H[X_0^{L'}] - H[X_0^{L'-1}] + h_\mu) \\
&= \lim_{L\to\infty} 2\sum_{L'=1}^{L} (h_\mu(L') - h_\mu) - \sum_{L'=1}^{2L} (h_\mu(L') + h_\mu) \\
&= 2\lim_{L\to\infty} \sum_{L'=1}^{L} (h_\mu(L') - h_\mu) - \lim_{L\to\infty} \sum_{L'=1}^{2L} (h_\mu(L') + h_\mu) \\
&= 2\mathbf{E} - \mathbf{E} = \mathbf{E}
\end{aligned}$$

The second line follows from the definition of mutual information. The third line is a result of stationarity. We insert some copies of $h_\mu$ and rearrange to form the definition of $\mathbf{E}$. To split the limit into two, we assume the existence of the individual limits. This really amounts to assuming

[18] Extensions as benign as the addition of an extra dimension—to a 2D process—necessitate more care with these equivalences **cite feldman 2D**.

the existence of one limit—**E**. In recent work, see Ref. **nick**, it is shown that this limit exists for all $\epsilon$-machines with a finite number of states.

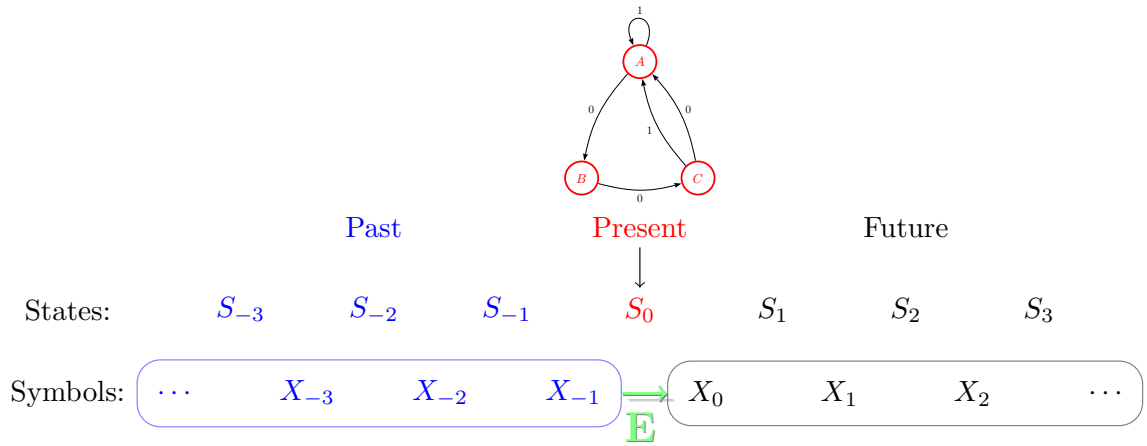


Figure 1.6: A stochastic process can be viewed as a communication channel. The data in the past is the input to the channel. The channel itself is the dynamical system, or $\epsilon$-machine, which transmits information to the future. The total information transmitted from past to future is equal to the excess entropy.

We can begin to collect our understanding of the information theoretic relationships among $\epsilon$-machine variables using an I-diagram (see Fig. 1.7). For a review of I-diagrams, see App. B. The other two quantities shown in this diagram are $H[\overleftarrow{X} \,|\, \overrightarrow{X}]$ and $H[\overrightarrow{X} \,|\, \overleftarrow{X}]$. Since these will generally be infinite quantities, it can be useful to think of the random variables in their finite forms, $X_{-k}^{k}$ and $X_0^k$. The rate of growth of these agglomerated variables (with $L$) is bounded above by $H[X_0]$ and below by $h_\mu$.

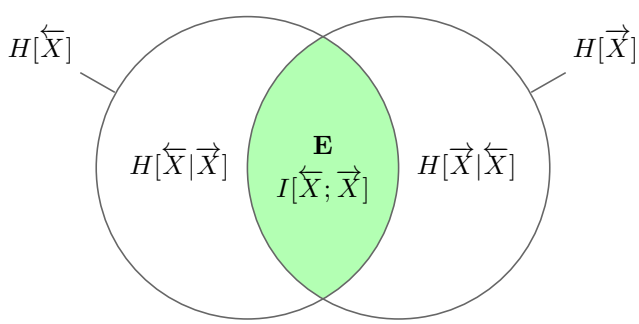


Figure 1.7: This I-diagram highlights the role of excess entropy as the mutual information between past and future data.

To form a complete I-diagram for an $\epsilon$-machine, we must introduce a state variable. Of course this diagram is not a complete description, but it does aid our thinking in several ways.

We start by adding a generic state, actually any random variable at all will do. In Fig. 1.8, all possible information relations among the variables are listed.

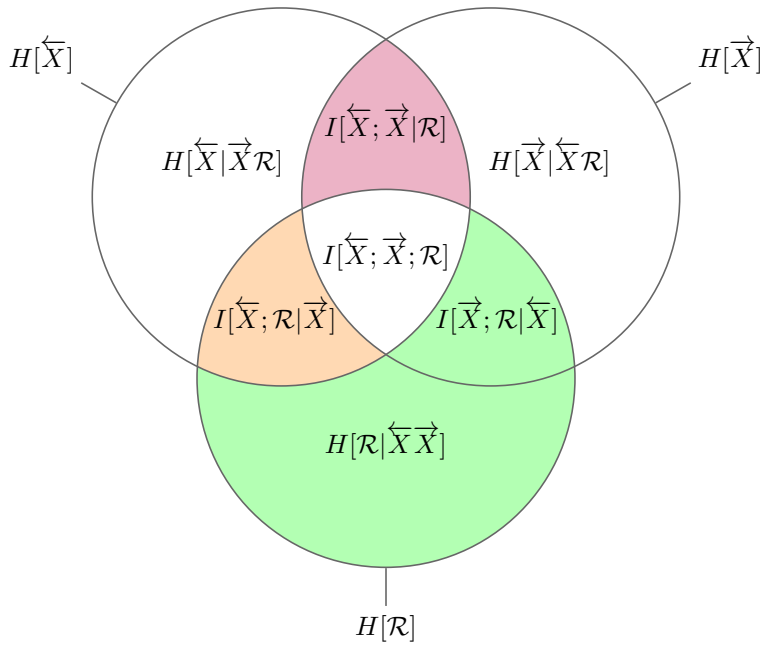


Figure 1.8: The generic relation among the past, future and a state, $\mathcal{R}$ includes 15 nontrivial information quantities. Demanding that the state involved is a causal state effects 4 of these quantities. The green area (which corresponds to two information atoms) is zero because the causal state is a single-valued function of the past. The purple area is zero because causal states are prescient. The orange area is not zero, but is the minimum value possible, given that green and purple are zero.

Note that some of the region are colored. This is to indicate that there is a difference between a generic state variable and a *causal* state in so far as these areas [19] are concerned. Let us explore these individually, substituting a causal state, $\mathcal{S}$ for the generic state, $\mathcal{R}$.

The green area is zero because the causal state is a single-valued function of the infinite past, $\overleftarrow{x}$. Since $H[\mathcal{S}|\overleftarrow{X}\,\overrightarrow{X}]$ is also a conditional entropy, and therefore positive, we have that the green subregions are individually zero.

The purple region is zero because the probability distribution over futures given a past is the same as that given the induced causal state. This implies that a past and the causal state it induces share the same amount of information with the future. Since we already have that

[19] The words 'area', 'region', '[information] quantity' and '[information] atom' are used interchangeably here in light of the correspondence between Venn diagrams and information theory. For more about this relationship refer to App. B.

$I[\overrightarrow{X};\mathcal{S}|\overleftarrow{X}] = 0$, this shared information must be the *same* information. Recalling that $I[\overleftarrow{X};\overrightarrow{X}] = \mathbf{E}$, we then arrive at $I[\mathcal{S};\overrightarrow{X}] = \mathbf{E}$.

The orange region is generically not zero, although the $\epsilon$-machine ensures that it is, given that the previously describe regions are zero, the smallest possible value. It is this orange region that will be the subject of much discussion later on. It is a quantity governed by opposing forces; on the one hand, it must be large enough to accomodate capturing all of the information relevant to the future ($\mathbf{E}$), while on the other hand it is asked to be as small as possible, giving the minimal unifilar optimally predictive representation. This quantity is called the crypticity.

## §1.7 Estimation of Excess Entropy

The difficulty in obtaining accurate estimates of the excess entropy in even relatively benign systems was the primary (initial) impetus behind our effort to reframe this problem. This section is not intended to provide a comprehensive accounting of the various ways in which estimation can be difficult, nor will it quantify exactly how difficult the estimation is. We will see through a simple example that it is indeed difficult, and argue that this is generic enough to warrant searching for an alternate method. The method having been discovered and detailed in Ref. **PRATISP** obviates the need to revisit and detail the previous study of difficulty in estimation. **a little boring**

### §1.7.1 Example of Sharp Convergence : Order-3 Markov

In order to set the stage for the difficult estimation task in the next section, we interrupt to offer an apparently substantial process to contend with (see Fig. 1.9). This process, being an order-3 Markov process, is a fair test case for estimation algorithms as finite order Markov models are used in a wide variety of settings **cite some Markov modeling refs**.

Calculating the standard excess entropy estimates, we see in Fig. 1.10 that for all instances of the class, there is a sharp convergence at $L = 3$. This is a consequence, and additionally an indicator, of the process being order-3 Markov. In fact, excess entropy can be calculated for finite order Markov processes in a finite way. Since we have that the entropy rate estimate becomes exact at $L = R + 1$ for an order-R markov process. This has the effect of truncating the infinite

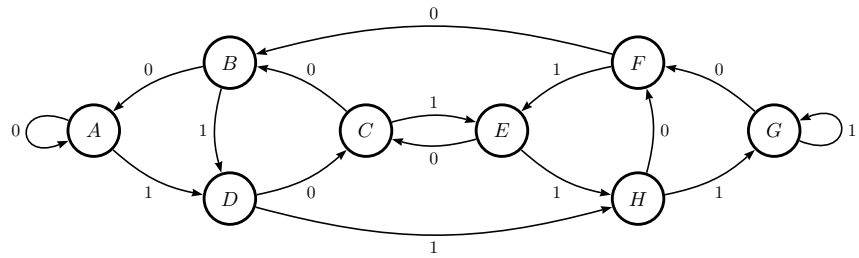


Figure 1.9: A process with 8 causal states. Since each state has two outgoing transitions, each of which has one free parameter, we suppress the probabilities here. The reader may verify that this is the structure of an order-3 Markov process—any 3 symbols will uniquely define a state (the converse happens to also be true in this instance).

sum,

$$\begin{aligned}\mathbf{E} &= \sum_{L=1}^{\infty}(h_\mu(L) - h_\mu) \\ &= \sum_{L=1}^{R}(h_\mu(L) - h_\mu) \\ &= H[X_0^R] - Rh_\mu.\end{aligned}$$

The last step is accomplished by collapsing the telescoping sum of entropy rate estimates. It appears that for finite order Markov processes, certainly for small orders, the excess entropy is easily calculable. We should now ask the questions: "What happens as the Markov order becomes large?", and "What happens when the process is not finite order Markov?"

To answer the first question, we have to calculate the probabilities of roughly $|\mathcal{A}|^R$ different length-$R$ words. The number of words will be smaller than this depending on the process's forbidden words. For large alphabets and large orders, this can quickly become a challenging task. For instance, if we treat the English language as a Markov process over letters $\{a, \ldots, z\}$and allow for a very modest correlation length of 6, we find that there is not even enough space on a modern computer to store the resulting probability distribution.

To illustrate the reponse to the second question, let us investigate a very simple non-Markov process.

## §1.7.2 Example of Slow Convergence : Even Process

The example we use to illustrate this complication is the Even Process, as seen in Fig. 1.11. The reason that this process is an appropriate choice for illustrating difficulty with convergence is

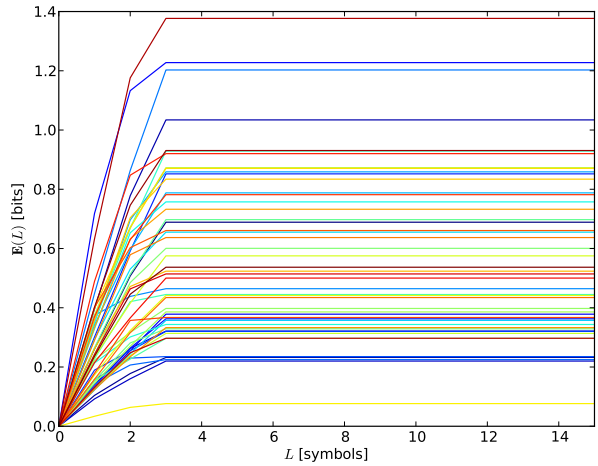


Figure 1.10: Excess entropy estimates for 50 instances of full order-3 Markov chains (see Fig. 1.9 for the topology). Notice that the estimates become extremely good (actually exact) at $L = 3$. This is a consequence of the process being finite order Markov.

that it is not Markovian; we might say that it is infinite-order Markov. This can be intuitively understood in the following way: If we have access to only a finite symbol history, say $N$ symbols, then when we encounter a word of $N$ ones, we cannot provide the proper distribution over futures. At a coarse level, we don't know whether the sequence is currently even in length and therefore has the option of terminating with a zero, or if it is currently odd in length and therefore *must* continue with at least one more one. Therefore no finite history (finite order Markov) model can properly generate the Even Process. This is a fundamental difference[20] between Markov models, or chains; and hidden Markov models, or functions of Markov chains **cite Markov vs hidden Markov**.

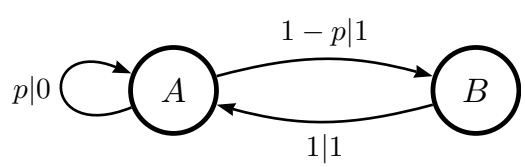


Figure 1.11: The Even Process requires that all blocks of uninterrupted ones, with zeros on either side, be even in length.

[20]This fundamental difference is *not* the difference between node output models and edge output models. Given a unifilar edge output model (this is a 'hidden model') with $N$ nodes and $M$ symbols, there are at most $N \times M$ edges. The corresponding node output model (also a 'hidden' model) then trivially has at most $N \times M$ nodes. This difference is of course important, but never involves transforming a 2 state model to an infinite state model

| Prediction | $N$ even | $N$ odd |
|---|---|---|
| $\Pr(X_N = 0 \vert X_0 = 1, \ldots, X_{N-1} = 1)$ | $1 - \frac{p}{2}$ | $\frac{p}{2}$ |
| $\Pr(X_N = 0 \vert X_{-1} = 0, X_0 = 1, \ldots, X_{N-1} = 1)$ | $p$ | 0 |
| $\Pr(X_N = 0 \vert \mathcal{S}_N = A)$ | $p$ | – |
| $\Pr(X_N = 0 \vert \mathcal{S}_N = B)$ | – | 0 |

Table 1.1: The above table illustrates that for the Even Process, for any length $N$, there exists a word (all ones) such that prediction based on that word alone is different than prediction based on that word knowing that the previous symbol is 0. Notice that the optimal probabilities, those predicted after the block of ones is begun by a zero, are the same as those predicted by the appropriate induced causal state.

It is plain to see in the left pane of Fig. 1.12 that, for a substantial subset of instances, the relative errors in the excess entropy estimates do not fall within acceptable bounds even when considering correlation lengths up to 10. The excess entropy, defined as the infinite sum of the entropy rate overestimates, is continually being fed by new overestimates, as is seen in the right pane of Fig. 1.12. One should probably object at this point saying that with appropriate algorithms and compute power, lengths far beyond 10 must certainly be accessible. This objection is certainly valid, but misses the point of the illustration. First, this is only the 'simplest of the difficult' examples. Depending on the application, model sizes will have dozens or hundreds of nodes. Second, as a matter of theoretical investigation, we would like to be able to calculate the excess entropy for infinite $\epsilon$-machines. Any previous algorithm will fail in this task. Third, the brute force griding out of ever better approximations does not strike us as the proper way to really understand how excess entropy behaves. To illustrate, reverse-type questions about $\mathbf{E}$ are difficult to resolve numerically: The question, "for what value of $p$ does $\mathbf{E} = 1/\pi$?" poses a reasonable computational challenge. Presumably one would have to sample points in the range of $p$, estimate $\mathbf{E}$ for each, and through interpolation and possibly some manner of successive approximation, hone in on the correct value. This is of course possible, but completely non-generalizable. To examine some other $\mathbf{E}$ value might require resampling a different region of $p$ values. Furthermore, the addition of a single state would require starting the whole procedure from scratch.

We would also like to note that the ability to even represent the relative error in the excess entropy estimates, as in Fig. 1.12, is only made possible by making use of the algorithm we developed to determine the exact value of $\mathbf{E}$. Before access to the limit that our estimates were al-

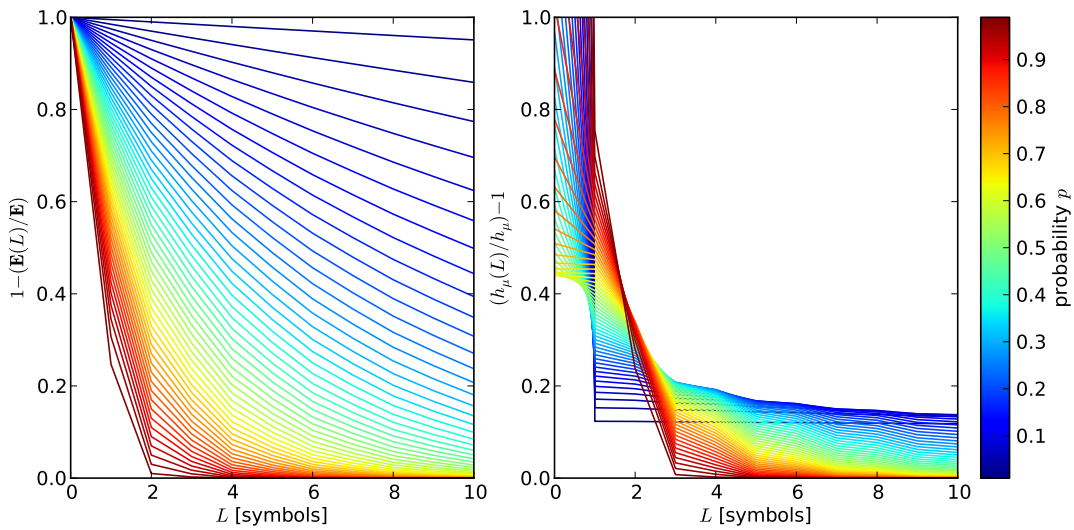


Figure 1.12: The non-Markovianness of the Even Process leads to some members of the family having very slow convergence. (Left) Relative errors in the excess entropy estimates show that even considering correlation lengths up to 10 is grossly inadequate for a large collection of processes. (Right) Relative error of entropy rate estimates are very slow to approach zero for members on the blue end of the spectrum. This process serves as a key motivating example in the search for analytic forms for **E**.

legedly approaching, excess entropy estimate plots were much more undetermined. Figure 1.13 demonstrates the slow, indeterminate growth of the estimates. There is of course the bound from **shalizi** $\mathbf{E} \leq C_\mu$, but from this simulation, there is no clear way to bound any instance away from $C_\mu$ at all.

In Ch. **??** we present our method for calculating **E** for an entire parameterized family of $\epsilon$-machines at once. Moreover, this method is finitely terminating even for infinite order Markov processes [21]. To put some concreteness to the technique, it can be rapidly calculated by hand that the solution to the above challenge is the result of this equation,

$$\frac{1}{\pi} = \log(2-p) - \frac{1-p}{2-p}\log(1-p).$$

[21] There are some questions remaining as to what happens in the case of infinite transient states. It appears that when the recurrent states are reachable, that this algorithm will be finite despite the infinite transients. The algorithm needs only to reach all recurrent states (not to reach them via all possible transient paths). When the recurrent states are not reachable, it is known that in some cases, one can define an infinite sequence that converges to the correct result and find the limit analytically. It is hoped that this procedure can be made general. As transient states and infinite states are not discussed here, the reader should look for results in the upcoming Ref. **cite Extension of E algo to infinite**

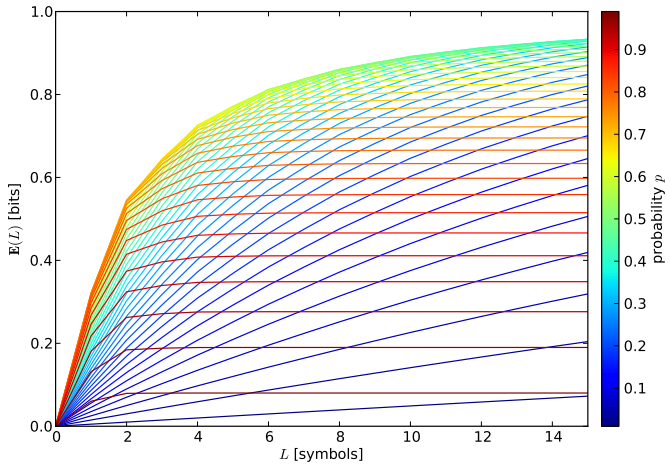


Figure 1.13: Excess entropy estimates for the Even Process without access to the actual limit **E**. Its estimates increase in a very slow manner making claims about convergence, except for very trivial ones, difficult.

## §1.8 Crypticity and Cryptic Order

The study of the structure of stochastic processes through their $\epsilon$-machine representations has lead to the the recognition of two new and important quantities: the crypticity, and the cryptic order. Intuitively, these two ideas spring from focusing one's attention not on the information region associated most directly with prediction—the excess entropy or predictive information—but rather with the region that characterizes the information above and beyond **E** necessary for determining the causal state, and thereby for making predictions. This is what we call the crypticity.

### §1.8.1 Crypticity

**Definition.** *The* crypticity *,* $\chi$ *, of a process is defined,*

$$\chi = H[\mathcal{S}_0 | \overrightarrow{X}_0]$$

We represent this quantity in our I-diagram as the difference between the statistical complexity and the excess entropy (see Fig. 1.14). At first, it might seem as though the definition of the $\epsilon$-machine ought to obviate any information except for that which is predictive information. The $\epsilon$-machine is, after all, the causal representation of the process. How can we reconcile these intuitions? The essential idea is this: optimal prediction, which is what causal states are built for,

requires not only the ability to match up histories with the appropriate future, or set of futures; it also requires the ability to match up histories with the appropriate distribution over futures, and *these pairings can overlap*.

A positive crypticity means that despite all your hard work in noting the relations between pasts and futures, and determining which class of pasts you are in, there exists a particular future which can follow more than one class (even all classes) of pasts. Supposing that future is realized, you might wish you had been less careful, as the result might[22] have been the same. To say this a little differently, and mathematically,

$$\underbrace{H[X_0^L] - H[X_0^L|\mathcal{S}_0]}_{\text{net earnings}} = \underbrace{H[\mathcal{S}_0]}_{\text{gross earnings}} - \underbrace{H[\mathcal{S}_0|X_0^L]}_{\text{taxes}}.$$

To expand upon this interpretation a little, the 'gross earnings' is the amount that enters consideration. The 'net earnings' is the amount of useful resource. Trivially the, the difference is what is given up in 'taxes'. This analogy is appealing and correct in that only in very rare cases can you get away with paying no taxes [23].

It is the last term—taxes—which is our crypticity. To assure the reader that this information waste is not just a corner-case, note that the Golden Mean, a process we have already introduced, has a crypticity $\chi = 2/3$, a significant fraction of the total stored information, $C_\mu = \log(3) - 2/3 \simeq 0.918$.

## §1.8.2 Cryptic Order

As the crypticity is a newly defined quantity, it is natural to attempt to tease it apart in ways similar to quantities we have dealt with in the past. The crypticity can be interpreted as the state-based companion information to the predicive information, **E**. That said, the primary dissection tool used to understand processes by thinking about their predictive information has been the Markov order. The Markov order describes the length scale of the correlations among symbols that give rise to probabilistic conditional independence; this independence is another way of describing an optimal predictor.

[22] There's the rub.

[23] It is not claimed that this is a deep analogy, but the 'economics of information' is an attractive thought; it suggests competition and optimization. It also encourages us to search for off-shore [quantum] information accounts. Other analogies have been explored for the crypticity, most notably in the context of heat engines. There, the statistical complexity is likened to heat transference and excess entropy to the derived work. The ratio then is a measure of the 'thermodynamic efficiency of the machine. There is much work yet to be done to firmly establish an economic or thermodynamic relationship.

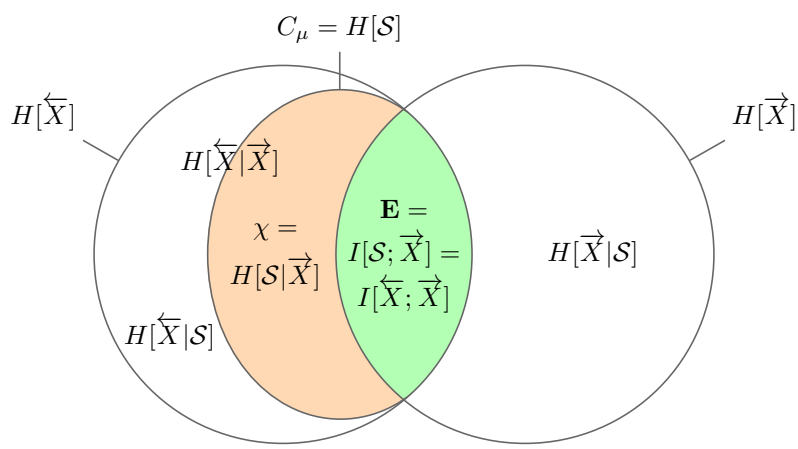


Figure 1.14: This highlights the crypticity $\chi$ in orange as the difference between the state information $C_\mu$ and the predictive information $\mathbf{E}$. In this sense, crypticity can be thought of as 'modeling overhead'.

To begin, it is useful to have a picture of what Markov order is in terms of our I-diagram. We can stratify the past in terms of the random variables $\{X^1_{-1}, X^2_{-2}, X^3_{-3}, ...\}$. The intersection of this stratification with the future is shown in Fig. 1.15. For the details as to why intersection with a stratification is allowed in this way, and for why it is also non-trivial, see App. B. One feature of $\epsilon$-machines is illustrated by the fact that an equivalent way of understanding Markov order is the depth of history required for determining the causal state. It is this which speaks to the fundamental nature of causal states, and which allows us to make the statement that the cryptic order is a 'companion' order.

Cryptic order is a new length scale introduced to characterize the way in which the information associated with crypticity is distributed in the process. We argue that the cryptic order is as fundamental to the nature of processes as the Markov order. It has a slightly different flavor in that it involves causal states, whereas Markov order can be defined without them. For an illustration as to how the cryptic and Markov orders can be different, see Fig. 1.17. The cryptic order is the length scale appropriate for capturing the uncertainty in the causal state *given the future*—it captures the crypticity.

Properties of the crypticity and the cryptic order are the subjects of Ch. **??**. This collection of definitions and proofs marks the beginning of a new and fundamental characterization of stochastic process. This characterization is thought to be fundamental as it is such a close analog to the Markov order. It is deemed impactful because, unlike the Markov order, it makes reference

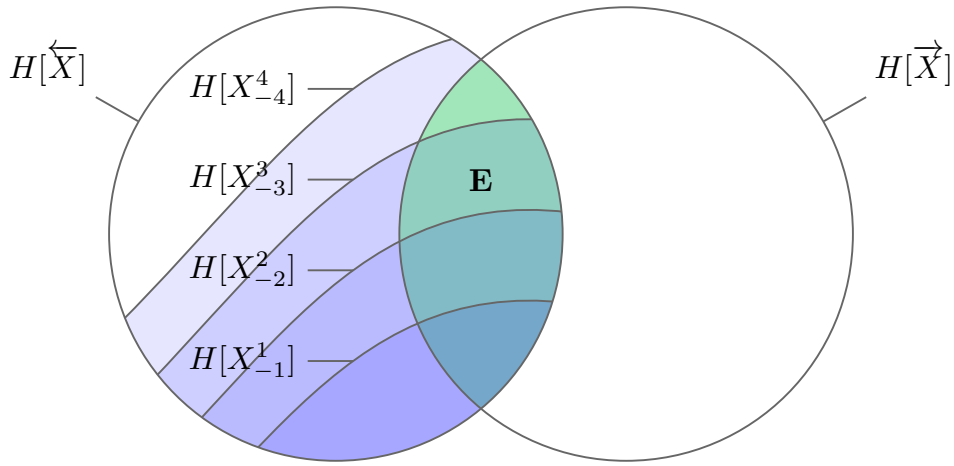


Figure 1.15: An illustration of a process which is order-4 Markov. The past $H[\overleftarrow{X}]$ is shown as being stratified in the standard way. We can see that conditioning on the past 4 variables reduces as much uncertainty in the future as does conditioning on the entire past. Conditioning on only the past 3 variables, however, neglects the upper tip of the mutual information, $\mathbf{E} = I[\overleftarrow{X}; \overrightarrow{X}]$.

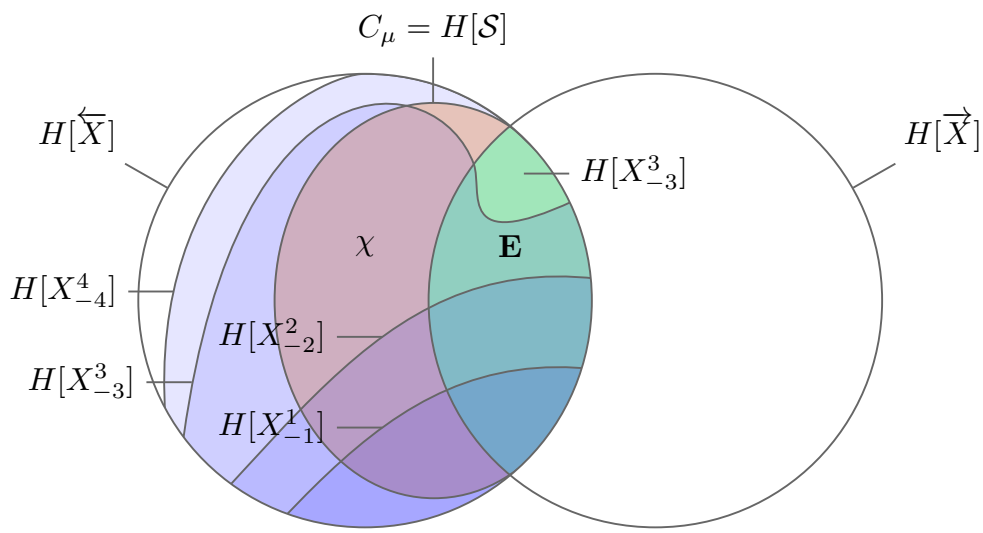


Figure 1.16: This is another illustration of an order-4 Markov process. The causal state has been added to the diagram and the boundaries made a little more curvy to anticipate future I-diagrams. Notice that in addition to the length 4 statistics being sufficient for capturing $\mathbf{E}$, the same is true for capturing $\chi$ which is the remainder of $C_\mu$. In contrast, the length 3 statistics are insufficient for both $\mathbf{E}$ and $\chi$. Being insufficient for $\mathbf{E}$ is why the process is order-4 Markov. Being insufficient for $\chi$ is why the process is order-4 cryptic.

to states, something that the generically non-zero crypticity strongly suggests we do. Also, the states refered to are not any state, but causal states, and so the naturalness of the $\epsilon$-machine in its ability to deliver quantities such as $h_\mu$ and $C_\mu$ extends this naturalness to the cryptic order.

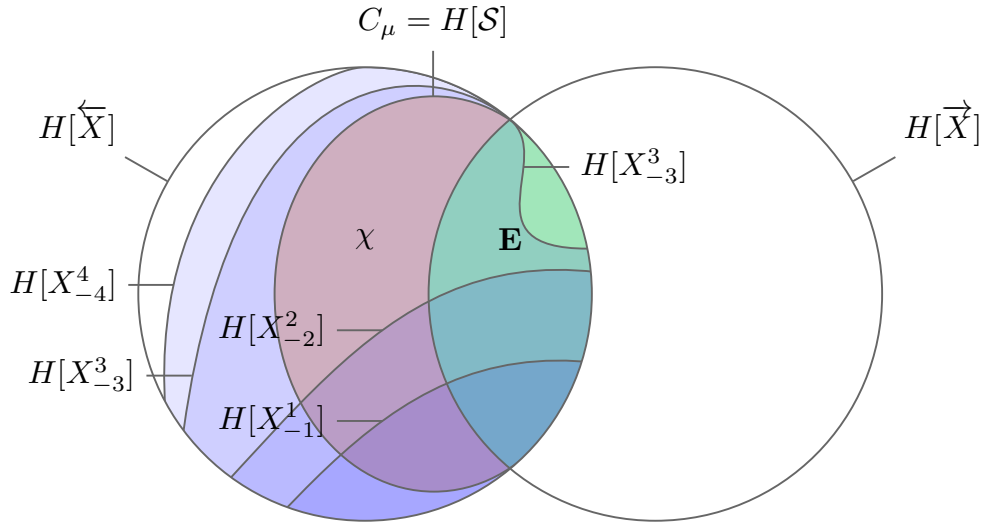


Figure 1.17: An illustration of a process with differing cryptic and Markov orders. The Markov order is 4; this is the first history length which contains all of the predictive information. Notice that the length 3 history curves back again missing a portion of **E**. The cryptic order is 3 because although the length 3 history misses some portion of **E**, it does determine the causal state conditioned on the future. Note that $H[X^3_{-3}]$ is labeled twice for clarity.

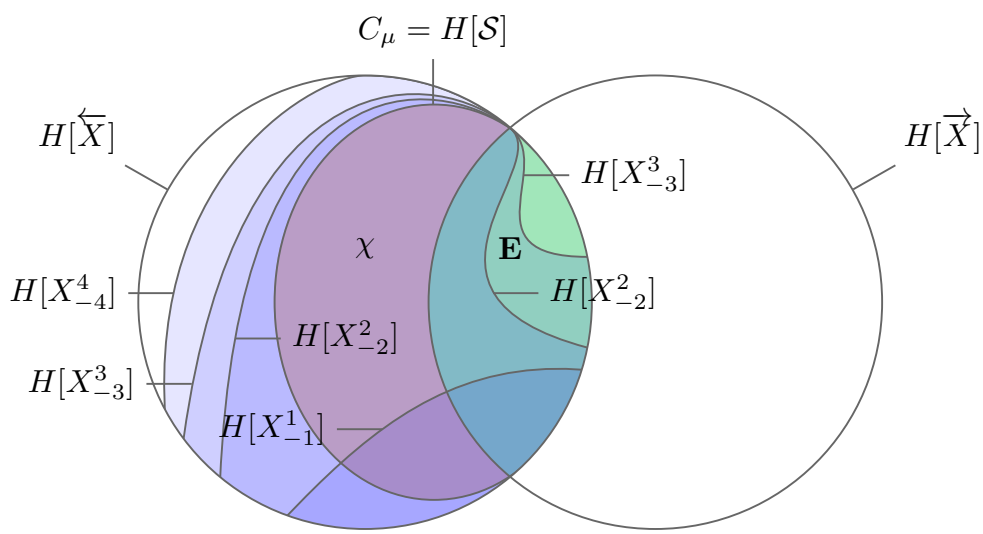


Figure 1.18: The Markov and cryptic orders may differ by more than one. This is an instance where the Markov order is 4, yet the cryptic order is 2. Two entropies are labeled twice for clarity.

CHAPTER 2

# Prediction, Retrodiction and the Amount of Information Stored in the Present

"Life can only be understood backwards, but it must be lived forwards." - S. Kierkegaard

## §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 extend *computational mechanics* [**?**, **?**] with 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. The importance of the excess entropy as a quantifier of stochastic processes has already been emphasized.

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.[1] 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 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 $\epsilon^+$ 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 $\epsilon$-machine $M^+$ from the reverse-scan $\epsilon$-machine $M^-$. They assign corresponding entropy rates, $h_\mu^+$ and $h_\mu^-$, and statistical complexities, $C_\mu^+ = H[\mathcal{S}^+]$ and $C_\mu^- = H[\mathcal{S}^-]$, respectively, to the process.

[1] There is a good puzzle here. While it is straightforward to show how the time-symmetric representation produces the correct forward and reverse processes—it projects onto them—it is not clear that the time-symmetric representation can be obtained through those constraints alone, even *given* the target dimension of the bidirectional machine. In fact, an analysis of the benign Golden Mean Process should illustrate this. It seems reasonable that the unspecified degrees of freedom will be well understood in the context of **cite SyncControl**—which describes the information quantities associated with various presentations of the same process. It will be interesting to know what the additional information is, and if we can then understand why the projections are not completely specifying.

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 both directions are equally predictable (equivalently, equally surprising):

**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:*

$$H[X_{-L+2}, \ldots, X_0] = H[X_{-L+1}, \ldots, X_{-1}] \,. \tag{2.1}$$

*Keeping this in mind, we directly calculate:*

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

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.*

This is a somewhat curious result that is worth absorbing. Note that **E** is mute on the prediction vs. retrodiction 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. In fact, it is the potential asymmetry here that opens the door for a new measure introduced later.

## §2.3 Excess Entropy from Causal States

Let us return to the excess entropy as a point of entry for the employment of our prediction / retrodiction machinery. Having this foothold will allow us to complete the calculation of all new quantities introduced here.

Until recently, **E** could not be directly calculated from the $\epsilon$-machine— in contrast to 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—$h_\mu$, $C_\mu$, and **E**. Ref. [**?**] announced the solution to this long-standing problem by deriving explicit expressions for **E** in terms of the $\epsilon$-machine, providing a unified information-theoretic analysis of general processes. Here we provide a detailed account of the underlying methods and results.

We should briefly recall what is already known about the relationships between these various quantities, specifically those relevant to **E**. First, some time ago, an explicit expression was developed from the Hamiltonian for one-dimensional spin chains with range-$R$ interactions [**?**]:

$$\mathbf{E} = C_\mu - R\, h_\mu \,. \tag{2.2}$$

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 [**?**]:

$$\mathbf{E} \leq C_\mu \,. \tag{2.3}$$

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.

The relationship between predicting and retrodicting a process, and ultimately **E**'s role, requires teasing out how the states of the forward and reverse $\epsilon$-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.[2] Practically, there is a direct correspondence between set theory and these information measures. Using this, Ref. [**?**] developed an *$\epsilon$-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 $h_\mu$, $C_\mu$, and **E** for general stochastic processes.

In a generic four-variable information diagram, there are 15 independent quantities. These quantities can be seen in Fig. 2.1 as atoms, or regions of the I-diagram. Fortunately, this greatly simplifies in the case of using predictive and retrodictive $\epsilon$-machines to represent the process; there are only 5 independent variables in this special case (see Fig .2.2). Reference [**?**] contains more details of this reduction. Here we present the main ideas.

The first attack on Fig. 2.1 is using the fact that causal states are a function of the infinite past. That is, each infinite past induce one and only one causal state.[3] Moreover, additional conditioning cannot reduce this (complete lack of) uncertainty any further. The following 4 independent equations are the consequences.

[2] See App. B for more background on the relatioship between information theory and Venn diagrams.
[3] This is not true for all representations. See **sync control** for more details.

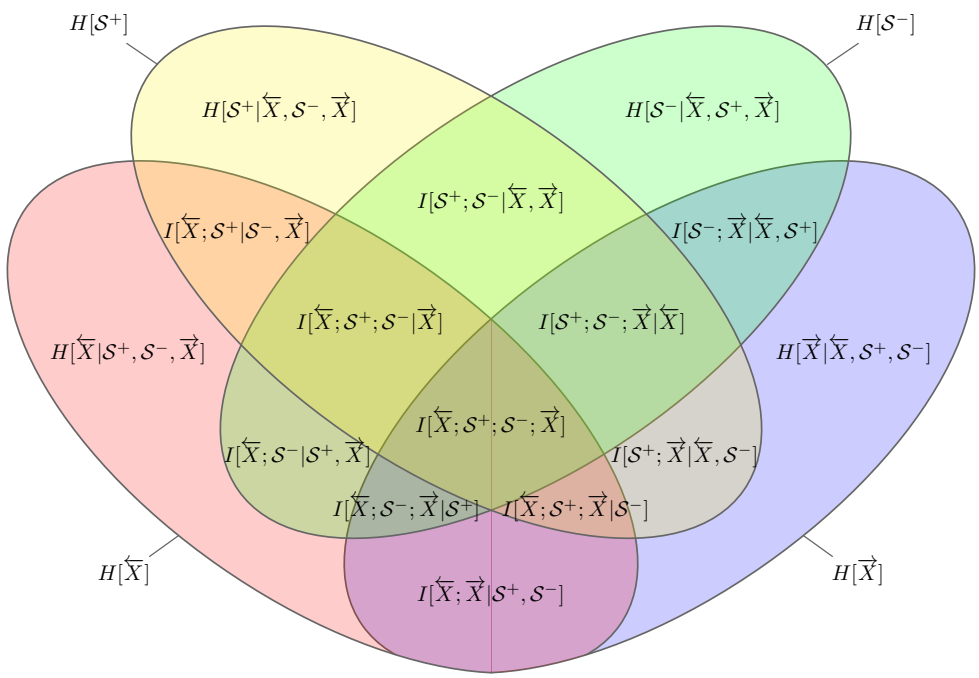


Figure 2.1: The generic (un-reduced) I-diagram for 4 random variables, where the names of the variables of interest have been inserted.

$$H[\mathcal{S}^+|\overleftarrow{X}] = 0$$

$$H[\mathcal{S}^+|\overleftarrow{X},\overrightarrow{X}] = 0$$

$$H[\mathcal{S}^+|\overleftarrow{X},\mathcal{S}^-] = 0$$

$$H[\mathcal{S}^+|\overleftarrow{X},\overrightarrow{X},\mathcal{S}^-] = 0$$

Using the I-diagram for reference, we can see that these four constraints reduce the four independent quantities: $H[\mathcal{S}^+|\overleftarrow{X},\mathcal{S}^-,\overrightarrow{X}]$, $I[\mathcal{S}^+;\mathcal{S}^-|\overleftarrow{X},\overrightarrow{X}]$, $I[\mathcal{S}^+;\mathcal{S}^-;\overrightarrow{X}|\overleftarrow{X}]$, $I[\mathcal{S}^+;\overrightarrow{X}|\overleftarrow{X},\mathcal{S}^-]$ each to zero.

The time reversed analysis proceeds identically finding the four quantities: $H[\mathcal{S}^-|\overleftarrow{X},\mathcal{S}^+,\overrightarrow{X}]$, $I[\mathcal{S}^+;\mathcal{S}^-|\overleftarrow{X},\overrightarrow{X}]$, $I[\overleftarrow{X};\mathcal{S}^+;\mathcal{S}^-|\overrightarrow{X}]$, $I[\overleftarrow{X};\mathcal{S}^-|\mathcal{S}^+,\overrightarrow{X}]$ to be zero. One of these, $I[\mathcal{S}^+;\mathcal{S}^-|\overleftarrow{X},\overrightarrow{X}]$, is accounted for twice. We have now removed 7 atoms from the diagram. A significant improvement, but there is more to go.

Since the predictive causal states predict as well as the pasts that produce them, and similarly for the retrodictive states, we have,

$$I[\overleftarrow{X};\overrightarrow{X}] = I[\mathcal{S}^+|\overrightarrow{X}]$$

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

These lead us to the following constraints on our atoms,

$$I[\overleftarrow{X};\mathcal{S}^-;\overrightarrow{X}|\mathcal{S}^+] + I[\overleftarrow{X};\overrightarrow{X}|\mathcal{S}^+,\mathcal{S}^-] = 0$$

$$I[\overleftarrow{X};\mathcal{S}^+;\overrightarrow{X}|\mathcal{S}^-] + I[\overleftarrow{X};\overrightarrow{X}|\mathcal{S}^+,\mathcal{S}^-] = 0.$$

But since $I[\overleftarrow{X};\overrightarrow{X}|\mathcal{S}^+,\mathcal{S}^-]$ is a conditional mutual information and is positive semidefinite, we have that all three quantities, $I[\overleftarrow{X};\mathcal{S}^-;\overrightarrow{X}|\mathcal{S}^+]$, $I[\overleftarrow{X};\mathcal{S}^+;\overrightarrow{X}|\mathcal{S}^-]$, $I[\overleftarrow{X};\overrightarrow{X}|\mathcal{S}^+,\mathcal{S}^-]$ are zero.

This leaves us with the following elegant description of the dependences among forward and reverse $\epsilon$-machines and the processes they model (see Fig. 2.2).

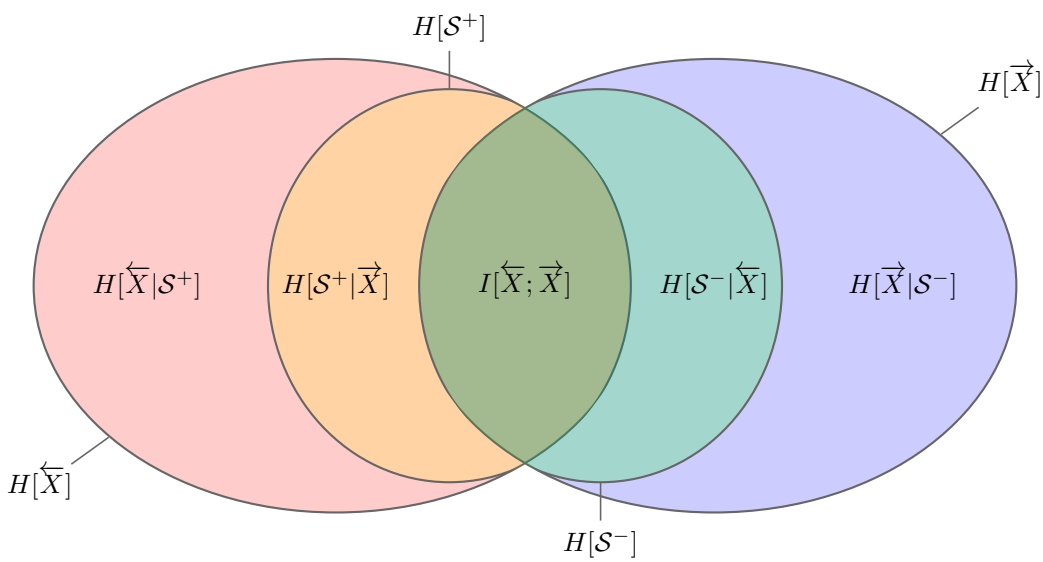


Figure 2.2: The I-diagram for the forward and reverse $\epsilon$-machines. Only 5 of the 15 independent information quantities remain. This image is a central reference for the work following.

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:*

$$\mathbf{E} = I[\mathcal{S}^+;\mathcal{S}^-]\,. \tag{2.4}$$

**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*

$$\begin{aligned} I[\overleftarrow{X}; \overrightarrow{X}; \mathcal{S}^+; \mathcal{S}^-] &= I[\overleftarrow{X}; \overrightarrow{X}] \\ &= \mathbf{E}\,, \end{aligned} \tag{2.5}$$

*on the one hand, and*

$$I[\overleftarrow{X}; \overrightarrow{X}; \mathcal{S}^+; \mathcal{S}^-] = I[\mathcal{S}^+; \mathcal{S}^-]\,, \tag{2.6}$$

*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 $\epsilon$-machine states.

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

$$C_\mu^+ = \mathbf{E} + H[\mathcal{S}^+|\mathcal{S}^-] \text{ and} \tag{2.7}$$

$$C_\mu^- = \mathbf{E} + H[\mathcal{S}^-|\mathcal{S}^+]\,. \tag{2.8}$$

**Proof.** $\mathbf{E} = I[\mathcal{S}^+; \mathcal{S}^-] = H[\mathcal{S}^+] - H[\mathcal{S}^+|\mathcal{S}^-]$. *Since the first term is* $C_\mu^+$*, 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 $\epsilon$-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.

## §2.4 The Bidirectional Machine

At this point, we have two separate $\epsilon$-machines—one for predicting ($M^+$) and one for retrodicting ($M^-$). We will now show that one can do better[4], by simultaneously utilizing causal information from the past and future.

**Definition.** *Let $M^\pm$ denote the* bidirectional machine *given by the equivalence relation $\sim^\pm$ [5]:*

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

*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 $\epsilon$-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:

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 $\epsilon$-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 $\mathcal{S}_t^\pm = (\epsilon^+(\overleftarrow{x}_t), \epsilon^-(\overrightarrow{x}_t))$. 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

[4] What we mean by *better* is that the two models are not independent from each other, and therefore can be compressed. We discuss this compression in a later section.

[5] Interpret the symbol $\pm$ as "plus *and* minus".

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 $\sigma^-$ that is not in the range of $r(\mathcal{S}^+)$. This means that no forward causal state is paired with $\sigma^-$ 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, $\sigma^-$ 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:*

$$C_\mu^{\pm} \equiv H[\mathcal{S}^{\pm}] = H[\mathcal{S}^+, \mathcal{S}^-] \,. \tag{2.9}$$

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 $C_\mu$s 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.

### §2.4.1 Upper Bounds

Here we give new, tighter bounds for $\mathbf{E}$ than Eq. (2.3) 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* $C_\mu^-$. *Similarly, since* $I[\mathcal{S}^-;\mathcal{S}^+] \leq H[\mathcal{S}^+]$, *we have* $\mathbf{E} \leq C_\mu^+$. □

### §2.4.2 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:*

$$\Xi(\mathcal{P}) = C_\mu^+ - C_\mu^- \ . \tag{2.10}$$

**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.

### §2.4.3 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:*

$$\chi^{\pm}(M^{+}, M^{-}) = H[\mathcal{S}^{+}|\mathcal{S}^{-}] + H[\mathcal{S}^{-}|\mathcal{S}^{+}] \,. \tag{2.11}$$

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

$$C_{\mu}^{\pm} = \mathbf{E} + \chi^{\pm} \,. \tag{2.12}$$

**Proof.** *This follows directly from the corollary and the predictive and retrodictive statistical complexity relations, Eq. (2.7) and (2.8).* □

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 ($\mathbf{E}$). 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:*

$$C_{\mu}^{\pm} \geq \chi^{\pm} \,. \tag{2.13}$$

**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 $\epsilon$-machine representation of a process; see Ref. [**?**] and Ref. [**?**]. They also suggest a classification scheme based on crypticity, to complement the Markov-order classification; see Ref. [**?**]. In the following, we phrase the calculation in terms of $\mathbf{E}$, and $\chi^{+}$, $\chi^{-}$, $\chi^{\pm}$, $C_{\mu}^{\pm}$, and $\Xi$ follow straightforwardly.

## §2.5 Alternative Presentations

The $\epsilon$-machine is a process's unique, minimal unifilar presentation. Now we introduce two alternative presentations, which need not be $\epsilon$-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 $t$.

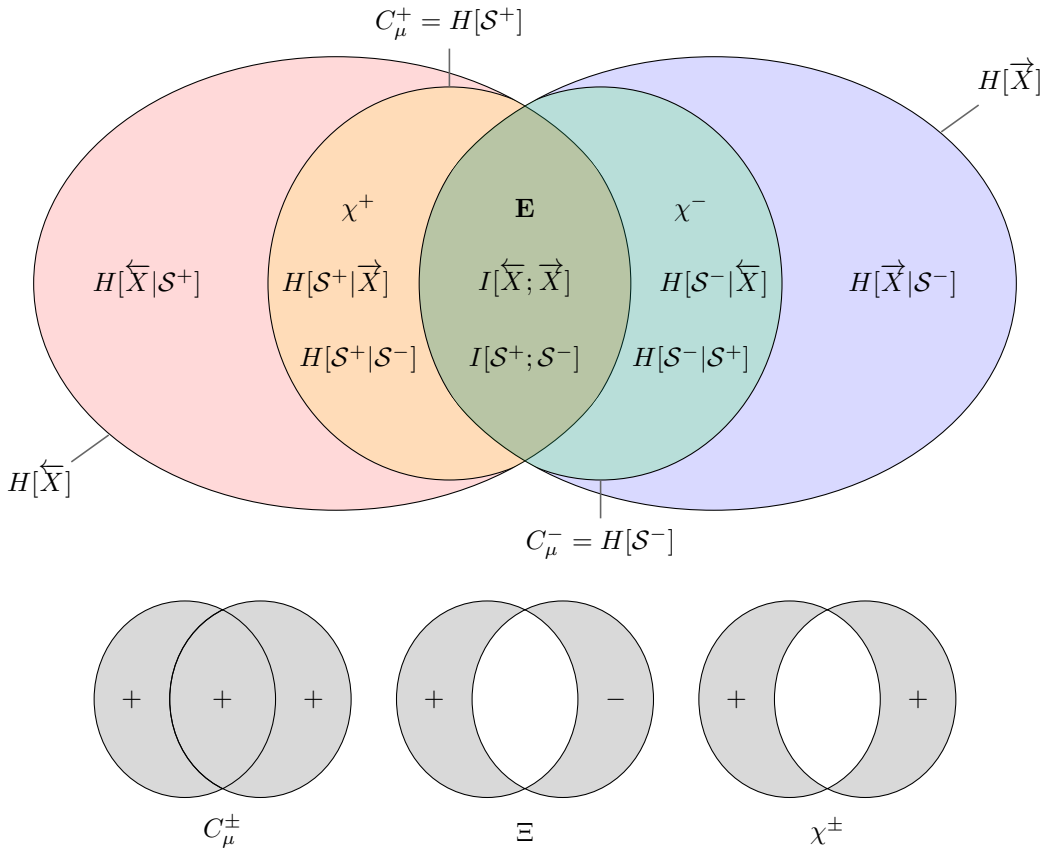


Figure 2.3: This diagram summarizes the measures and relationships derived in this chapter. The upper part of the figure should already be familiar—some relationships have been added. The bottom three icons illustrate which portions of the above diagram are added (or subtracted) to obtain the three newly defined measures: $C_\mu^\pm$, $\Xi$, and $\chi^\pm$. These represent the process's bidirectional information storage, irreversibility, and information overhead, respectively.

### §2.5.1 Time-Reversed Presentation

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

$$T^{(x)}_{\mathcal{R}\mathcal{R}'} \equiv \Pr(X = x, \mathcal{R}'|\mathcal{R})\,. \tag{2.14}$$

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

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:

$$\begin{aligned} \widetilde{T}^{(x)}_{\mathcal{R}'\mathcal{R}} &\equiv \Pr(X = x, \mathcal{R}|\mathcal{R}') \\ &= T^{(x)}_{\mathcal{R}\mathcal{R}'} \frac{\Pr(\mathcal{R})}{\Pr(\mathcal{R}')} \,. \end{aligned} \tag{2.15}$$

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 $\pi$ of $M$.*

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

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

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

## §2.5.2 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 $w$:

$$\mu(w) \equiv \Pr(\mathcal{R}_L | X_0^L = w) \tag{2.16}$$

$$= \frac{\Pr(X_0^L = w, \mathcal{R}_L)}{\Pr(X_0^L = w)} \tag{2.17}$$

$$= \frac{\pi T^{(w)}}{\pi T^{(w)} \mathbf{1}}, \tag{2.18}$$

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:

$$\Pr(x, \mu(wx) | \mu(w)) \equiv \frac{\Pr(wx)}{\Pr(w)} \tag{2.19}$$

$$= \mu(w) T^{(x)} \mathbf{1} \,. \tag{2.20}$$

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* $\mathcal{U}(M)$.

## §2.6 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 $\epsilon$-machines, yields a direct calculation of $\mathbf{E}$ 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:

$$\begin{array}{ccc} M^+ & \xleftarrow{\mathcal{U}} & \widetilde{M}^- \\ \mathcal{T}\big\downarrow & & \big\uparrow\mathcal{T} \\ \widetilde{M}^+ & \xrightarrow[\mathcal{U}]{} & M^- \end{array} \tag{2.21}$$

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 $\epsilon$-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 $\epsilon$-machine. We leave it as a conjecture that $\mathcal{U}(\mathcal{T}(M))$ is always minimal. Even so, App. D 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 $\nu(w)$:

$$\nu(w) \equiv \Pr(\mathcal{R}_0 | X_0^L = w) \tag{2.22}$$

$$= \frac{\Pr(\mathcal{R}_0, X_0^L = w)}{\Pr(X_0^L = w)} \tag{2.23}$$

$$= \frac{\pi T^{(\widetilde{w})}}{\pi T^{(\widetilde{w})} \mathbf{1}}, \tag{2.24}$$

where $\pi$ 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:

$$\Pr(x, \nu(xw) | \nu(w)) \equiv \frac{\Pr(xw)}{\Pr(w)} \tag{2.25}$$

$$= \nu(w) T^{(x)} \mathbf{1}. \tag{2.26}$$

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 $M^+$:

$$\begin{aligned}\nu(w) &= \Pr(\mathcal{R}_0 | X_0^L = w) \\ &= \Pr(\mathcal{S}_0^+ | X_0^L = w)\,.\end{aligned}$$

Whenever the mixed-state presentation is an $\epsilon$-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,

$$\begin{aligned}\nu(w) &= \Pr(\mathcal{S}_0^+ | X_0^L = w) \\ &= \Pr\left(\mathcal{S}_0^+ | \mathcal{S}_0^- = \epsilon^-(w)\right)\,.\end{aligned}$$

Hence, we can calculate $H[\mathcal{S}^+|\mathcal{S}^-]$ and obtain $\mathbf{E}$ via (2.4).

## §2.7 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 $\epsilon$-machine has three recurrent causal states $\boldsymbol{\mathcal{S}}^+ = \{A, B, C\}$ and is shown in Fig. 2.4(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. (**??**) and their associated future conditional distributions $\Pr(\overrightarrow{X} \mid \overleftarrow{x})$ to obtain the $\epsilon$-machine causal states and transitions.)

These matrices are used calculate the stationary distribution $\pi$ 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 $\pi$, we create the time-reversed presentation $\widetilde{M}^+ = \mathcal{T}(M^+)$. This is shown in Fig. 2.4(b). Notice that the machine is not unifilar, and so it is clearly not an $\epsilon$-machine. The transition matrices for the time-reversed presentation are given by:

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

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

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.

Now we are in a position to calculate the mixed-state presentation, $M^- = \mathcal{U}(\widetilde{M}^+)$, shown in Fig. 2.4(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

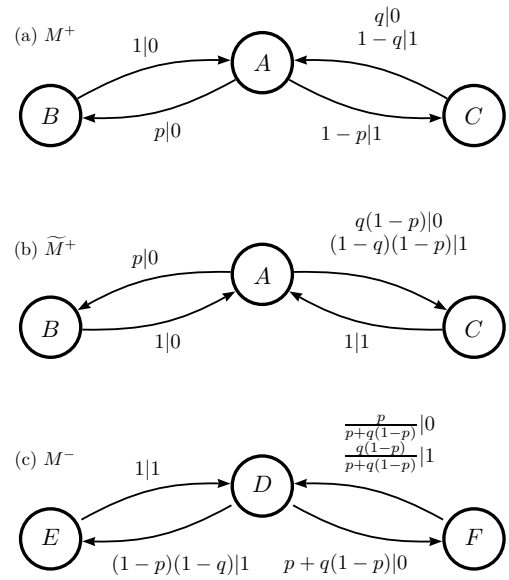


Figure 2.4: The presentations used to calculate the excess entropy for the RnC Process: (a) $M^+$, (b) $\widetilde{M}^+ = \mathcal{T}(M^+)$, and (c) $M^- = \mathcal{U}(\widetilde{M}^+)$. Edge labels $t\,|x$ give the probability $t = T^{(x)}_{\mathcal{R}\mathcal{R}'}$ of making a transition and seeing symbol $x$.

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

$$\nu(10) = \frac{\nu(0)\widetilde{T}^{(1)}}{\nu(0)\widetilde{T}^{(1)}\mathbf{1}} . \tag{2.27}$$

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 $\nu(100)$ [6]. By Eq. (2.26), we have

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

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 $\boldsymbol{\mathcal{S}}^- = \{D, E, F\}$:

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

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

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

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

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.

## §2.8 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.

### §2.8.1 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 $\epsilon$-machine representation. See Figure 3.3(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 $\epsilon$-machine has two recurrent causal states $\boldsymbol{\mathcal{S}}^+ = \{A, B\}$ and transition matri-

ces [**?**]:

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

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

Figure 3.3(a) gives $M^+$, while 3.3(b) gives $M^-$. We see that the $\epsilon$-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 $\mathbf{E} = C_\mu^\pm$.

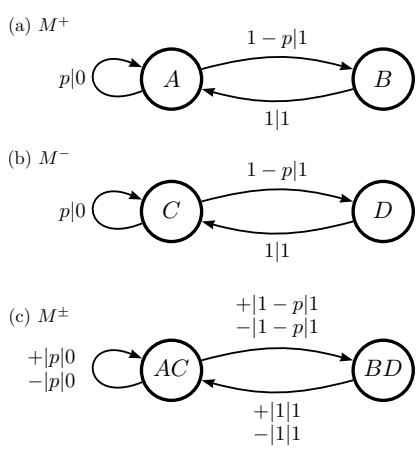


Figure 2.5: Forward and reverse $\epsilon$-machines for the Even Process: (a) $M^+$ and (b) $M^-$. (c) The bidirectional machine $M^\pm$. Edge labels are prefixed by the scan direction $\{-,+\}$.

Now, let's analyze its bidirectional machine, which is shown in Fig. 3.3(c). The reverse and

forward maps are given by:

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

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

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.

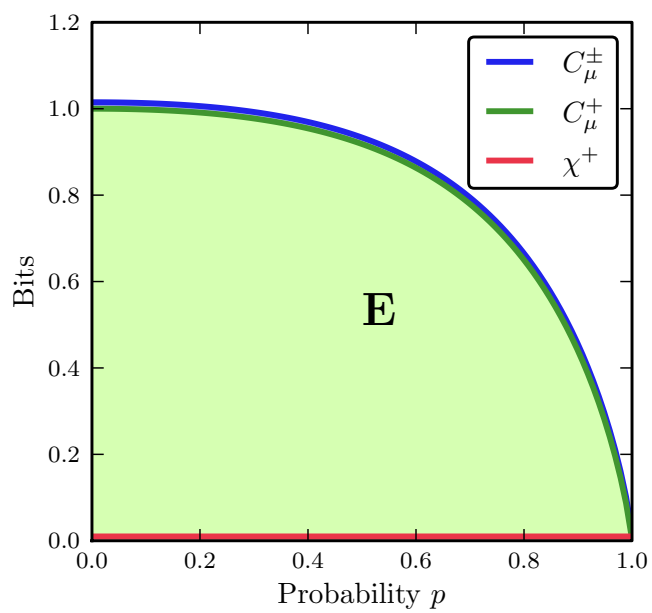


Figure 2.6: The Even Process's information processing properties—$C_\mu^\pm$, $C_\mu^+$, and $\chi^+$—as its self-loop probability $p$ varies. The colored area bounded by the curves show the magnitude of $\mathbf{E}$.

Without going into details to be reported elsewhere, the Even Process is also notable since it is difficult to empirically estimate its $\mathbf{E}$. (The convergence as a function of the number of measurements is extremely slow.) Viewed in terms of the quantities $C_\mu^+$, $C_\mu^-$, $\chi^+$, $\chi^-$, and $\Xi$, 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. 2.6) 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 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. 3.3.

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 $C_{\mu}^{+}$, $C_{\mu}^{\pm}$, and $\chi^{+}$, Fig. 2.6 rather directly illustrates these properties and shows that they are maintained across the entire process family as the self-loop probability $p$ is varied.

### §2.8.2 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 $\varphi$ 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 $\epsilon$-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 $\epsilon$-machine has two recurrent causal states $\mathcal{S}^+ = \{A, B\}$ 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 2.7(a) gives $M^+$, while (b) gives $M^-$. We see that the $\epsilon$-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 $\epsilon$-machines.

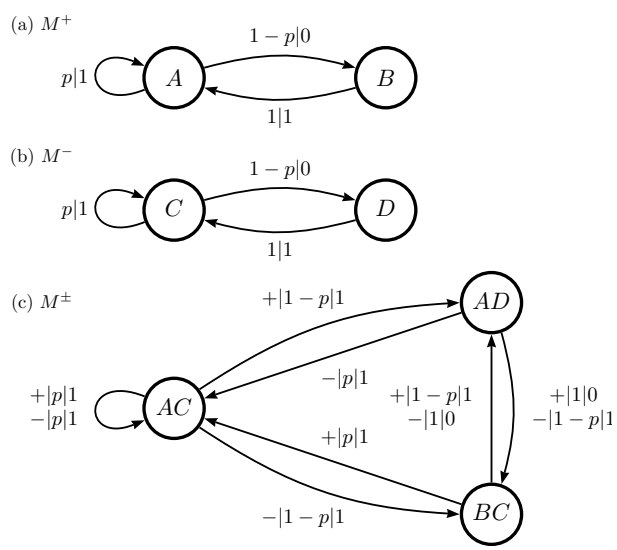


Figure 2.7: Forward and reverse $\epsilon$-machines for the Golden Mean Process: (a) $M^+$ and (b) $M^-$. (c) The bidirectional machine $M^\pm$.

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. (2.2).) Thus,

$$\mathbf{E} = H\left(\frac{1}{2-p}\right) - \frac{H(p)}{2-p}\,. \tag{2.28}$$

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. 2.8.

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

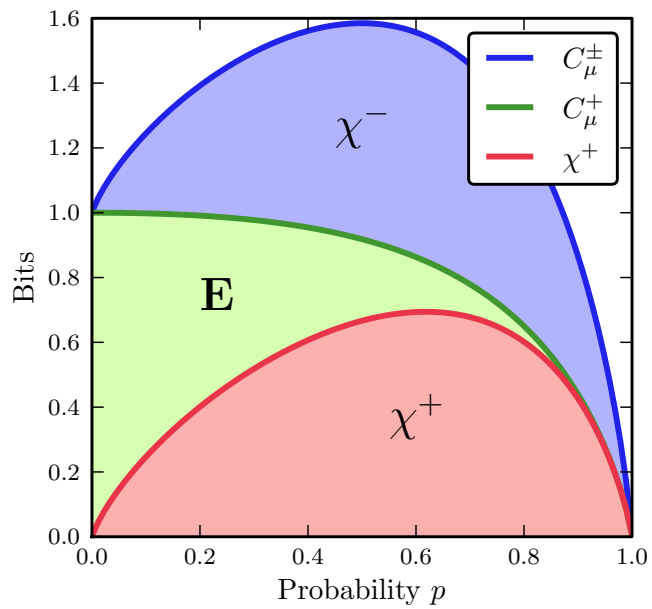


Figure 2.8: The Golden Mean Process's information processing measures—$C_\mu^{\pm}$, $C_\mu^{+}$, and $\chi^{+}$—as its self-loop probability $p$ varies. Colored areas bounded by the curves give the magnitude at each $p$ of $\chi^{-}$, $\mathbf{E}$, and $\chi^{+}$.

The origin of its crypticity is found by analyzing the bidirectional machine, which is shown in Fig. 2.7(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: $\Pr(\mathcal{S}^{\pm}) = \Pr(AC, AD, BC) = \big(p, 1-p, 1-p\big)/(2-p)$. 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 $C_{\mu}$s, 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, 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 $\chi^{+}$ bits of memory and a generation of $\chi^{-}$ 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 2.8 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: $C_{\mu}^{\pm} \approx \chi^{+} + \chi^{-}$.

### §2.8.3 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 $\epsilon$-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 2.9(b) shows $M^-$ which has four recurrent causal states $\boldsymbol{\mathcal{S}}^- = \{D, E, F, G\}$. We see that the $\epsilon$-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.

Let's analyze the RIP bidirectional machine, which is shown in Fig. 2.9(c) for $p = q = 1/2$.

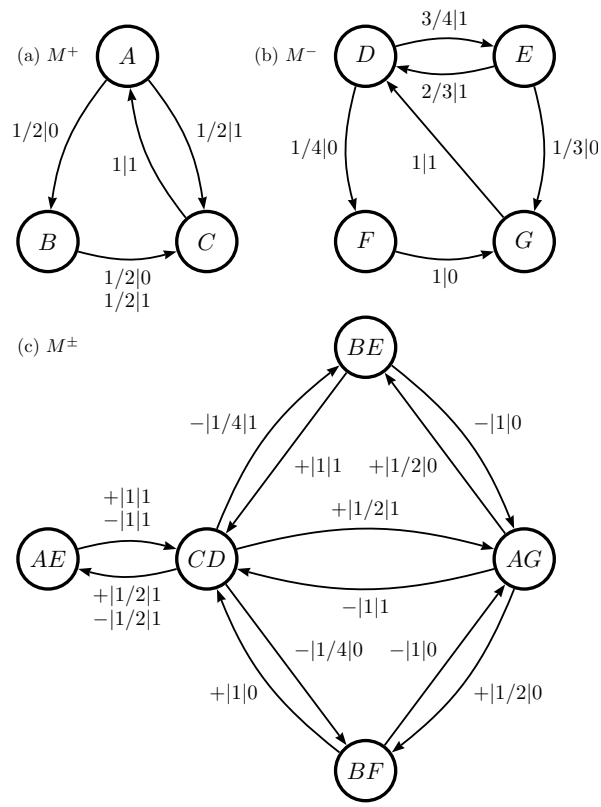


Figure 2.9: Forward and reverse $\epsilon$-machines for the RIP with $p = q = 1/2$: (a) $M^+$ and (b) $M^-$. (c) The bidirectional machine $M^\pm$ also for $p = q = 1/2$. (Reprinted with permission from Refs. [**?**].)

The reverse and forward maps are given by:

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

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

Or, for general $p$ and $q$, we have

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

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 $C_\mu^+$ and the last is $\chi^+$.

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 2.10 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.

## §2.9 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 $\epsilon$-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 $\epsilon$-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 $\epsilon$-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 $C_\mu$). 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 $\epsilon$-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 $\Xi$. 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 $C_\mu^+$ and $C_\mu^-$ 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 $\epsilon$-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 $\epsilon$-machine analysis beyond one-dimensional processes.

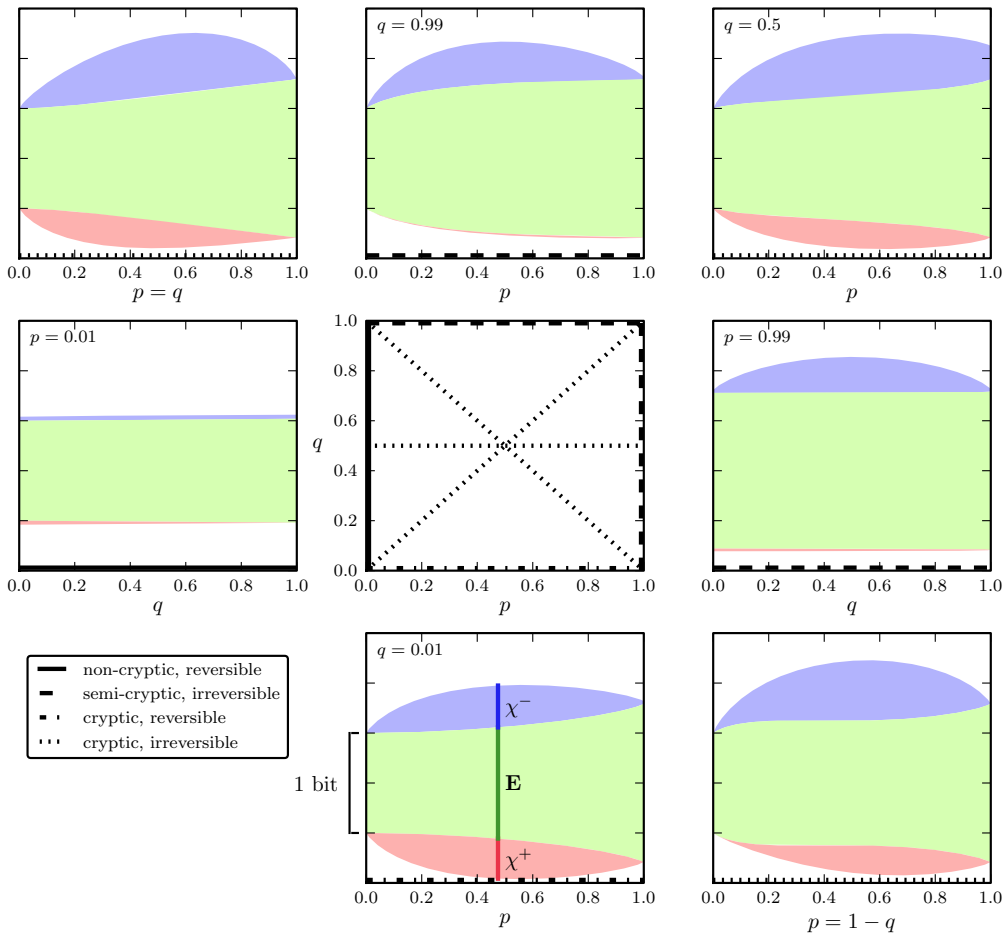


Figure 2.10: The Random Insertion Process's information processing measures as its two probability parameters $p$ and $q$ vary. The central square shows the $(p, q)$ parameter space, with solid and dashed lines indicating the paths in parameter space for each of the other information versus parameter 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 $\chi^-$, $\mathbf{E}$, and $\chi^+$.

CHAPTER 3

# Information Accessibility and Cryptic Processes

## §3.1 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. This expansion will lead to a classification of processes orthogonal to that provided by Markov order.

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

$$\mathbf{E} = I[\mathcal{S}^+; \mathcal{S}^-] \,, \tag{3.1}$$

and an accompanying contructive algorithm 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.

The complementary viewpoint, which we will take in this paper, is also provided by this result. That is, we very straightforwardly have,

$$\begin{aligned}
\chi^+ &= H[\mathcal{S}^+ | \overrightarrow{X}] \\
&= C_\mu - \mathbf{E} \\
&= C_\mu - I[\mathcal{S}^+; \mathcal{S}^-]
\end{aligned}$$

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

$$\begin{aligned}
\chi^+ &= H[\mathcal{S}^+ | \mathcal{S}^-] \text{ or} \\
\chi^- &= H[\mathcal{S}^- | \mathcal{S}^+] \,.
\end{aligned}$$

In the following we will not concern ourselves with reverse representations and so can simplify the notation, using $C_\mu$ for $C_\mu^+$ and $\chi$ for $\chi^+$.

Here we show that, for a restricted class of processes, the crypticity in Eqn. **??** can be systematically expanded to give an alternative closed-form to the excess entropy in Eqn. 3.1. One ancillary benefit is a new and, we argue, natural hierarchy of processes in terms of information accessibility.

## §3.2 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*

$$H[\mathcal{S}_k | \overrightarrow{X}_0] = 0 \,. \tag{3.2}$$

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

**Definition.** *An* $\infty$-cryptic *process is one for which the process's $\epsilon$-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 11.** *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. 11 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. 11 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 $\chi$ at different orders. To this end, we define the $k$-cryptic approximation.

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

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

### §3.2.1 The $k$-Cryptic Expansion

We will now develop a systematic expansion of $\chi$ 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*

$$\chi = \chi(k) + H[\mathcal{S}_k|\overrightarrow{X}_0] \,. \tag{3.3}$$

*Proof.* We calculate directly, starting from the definition, adding and subtracting the $k$-crypticity criterion term from $\chi$'s definition, Eqn. **??**:

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

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 $\epsilon$-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 $\chi$ 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. 3.3:

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

Likewise, assuming $\chi = \chi(k)$ requires, by Eqn. 3.3 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.*

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

□

## §3.2.2 Convergence

**Proposition 12.** *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* $\chi$, $\chi(j) = \chi$ *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. 12, $\chi(1) \leq \chi(k) = 0$. 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. 12, $\chi(k+1) \geq \chi(k)$. 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 $\chi$ at a finite $k$, remains at that value for all larger $k$.

**Proposition 13.** *The cryptic approximation $\chi(k)$ converges to $\chi$ 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 $\epsilon$ map from pasts to causal states of Eqn. **??**. Therefore, as $k \to \infty$, $\chi(k) \to \chi$. □

**Proposition 14.** *The cryptic approximation $\chi(k)$ is a concave function.*

We take a somewhat round-about method of proof here because it is intuitively easier to follow and it uses as a starting point, the block-state entropy function $H[X_0^L, \mathcal{S}_L]$, a function that parallels the block entropy function analyzed in Ref. **ruro**.

*Proof.* We begin by proving the convexity of the block-state entropy. The statement of convexity is,

$$H[X_0^{L+1}\mathcal{S}_{L+1}] - H[X_0^L\mathcal{S}_L] \geq H[X_0^L\mathcal{S}_L] - H[X_0^{L-1}\mathcal{S}_{L-1}]$$

By stationarity we have,

$$H[X_{-1}^{L+1}\mathcal{S}_L] - H[X_0^L\mathcal{S}_L] \geq H[X_{-1}^L\mathcal{S}_{L-1}] - H[X_0^{L-1}\mathcal{S}_{L-1}]$$

Equivalently,

$$H[X_{-1}|X_0^L\mathcal{S}_L] \geq H[X_{-1}|X_0^{L-1}\mathcal{S}_{L-1}]$$

We can use an I-diagram to help understand the convexity statement (Fig. 3.9). The convexity is now translated to,

$$\alpha + \gamma \geq \alpha + \beta$$

$$\gamma \geq \beta$$

Using the fact that the causal state is an optimal representation of the past, we have the following equivalent expressions for the entropy rate:

$$h_\mu =$$

$$H[X_{L-1}\mathcal{S}_L|\mathcal{S}_{L-1}] = \beta + \epsilon + \delta + \zeta$$

$$H[X_{L-1}\mathcal{S}_L|\mathcal{S}_{L-1}X_0^{L-1}] = \beta + \zeta$$

$$H[X_{L-1}\mathcal{S}_L|\mathcal{S}_{L-1}X_{-1}] = \epsilon + \zeta$$

$$H[X_{L-1}\mathcal{S}_L|\mathcal{S}_{L-1}X_{-1}X_0^{L-1}] = \zeta$$

Note that we relied on the shielding property of the causal states, and also the unifilarity of the $\epsilon$-machine. These four relations together yield,

$$\zeta = h_\mu$$

$$\beta = \delta = \epsilon = 0$$

And combining, we see that the convexity statement is simply $\gamma \geq 0$. Since $\gamma$ is a conditional mutual information, and is therefore positive semidefinite, we have shown that the block-state entropy function is convex. □

*Proof.* Now in order to prove that $\chi(L)$ is a concave function, we use the fact that it is the difference between the block-state entropy function and the state-block entropy function (which for an $\epsilon$-machine is linear).

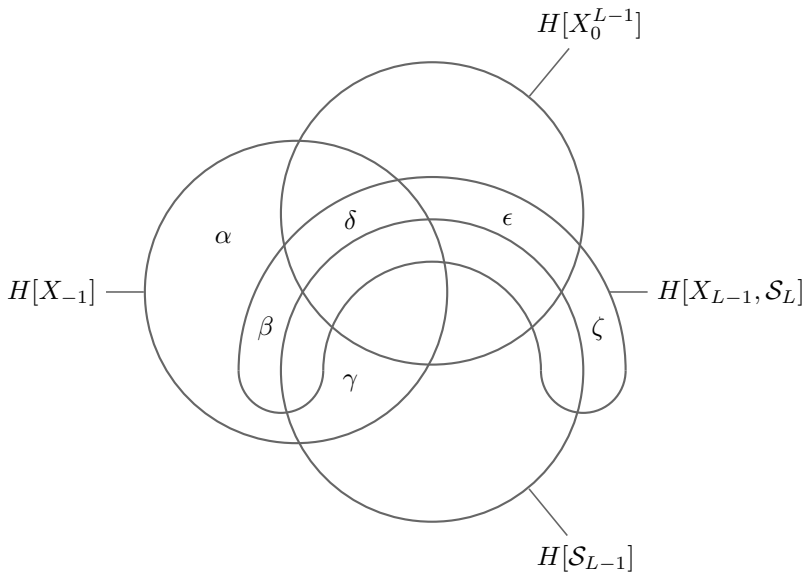


Figure 3.1: An I-diagram helps to organize the algebra. Note that we reduce the complexity of this diagram by making two of the variables aggregate variables. Also, we have opted for an alternate representation of the I-diagram keeping three of the regions circular.

The state-block entropy $H[\mathcal{S}_0, X_0^L]$ is quite clearly linear since for any $L$,

$$\begin{aligned} H[\mathcal{S}_0, X_0^{L+1}] - H[\mathcal{S}_0, X_0^L] &= H[X_L | \mathcal{S}_0, X_0^L] \\ &= H[X_L | \mathcal{S}_0, X_0^L, \mathcal{S}_L] \\ &= H[X_L | \mathcal{S}_L] \\ &= h_\mu \end{aligned}$$

Finally, noting that the cryptic approximation is the difference of these two new entropy functions,

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

We recall that the elementary result that a linear function less a convex function is equal to a concave function. This completes the proof. □

## §3.2.3 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.

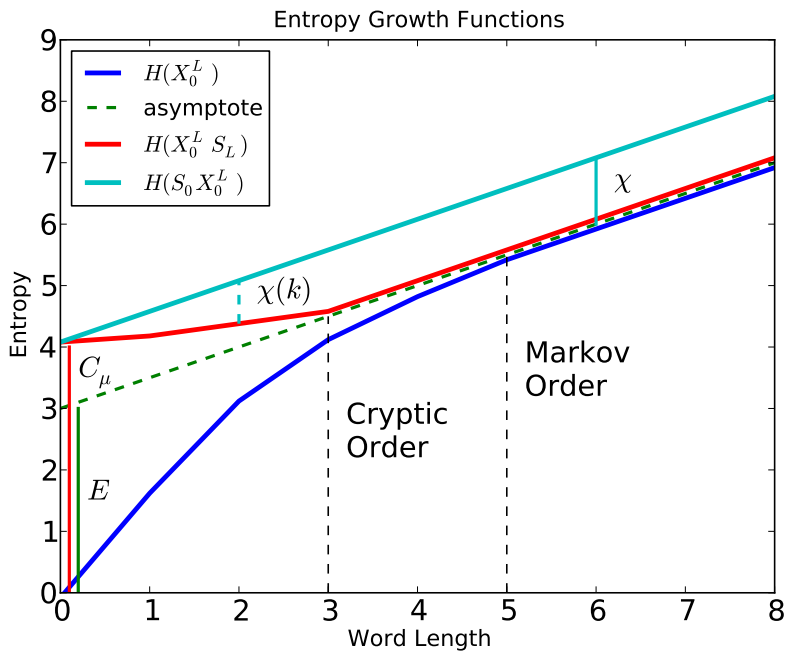


Figure 3.2: The entropy growth functions: block entropy $H[X_0^L]$, block-state entropy $H[X_0^L, \mathcal{S}_L]$, and state-block entropy $H[\mathcal{S}_0, X_0^L]$ provide a convenient way for understanding several of a process's properties. Previously, the entropy rate, excess entropy, and Markov order were seen on this diagram. We now add statistical complexity, crypticity, and cryptic order to that list. A pleasing feature of this figure is that it reproduces the I-diagram in Fig. 1.17 when viewed end on.

**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. □

### §3.2.4 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. **??**, which gives the excess entropy for order-$R$ Markov processes:

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

By Prop. 11, 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, $C_\mu \equiv H[S_R]$, 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:

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

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 15.**

$$H[X_0^R|\mathcal{S}_R] = 0 \quad \Longleftrightarrow \quad \chi(R) = R\, h_\mu\,, \tag{3.4}$$

*where $h_\mu$ 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 16.** *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 17.** *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. 15, 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. **??** 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 18.** *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. 17, this is not true. Therefore, the primitive orders of Markovity and crypticity are the same. □

## §3.3 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 $\epsilon$-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 $\infty$-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 $\chi$ without any knowledge of the cryptic order.

### §3.3.1 Even Process: 0-Cryptic

Figure 3.3 gives the $\epsilon$-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.

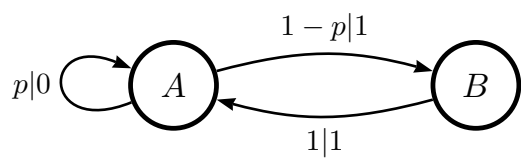


Figure 3.3: A 0-cryptic process: Even Process. The transitions denote the probability $p$ of generating symbol $x$ as $p|x$.

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 $C_\mu$ 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.

### §3.3.2 Golden Mean Process: 1-Cryptic

Figure 3.4 shows the $\epsilon$-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 $\epsilon$-machine looks remarkably similar to that for the Even Process. The informational analysis, however, shows that they have markedly different properties.

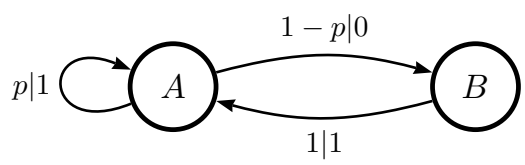


Figure 3.4: A 1-cryptic process: Golden Mean Process.

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 $C_\mu$ 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. 11, 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. **??** here, applies for an $R = 1$ Markov chain.

### §3.3.3 Butterfly Process: 2-Cryptic

The next example, the Butterfly Process of Fig. 4.1, 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 $\epsilon$-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 $\chi$

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

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

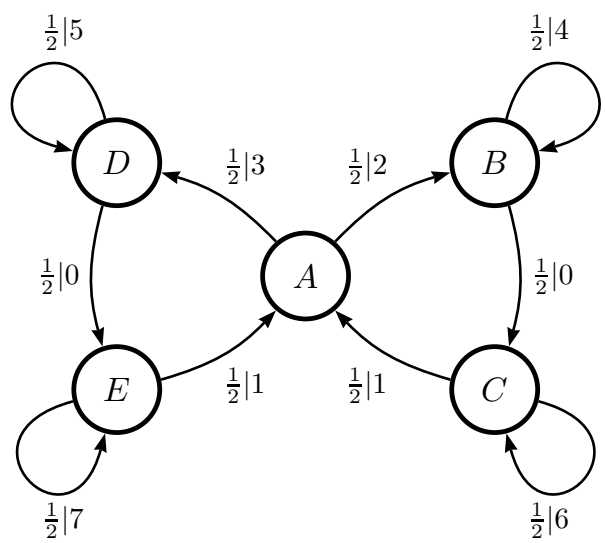


Figure 3.5: A 2-cryptic process: Butterfly Process over a 6-symbol alphabet.

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.

### §3.3.4 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. 4.4 for its $\epsilon$-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 11 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.

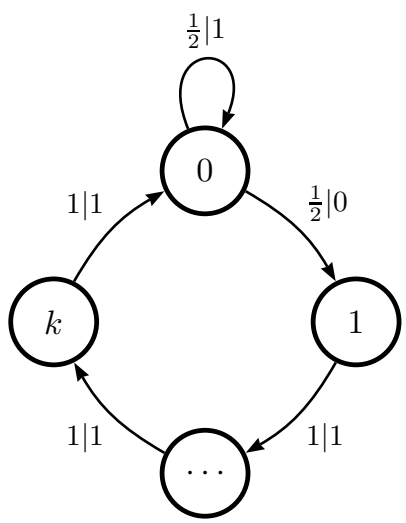


Figure 3.6: $k$-cryptic processes: Restricted Golden Mean Family.

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. [**?**].)

## §3.3.5 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 $\epsilon$-machine for this process is shown in Fig. 3.7.

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.

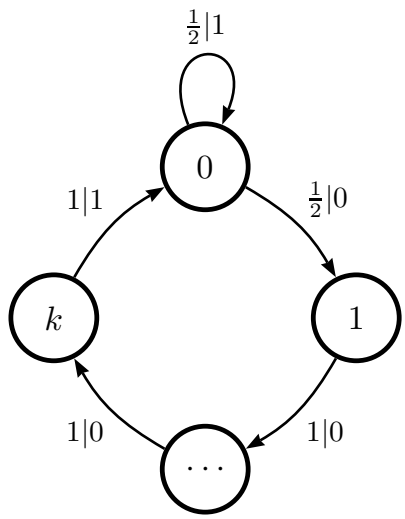


Figure 3.7: $k$-cryptic processes: Stretched Golden Mean Family.

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,

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

## §3.3.6 Nemo Process: $\infty$-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 $\epsilon$-machine is shown in Fig. 4.7.

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

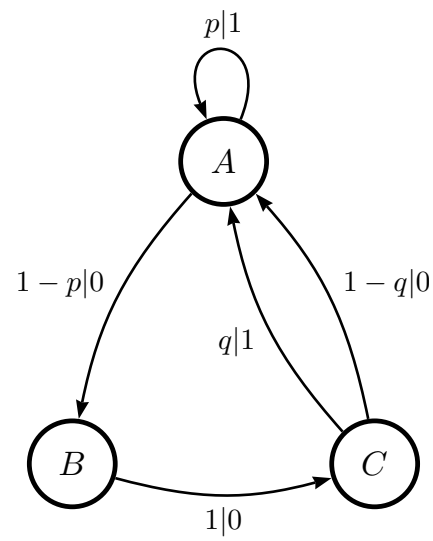


Figure 3.8: The $\infty$-cryptic Nemo Process.

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 $\epsilon$-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 $T^{(x)}_{\sigma\sigma'} = \Pr(\mathcal{S}' = \sigma', X = x|\mathcal{S} =$

$\sigma$). 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 $\infty$-cryptic. This means that we cannot make use of the $k$-cryptic approximation to calculate $\chi$ or **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 $C_\mu^+$. 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. [**?**].)

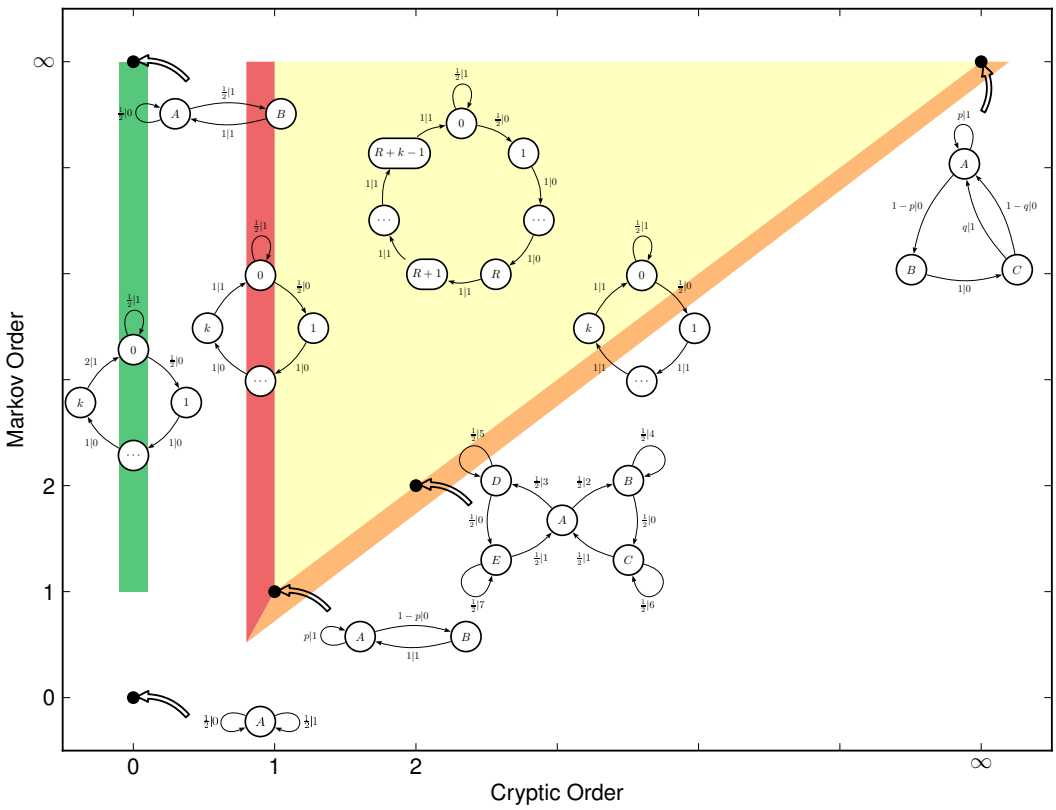


Figure 3.9: This figure shows a bird's-eye view of process space. Some sample processes were chosen and placed on a plot of Markov vs cryptic order. Some $\epsilon$-machines point to particukar points in the space while others are parameterized $\epsilon$-machines and refer to a colored region. We can readily see that aside from the bound $R \geq k$, the space is filled. This means that the cryptic order is a nontrivial complement to Markov order.

## §3.4 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 $\epsilon$-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 $\mathbf{E}$ 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 $\mathbf{E}$, 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 $\chi$ 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-$\infty$ 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 $\infty$-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 $\infty$-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 $\infty$-cryptic processes, the general theory introduced in Ref. [**?**] and Ref. [**?**] is seen to be necessary. It allows one to directly calculate **E** and crypticity and to do so efficiently.

CHAPTER 4

# Information Accessibility and Cryptic Processes: Linear Combinations of Causal States

## §4.1 Introduction

We introduced a new system "invariant"—the *crypticity* $\chi$—for stationary hidden stochastic processes to capture how much internal state information is directly accessible (or not) from observations [**?**, **?**, **?**]. Two approaches to calculate $\chi$ 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 $\epsilon$-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 **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 $\epsilon$-machine $M$ of a process, calculate its reverse-time representation $M^-$. (The latter is not always minimal and so not, in that case, an $\epsilon$-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

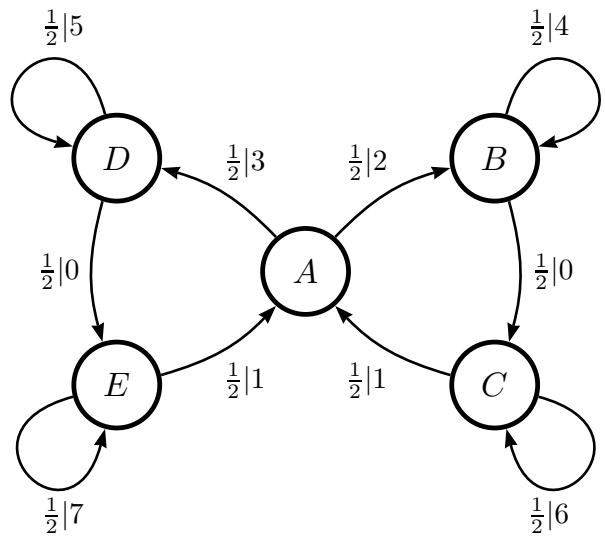


Figure 4.1: A 2-cryptic process: The $\epsilon$-machine representation of the Butterfly Process. Edge labels $t|x$ give the probability $t = T^{(x)}_{\sigma\sigma'}$ of making a transition and from causal state $\sigma$ to causal state $\sigma'$ and seeing symbol $x$.

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.

## §4.2 Butterfly Process

Figure 4.1 shows the $\epsilon$-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. 4.2.) Note that the time-reverse presentation is no longer unifilar and, therefore, it is not the reversed process's $\epsilon$-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 4.1.

The result is the reverse $\epsilon$-machine shown in Fig. 4.3. Note that it has two more states than the original (forward) $\epsilon$-machine of Fig. 4.1.

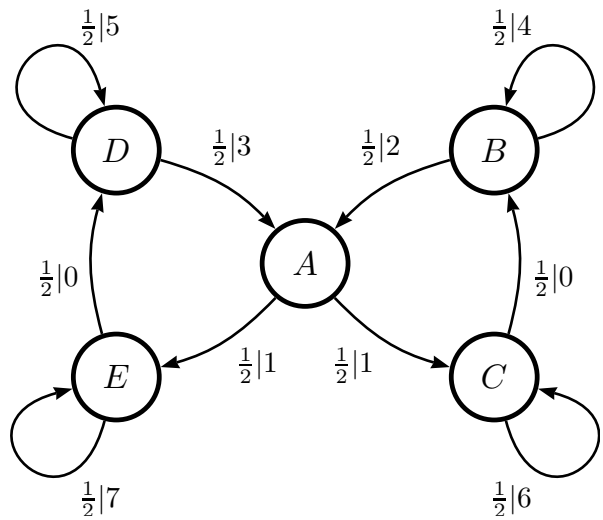


Figure 4.2: Time-reversed Butterfly Process.

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. [**?**]:

$$\mathbf{E} = C_\mu - \chi \tag{4.1}$$

$$\mathbf{E} = C_\mu - H[\mathcal{S}^+ | \overrightarrow{X}] \tag{4.2}$$

$$= C_\mu - H[\mathcal{S}^+ | \mathcal{S}^- = \epsilon^+(\overrightarrow{X})] \,. \tag{4.3}$$

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. [**?**].

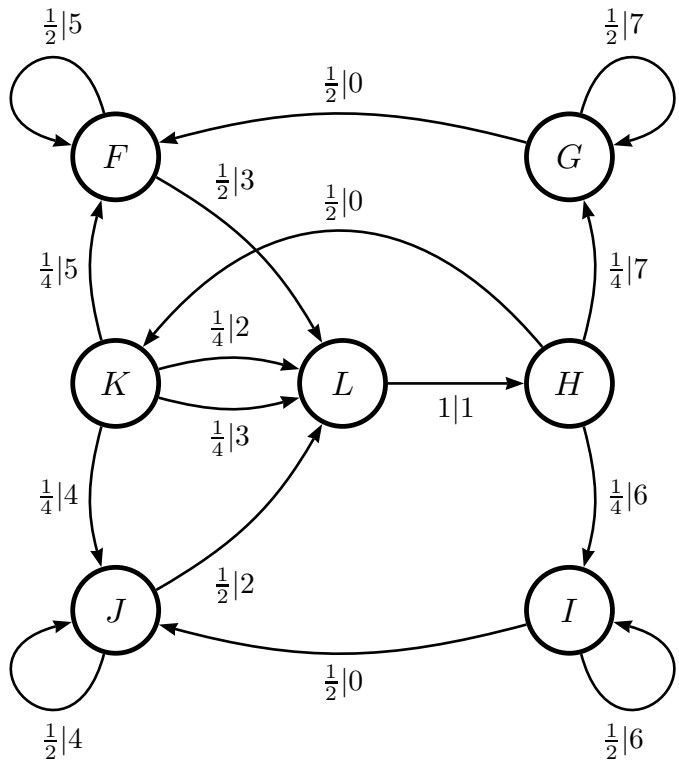


Figure 4.3: Reverse Butterfly Process.

## §4.3 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 $\epsilon$-machine is given in Fig. 4.4 and its stationary distribution is:

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

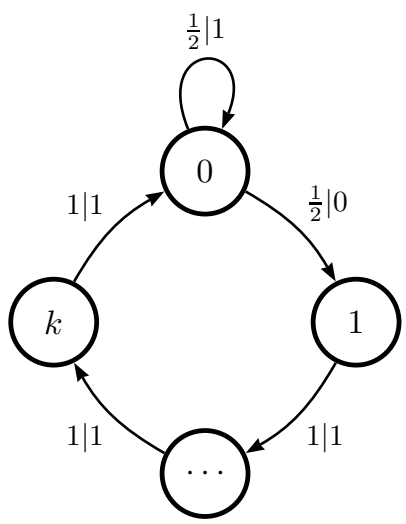


Figure 4.4: The $\epsilon$-machine for the Restricted Golden Mean Process.

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 $\epsilon$-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 4.2 gives the mixed states for these allowed words. It is also reason-

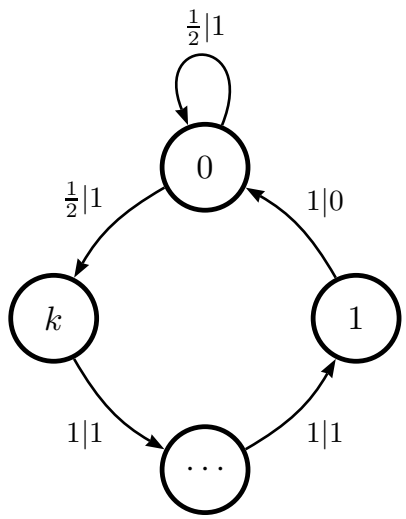


Figure 4.5: Time-reversed presentation of the Restricted Golden Mean Process.

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

With this, we can now compute $\chi$ 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:

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

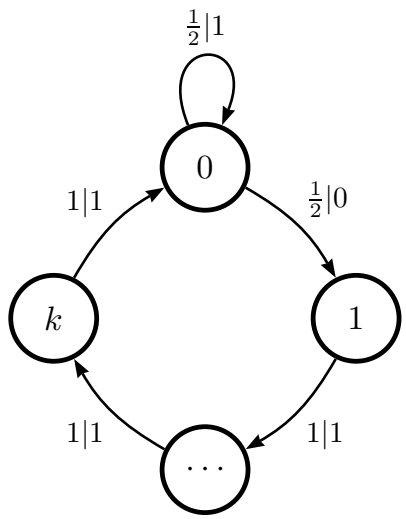


Figure 4.6: Reverse Restricted Golden Mean Process.

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

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

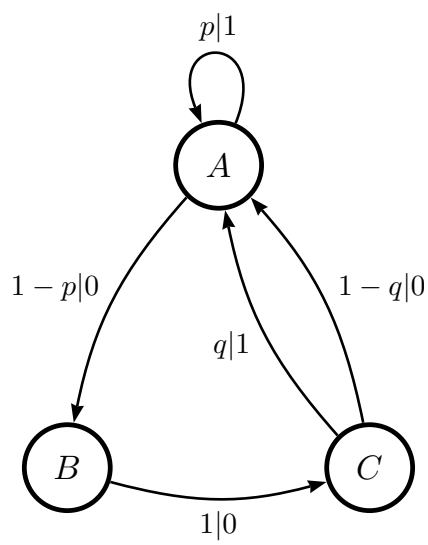


Figure 4.7: The $\epsilon$-machine for the $\infty$-cryptic Nemo Process.

## §4.4 Nemo Process

We now demonstrate how to calculate $\chi$ and $\mathbf{E}$ for Ref. [**?**]'s $\infty$-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 4.7 shows $M^+$, the $\epsilon$-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. 4.8.

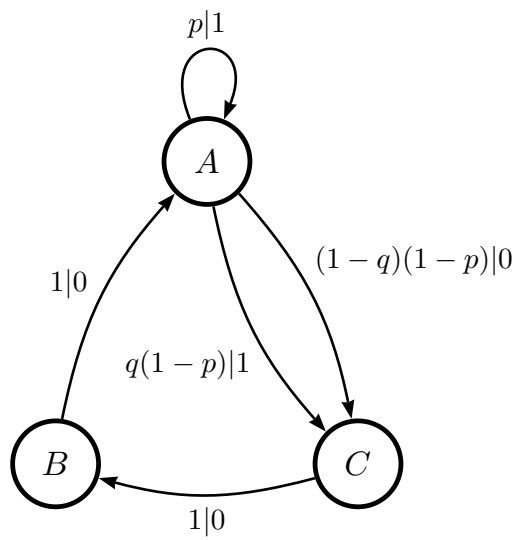


Figure 4.8: The time-reversed presentation, $\widetilde{M}^+ = \mathcal{T}(M^+)$, of the Nemo Process.

The transition matrices of this machine are:

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

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

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

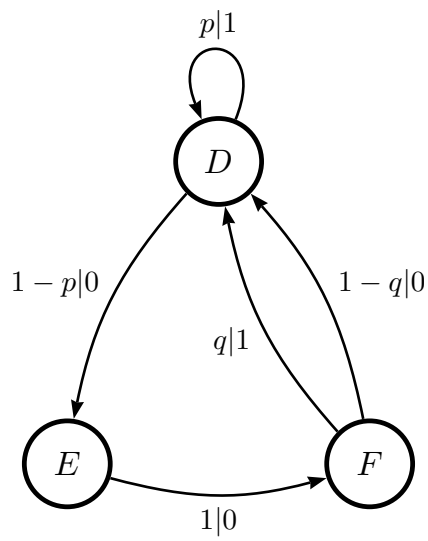


Figure 4.9: The reverse $\epsilon$-machine for the Nemo Process.

which is shown in Fig. 4.9. 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 $\epsilon$-machine causal states, which are exactly the same as the forward $\epsilon$-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:

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

We use this to directly compute:

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

Finally, we have:

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

## §4.5 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 | $\mu$ 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 4.1: Calculating the time-reversed Butterfly Process's $\epsilon$-machine via the forward $\epsilon$-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 | $\mu$ 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 $1 \leq n \leq k$ | $\frac{1}{k-n+1}(\frac{1}{2^n}, 1^{k-n}, \frac{1}{2^1}\frac{1}{2^2}\frac{1}{2^3}, \ldots, \frac{1}{2^n})$ |
| $0(1)^k$ | $(\frac{1}{2^k}, \frac{1}{2^1}\frac{1}{2^2}\frac{1}{2^3}, \ldots, \frac{1}{2^k})$ |
| $1(1)^k$ | $0(1)^k$ |
| $0(1)^k 0$ | 0 |
| $0(1)^k 1$ | $0(1)^k$ |

Table 4.2: Calculating the reversed RGMP using mixed states over the $\epsilon$-machine states.

# APPENDIX A
# Commentary on information measures

## §A.1 Entropy rate for bicyclists

We imagine the following scenario[1]—one not so far-fetched for a place such as Davis, CA:

A graduate student, G, new to Davis, is bicycling around town in an effort to get to know the place. G was never accused of overly developed spatial skills, but is, however, a diligent note-taker. Upon reaching each intersection, G stops and takes note of whether she decides to continue on straight (S), turn left (L), or turn right (R). If she were to begin again at the same place, she could reproduce her exact path through town by going from intersection to intersection and following the next instruction—S, R, S, S, L, etc.

At first, exploration is quite unpatterned. G has a very poor sense for city planning, so chooses from the possibilities {S, A, R} equally. The probability distribution over the choices at each intersection is uniform: $\{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}$.

A typical bike ride that crosses 30 intersections would look like:

```
L,S,S,R,S,S,S,L,L,S,L,R,R,R,R,L,R,L,R,L,S,L,L,S,R,R,L,L,S,S
```

Now imagine that G begins to learn her way around town and, desiring to go someplace in particular, will do so in a more or less direct manner. This will have some obvious manifestations in her notes. For instance, we are very unlikely to see L,L,L,L and much more likely to see long strings of Ss.

To make the case easier to analyze, let's suppose that she now knows the city grid well enough to never make even two consecutive lefts or rights.

Maybe one of her rides looks like:

[1] This is not intended as an affront to bicyclists or their intelligences. Nor is it implying that the following is particularly non-pedestrian.

S,R,S,S,L,S,S,S,S,S,S,S,S,S,L,S,S,S,S,S,S,S,L,S,S,S,S,S,S,S

## §A.2 Statistical Complexity for ...

# APPENDIX B
# Venn Diagrams and Information Theory

"The reader will find many figures in this work.[1]"

## §B.1 Venn Diagrams and Information Theory

Venn diagrams used to understand the information theoretic relationships among random variables, or *I-diagrams*, are a crucial member of any information theorist's tool belt (see Fig. B.2).

Given a set of random variables, there is a one-to-one mapping from the set of information quantities—entropies, mutual informations, conditional mutual informations—to an I-diagram.

Given that this map is one-to-one, how can I claim that the I-diagram is crucial? Is this not just another redundant representation to learn? The answer is most definitely: No.

As is the general theme of this thesis and the surrounding body of work, *natural* representations are given special status and attention. The I-diagram is not claimed to be natural in the same sense that the $\epsilon$-machine is. However, the association of random variables with geometric bodies, and their relationships with geometric intersection and occlusion, will *prove itself* to be natural for humans.

Why might this be the case? I claim that this is due to our visual hardware/firmware. We very naturally perceive connected homogeneous regions as objects. We also are skilled at understanding occlusion and how occlusions change as objects move. Then again, for some reason it seemed to take man a rather long time to figure out what the lunar eclipse was.

I submit that the student of information theory *knows* I-diagrams whether or not they have seen one.[2] Here we seek only to make this knowledge more explicit and to develop a little of the I-diagram calculus.

[1] "The reader will find no figures in this work" - J. L. Lagrange, Mécanique Analytique, 1888.

[2] "Before hearing Monge, I did not know that I knew descriptive geometry." - J. L. Lagrange, said after a lecture by Gaspard Monge. Monge was the inventer of descriptive geometry. This highly intuitive method is now integral to design and engineering.

## §B.2 How to read an I-diagram

Each random variable corresponds to one 'smooth' object in the diagram. This is not a requirement, but makes visual inspection easier. When possible, we will use circles or ellipses. Certain circumstances will encourage slightly more exotic smooth shapes.

The area of a smooth shape corresponds to the entropy of that random variable, $H[X]$ (see Fig. B.1). This is the total uncertainty in the variable $X$. We will loosely refer to these shapes as both random variables and their entropies. The meaning should be clear from context.

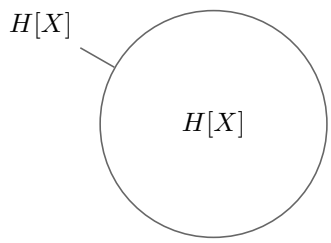


Figure B.1: The simplest I-diagram - one random variable, $X$.

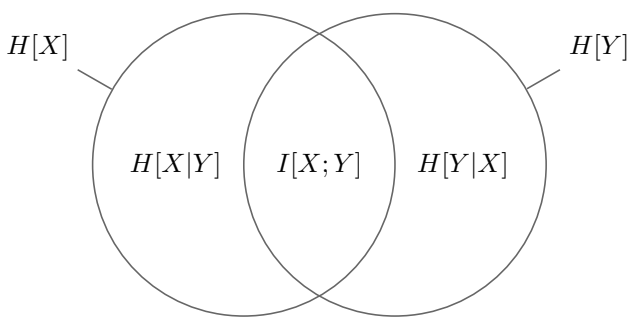


Figure B.2: I-diagram for two random variables, $X$ and $Y$.

In an I-diagram involving two random variables, the area of intersection of two random variable shapes, $X$, $Y$, corresponds to the mutual information between those random variables, $I[X;Y]$ (see Fig. B.3). This is the information that the two random variables share.

The area of random variable $X$ that is unoccluded by random variable $Y$ corresponds to the conditional entropy, $H[X|Y]$. This is the uncertainty in $X$ that remains when $Y$ is known.

Now consider a three variable I-diagram. What is the uncertainty in $X$ given *both* $Y$ and $Z$? This is notated $H[X|YZ]$ as in Fig. B.5.

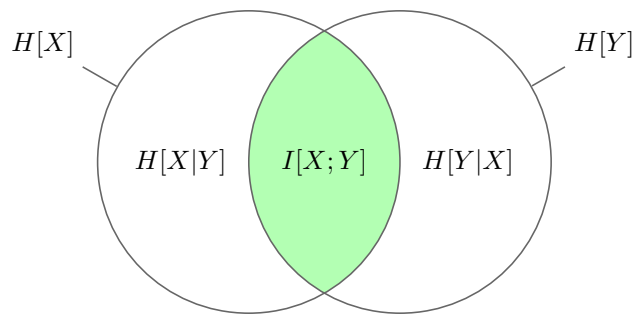


Figure B.3: The mutual information between $X$ and $Y$ is highlighted.

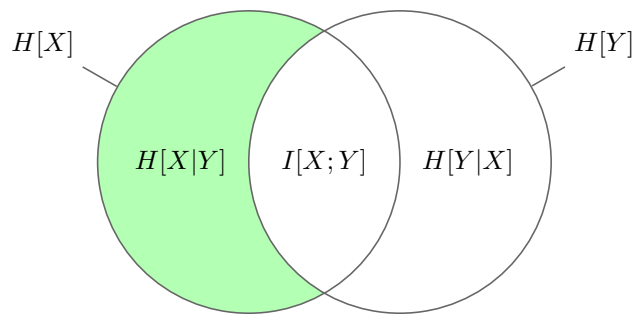


Figure B.4: The conditional entropy of $X$ given $Y$ is highlighted.

We can equivalently think of merging $Y$ and $Z$ to form a random variable, $W$, with a larger event space. In this light, our quantity of interest would look like $H[X|W]$. We can think of this graphically quite naturally by considering $W$ as the union of $Y$ and $Z$ (see Fig. B.6).

Now considering a three variable I-diagram, the area of intersection between $X$ and $Y$ that is unoccluded by $Z$ is the conditional mutual information, $I[X;Y|Z]$ (see Fig. B.7). This is the information shared by $X$ and $Y$ when $Z$ is known. Adding knowledge about $Z$ can either increase or decrease the (conditioned) mutual information.

From these diagrams, e.g. Fig. B.2, we can visually prove such identities as:

$$H[XY] = H[X|Y] + I[X;Y] + H[Y|X]$$

$$I[X;Y] = H[X] - H[X|Y]$$

## §B.2.1 Stratification of a Composite Variable

Just as more than one random variable can be considered as a composite variable, as in Fig. B.6, a composite variable can be stratified in a very intuitive way. Consider the set of random variables $\{X_0, X_1, X_2, X_3, X_4\}$. The composite random variable $X_0^5$ can be stratified by the set $\{X_0^1, X_0^2, X_0^3, X_0^4, X_0^5\}$. These results are generic for any set of 'stratifying' variables.

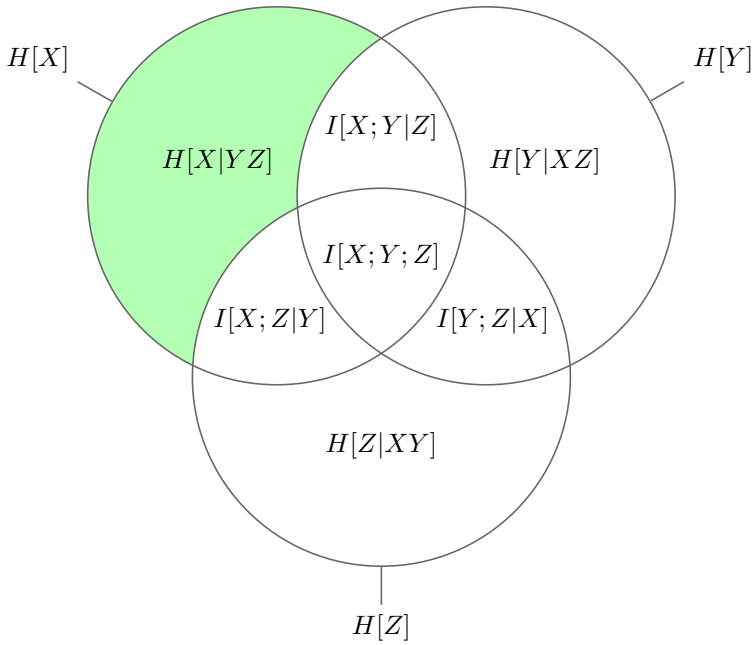


Figure B.5: The conditional entropy of $X$ given $Y$ and $Z$ is highlighted.

**Definition.** *Given a set* $\mathbf{X} = \{X_1, \ldots, X_N\}$ *for some integer,* $N$*, a set of subsets of* $\mathbf{X}$*,* $\mathbf{Y}$*, is called* stratifying *if a total ordering on* $\mathbf{Y}$ *is induced by the relation* $\subset$*.*

For any such $\mathbf{X}$, we have the stratifying set $\mathbf{Y} = \{\{X_1\}, \{X_1, X_2\}, \ldots, \{X_1, X_2, \ldots, X_N\}\}$.

We can stratify the past random variables of a process as in Fig. B.8.

That this is true follows straightforwardly from the duality between information measures and set operations demonstrated by **cite Yeung**.

Additionally, the I-diagram involving a stratified variable and one additional variable has the form in Fig. **??**. To see that this is not a trivial statement, consider the following scenario. Let random variables $X$ and $Y$ each be fair coins. Then let $Z$ be the exclusive-or (XOR) of $X$ and $Y$. If we were to begin with the I-diagram for just $X$ and $Y$, we would have the following picture. **graphic here**

Then naively 'adding' $Z$ to the I-diagram yields the following picture. **graphic here**

This fails to capture the true information theoretic structure. Specifically, it does not capture that $I[X;Y|Z] = 1$ or that $I[X;Y;Z] = -1$. The correct picture does not allow for regions to be eliminated before the addition of all variables to the diagram. This is why it is non-trivial to 'naturally' add $Z$ to the stratified variable I-diagram.

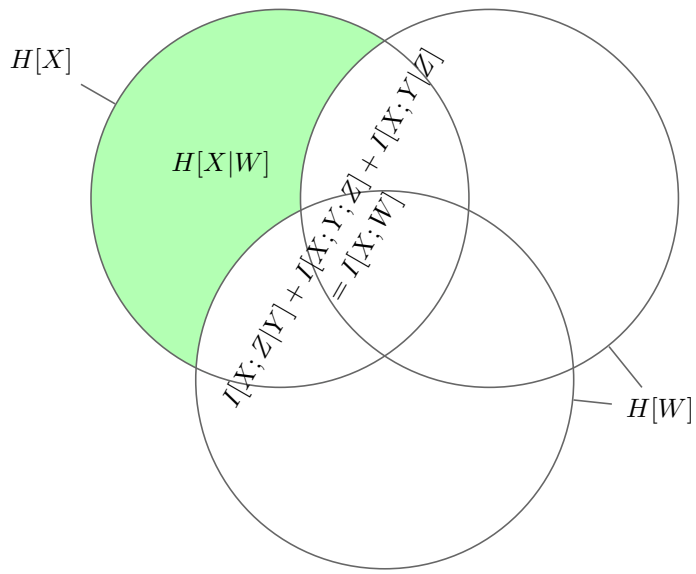


Figure B.6: The mutual information of $X$ and joint variable $W$ is highlighted.

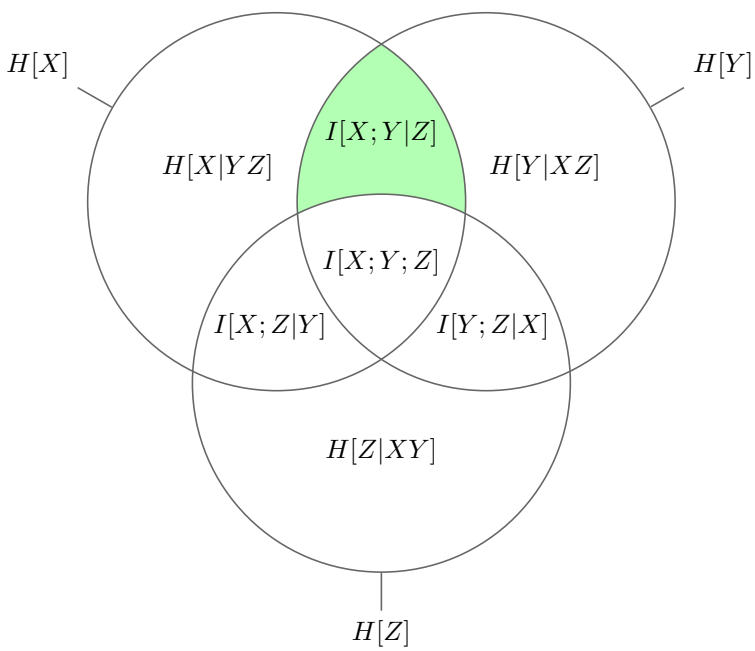


Figure B.7: The conditional mutual entropy of $X$ and $Y$ given $Z$ is highlighted.

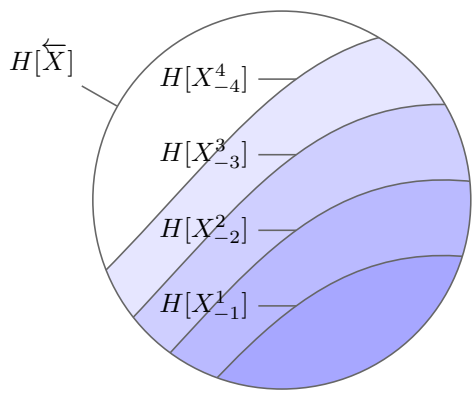


Figure B.8: The standard stratification of the conglomerate random variable $\overleftarrow{X}$.

# APPENDIX C
# Proofs

## §C.1 Entropic Independence ⇒ Probabilistic Independence

Show that $H[\mathcal{S}_R|X_0^R] = 0 \iff$ the process is order-R Markov.

$$\Pr(X_0X_1X_2\ldots|\ldots X_{-3}X_{-2}X_{-1}) = \Pr(X_0X_1X_2\ldots|X_{-R}\ldots X_{-2}X_{-1})$$

First we need a new equivalence relation.

$$*x_{-R}^{R'} \overset{R}{\sim} *x_{-R}^{R''} \iff \Pr(\overrightarrow{X}_0|x_{-R}^{R'}) = \Pr(\overrightarrow{X}_0|x_{-R}^{R''})$$

Assume order-R Markov, and show that this relation induces the same partition on histories as the original causal relation.

$$\begin{aligned}
&\forall \boldsymbol{\sigma} \in \boldsymbol{\mathcal{S}}, \forall \overleftarrow{x}', \overleftarrow{x}'' \in \boldsymbol{\sigma}, \overleftarrow{x}' \sim \overleftarrow{x}'' \\
&\iff \Pr(\overrightarrow{X}|\overleftarrow{x}') = \Pr(\overrightarrow{X}|\overleftarrow{x}'') \\
&\iff \Pr(\overrightarrow{X}|x_{-R}^{R'}) = \Pr(\overrightarrow{X}|x_{-R}^{R''}) \\
&\iff x_{-R}^{R'} \overset{R}{\sim} x_{-R}^{R''}
\end{aligned}$$

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

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

$$\sum_{w\in\mathcal{A}^R} \Pr(X_0^R = w)H[\mathcal{S}_R|X_0^R = w] = 0$$

Since we must only consider words with non-zero probability, we have

$$\forall w \in \mathcal{A}^R : \Pr(w) > 0, H[\mathcal{S}_R|X_0^R = w] = 0$$

In other words, all words of length $R$ induce a causal state. What is left is to show that the causal state induced is the same as any induced by a history ending with that word.

If a word $w$ induces a state, then

$$\sum_{i\ldots m} \pi_i T_{ij}^{w_1} T_{jk}^{w_2} \ldots T_{mn}^{w_R}$$

has only one non-zero entry.

Since $\pi_i \neq 0$ the inside sum

$$\sum_{jk\ldots m} T_{ij}^{w_1} T_{jk}^{w_2} \ldots T_{mn}^{w_R} \propto \mathcal{P}_{in}^{\sigma}$$

Where $\mathcal{P}_{in}^{\sigma}$ is a projector onto causal state $\sigma$.

Assume $sw$ induces a different state

$$\sum_{h\ldots m} \pi_h T_{hi}^{s} T_{ij}^{w_1} T_{jk}^{w_2} \ldots T_{mn}^{w_R}$$

We can perform the sum $\sum_h \pi_h T_{hi}^{s}$ to get another distribution, $\pi_i'$.

$$\sum_i \pi_i' \mathcal{P}_{in}^{\sigma} = 0, 1$$

If zero, this is not a valid word. If 1, the projector must project onto the same subspace and thus the same state is induced by this extended word.

# APPENDIX D

# Mixed-State Presentation is Sufficient to Calculate the Switching Maps

While we conjecture that the mixed-state operation $\mathcal{U}(\widetilde{M}^+)$ yields an $\epsilon$-machine, this remains an open problem. Our conjecture, however, is based on a rather large number of test cases in which it is an $\epsilon$-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 $\epsilon$-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 19.** *$B$'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 20.** $H[\mathcal{R}'|\mathcal{R}, X] = 0$ *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.* $\square$

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

**Proposition 21.** *Two pasts that induce the same state in $B$ must be pasts in the same causal state of $B$'s $\epsilon$-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 $\epsilon$-machines and apply this technique to calculate $\mathbf{E}$ in the case where the resultant machine $\mathcal{T}(M^+)$ is not an $\epsilon$-machine.

The input machine is the reversed $\epsilon$-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 $\mathbf{E}$ using:

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