CHAPTER 1

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

§1.1 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.

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

[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 1.1.

									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 1.1: 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:*

$$H[X_{-L+2}, \ldots, X_0] = H[X_{-L+1}, \ldots, X_{-1}] \,. \tag{1.1}$$

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.

§1.3 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 ϵ-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_μ, C_μ, and **E**. Ref. [**?**] announced the solution to this long-standing problem by deriving explicit expressions for **E** in terms of the ϵ-machine, providing a unified information-theoretic analysis of general processes. Here we provide a detailed account of the underlying methods and results.

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 [**?**]:

$$\mathbf{E} = C_\mu - R\, h_\mu \,. \tag{1.2}$$

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 [**?**]:

$$\mathbf{E} \leq C_\mu \,. \tag{1.3}$$

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 ϵ-machines capture information from the past and the future. To do this we analyzed [**?**] a four-variable mutual information: $I[\overleftarrow{X}; \overrightarrow{X}; \mathcal{S}^+; \mathcal{S}^-]$. A large number of expansions of this quantity are possible. A systematic development follows from Ref. [**?**] which showed that Shannon entropy $H[\cdot]$ and mutual information $I[\cdot;\cdot]$ form a signed measure over the space of events.[2] Practically, there is a direct correspondence between set theory and these information measures. Using this, Ref. [**?**] developed an *ϵ-machine information diagram* over four variables, which gives a minimal set of entropies, conditional entropies, mutual informations, and conditional mutual informations necessary to analyze the relationships among h_μ, C_μ, and **E** for general stochastic processes.

In a generic four-variable information diagram, there are 15 independent quantities. These quantities can be seen in Fig. 1.1 as atoms, or regions of the I-diagram. Fortunately, this greatly simplifies in the case of using predictive and retrodictive ϵ-machines to represent the process; there are only 5 independent variables in this special case (see Fig .1.2). Reference [**?**] contains more details of this reduction. Here we present the main ideas.

The first attack on Fig. 1.1 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. **??** 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.

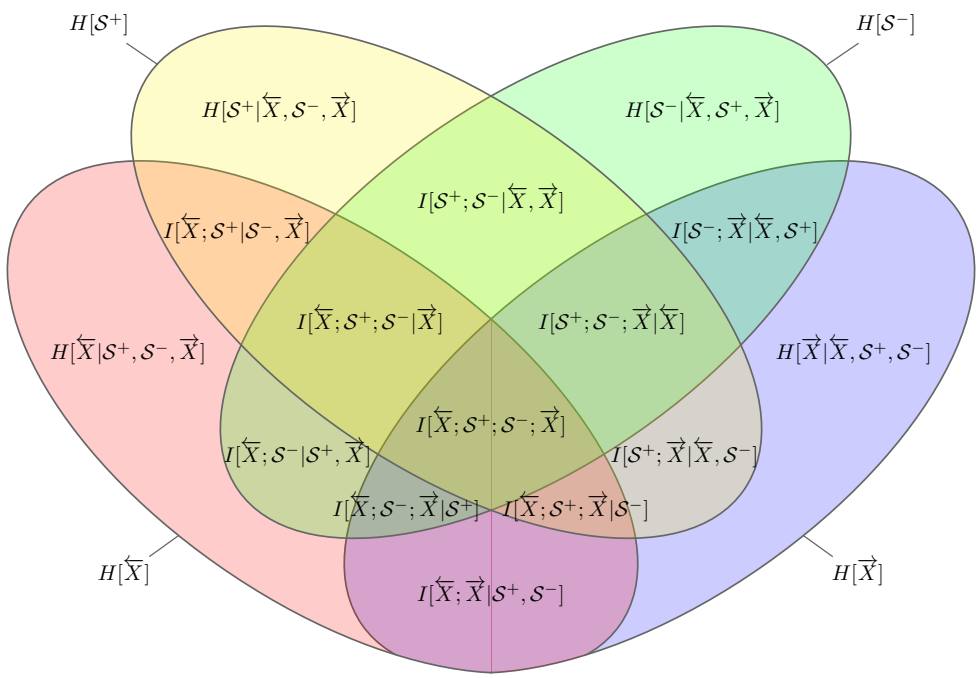

Figure 1.1: The generic (un-reduced) I-diagram for 4 random variables, where the names of the variables of interest have been inserted.

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

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

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

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

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 ϵ-machines and the processes they model (see Fig. 1.2).

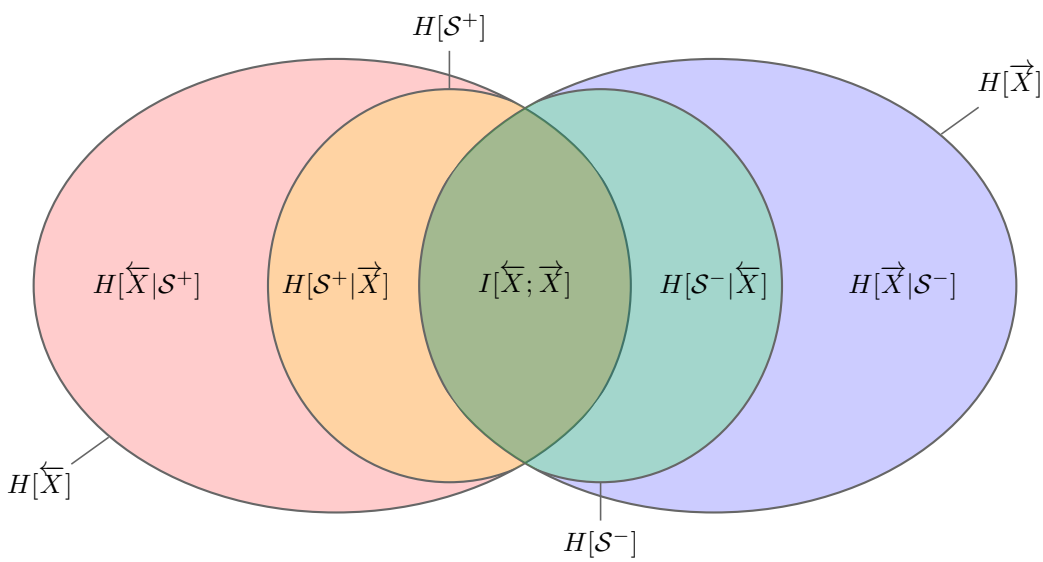

Figure 1.2: The I-diagram for the forward and reverse ϵ-machines. Only 5 of the 15 independent information quantities remain. This image is a central reference for the work following.

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:*

$$\mathbf{E} = I[\mathcal{S}^+;\mathcal{S}^-]\,. \tag{1.4}$$

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*

$$\begin{aligned} I[\overleftarrow{X}; \overrightarrow{X}; \mathcal{S}^+; \mathcal{S}^-] &= I[\overleftarrow{X}; \overrightarrow{X}] \\ &= \mathbf{E}\,, \end{aligned} \tag{1.5}$$

on the one hand, and

$$I[\overleftarrow{X}; \overrightarrow{X}; \mathcal{S}^+; \mathcal{S}^-] = I[\mathcal{S}^+; \mathcal{S}^-]\,, \tag{1.6}$$

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 ϵ-machine states.

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

$$C_\mu^+ = \mathbf{E} + H[\mathcal{S}^+|\mathcal{S}^-] \text{ and} \tag{1.7}$$

$$C_\mu^- = \mathbf{E} + H[\mathcal{S}^-|\mathcal{S}^+]\,. \tag{1.8}$$

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

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

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

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

§1.4 The Bidirectional Machine

At this point, we have two separate ϵ-machines—one for predicting (M^+) and one for retrodicting (M^-). We will now show that one can do better[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 ϵ-machines [**?**].

To illustrate, imagine being given a particular realization $\overleftrightarrow{x}$. In effect, the bidirectional machine $M^\pm$ describes how one can move around on the hidden process lattice of Table 1.1:

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 ϵ-machine. For example, if one moves forward following M^+ and ends in state $\mathcal{S}^+$, having seen $\overleftarrow{x}$ and about to see $\overrightarrow{x}$, then one moves to $\mathcal{S}^- = \epsilon^-(\overrightarrow{x})$.

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

A similar argument shows that f is onto. □

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

$$C_\mu^{\pm} \equiv H[\mathcal{S}^{\pm}] = H[\mathcal{S}^+, \mathcal{S}^-] \,. \tag{1.9}$$

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_μs as proportional to the size of the corresponding machine, we establish the representational efficiency of the bidirectional machine:

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

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

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

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

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

§1.4.1 Upper Bounds

Here we give new, tighter bounds for $\mathbf{E}$ than Eq. (1.3) 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_μ^-. *Similarly, since* $I[\mathcal{S}^-;\mathcal{S}^+] \leq H[\mathcal{S}^+]$, *we have* $\mathbf{E} \leq C_\mu^+$. □

§1.4.2 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:*

$$\Xi(\mathcal{P}) = C_\mu^+ - C_\mu^- \ . \tag{1.10}$$

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.

§1.4.3 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 $\mathbf{E}$. We now introduce a measure of this accessibility.

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

$$\chi^{\pm}(M^{+}, M^{-}) = H[\mathcal{S}^{+}|\mathcal{S}^{-}] + H[\mathcal{S}^{-}|\mathcal{S}^{+}] \,. \tag{1.11}$$

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

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

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

$$C_{\mu}^{\pm} = \mathbf{E} + \chi^{\pm} \,. \tag{1.12}$$

Proof. *This follows directly from the corollary and the predictive and retrodictive statistical complexity relations, Eq. (1.7) and (1.8).* □

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:*

$$C_{\mu}^{\pm} \geq \chi^{\pm} \,. \tag{1.13}$$

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 ϵ-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 χ^{+}, χ^{-}, $\chi^{\pm}$, $C_{\mu}^{\pm}$, and Ξ follow straightforwardly.

§1.5 Alternative Presentations

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

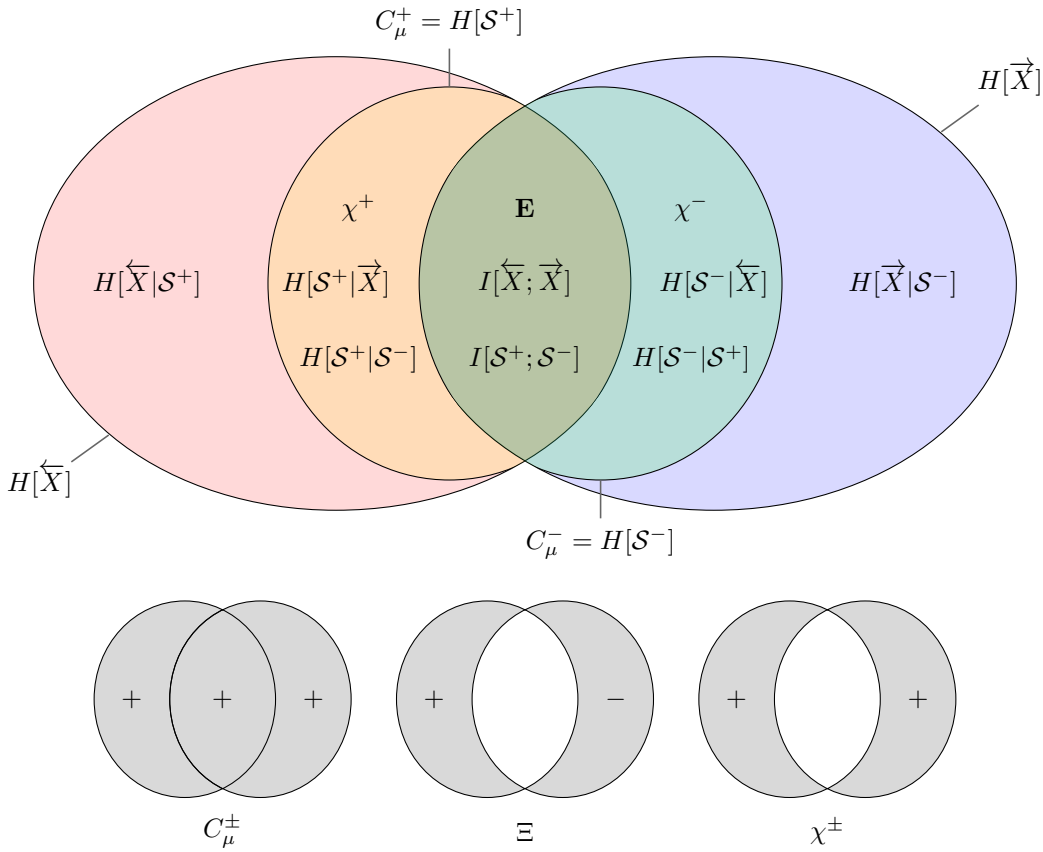

Figure 1.3: This diagram summarizes the measures and relationships derived in this chapter. The upper part of the figure should already be familiar—some relationships have been added. The bottom three icons illustrate which portions of the above diagram are added (or subtracted) to obtain the three newly defined measures: $C_\mu^\pm$, Ξ, and $\chi^\pm$. These represent the process's bidirectional information storage, irreversibility, and information overhead, respectively.

§1.5.1 Time-Reversed Presentation

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

$$T^{(x)}_{\mathcal{R}\mathcal{R}'} \equiv \Pr(X = x, \mathcal{R}'|\mathcal{R})\,. \tag{1.14}$$

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

Using standard probability rules to interchange $\mathcal{R}$ and $\mathcal{R}'$, we can construct a new set of transition matrices which defines a presentation of the process that generates the symbols in

reverse order. It is useful to consider a time-reversing operator acting on a machine. Denoting it $\mathcal{T}$, $\widetilde{M} = \mathcal{T}(M)$ is the *time-reversed presentation* of M. It has symbol-labeled transition matrices:

$$\begin{aligned} \widetilde{T}^{(x)}_{\mathcal{R}'\mathcal{R}} &\equiv \Pr(X = x, \mathcal{R}|\mathcal{R}') \\ &= T^{(x)}_{\mathcal{R}\mathcal{R}'} \frac{\Pr(\mathcal{R})}{\Pr(\mathcal{R}')} \,. \end{aligned} \tag{1.15}$$

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 π of M.*

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

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

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

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

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

§1.5.2 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:

$$\mu(w) \equiv \Pr(\mathcal{R}_L | X_0^L = w) \tag{1.16}$$

$$= \frac{\Pr(X_0^L = w, \mathcal{R}_L)}{\Pr(X_0^L = w)} \tag{1.17}$$

$$= \frac{\pi T^{(w)}}{\pi T^{(w)} \mathbf{1}}, \tag{1.18}$$

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:

$$\Pr(x, \mu(wx) | \mu(w)) \equiv \frac{\Pr(wx)}{\Pr(w)} \tag{1.19}$$

$$= \mu(w) T^{(x)} \mathbf{1} \,. \tag{1.20}$$

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

§1.6 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 ϵ-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:

$$\begin{array}{ccc} M^+ & \xleftarrow{\mathcal{U}} & \widetilde{M}^- \\ \mathcal{T}\big\downarrow & & \big\uparrow\mathcal{T} \\ \widetilde{M}^+ & \xrightarrow[\mathcal{U}]{} & M^- \end{array} \tag{1.21}$$

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

Also, note that $\mathcal{U}$ is not guaranteed to construct a minimal presentation of the process. However, this does not appear to be an issue when working with time-reversed presentations of an ϵ-machine. We leave it as a conjecture that $\mathcal{U}(\mathcal{T}(M))$ is always minimal. Even so, App. **??** 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 1.1 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 $\nu(w)$:

$$\nu(w) \equiv \Pr(\mathcal{R}_0 | X_0^L = w) \tag{1.22}$$

$$= \frac{\Pr(\mathcal{R}_0, X_0^L = w)}{\Pr(X_0^L = w)} \tag{1.23}$$

$$= \frac{\pi T^{(\widetilde{w})}}{\pi T^{(\widetilde{w})} \mathbf{1}}, \tag{1.24}$$

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

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

$$\Pr(x, \nu(xw) | \nu(w)) \equiv \frac{\Pr(xw)}{\Pr(w)} \tag{1.25}$$

$$= \nu(w) T^{(x)} \mathbf{1}. \tag{1.26}$$

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

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

Hence, we can calculate $H[\mathcal{S}^+|\mathcal{S}^-]$ and obtain $\mathbf{E}$ via (1.4).

§1.7 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 ϵ-machine has three recurrent causal states $\boldsymbol{\mathcal{S}}^+ = \{A, B, C\}$ and is shown in Fig. 1.4(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 ϵ-machine causal states and transitions.)

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

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

Using the $T^{(x)}$ and π, we create the time-reversed presentation $\widetilde{M}^+ = \mathcal{T}(M^+)$. This is shown in Fig. 1.4(b). Notice that the machine is not unifilar, and so it is clearly not an ϵ-machine. The transition matrices for the time-reversed presentation are given by:

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

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

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. 1.4(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

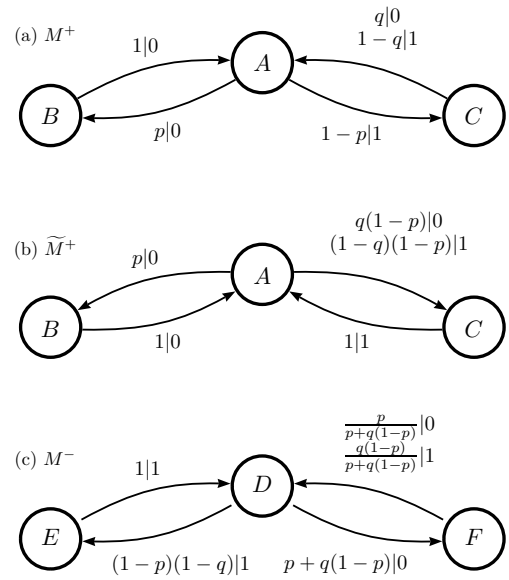

Figure 1.4: The presentations used to calculate the excess entropy for the RnC Process: (a) M^+, (b) $\widetilde{M}^+ = \mathcal{T}(M^+)$, and (c) $M^- = \mathcal{U}(\widetilde{M}^+)$. Edge labels $t\,|x$ give the probability $t = T^{(x)}_{\mathcal{R}\mathcal{R}'}$ of making a transition and seeing symbol 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

$$\nu(10) = \frac{\nu(0)\widetilde{T}^{(1)}}{\nu(0)\widetilde{T}^{(1)}\mathbf{1}} \,. \tag{1.27}$$

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. (1.26), 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.

§1.8 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.

§1.8.1 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 ϵ-machine representation. See Figure 1.5(a). Its behavior is characterized by consecutive 1s always appearing in even blocks. With probability p, each block of 1s can be followed by a 0, which can repeat until the next even block of 1s.

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

Its forward ϵ-machine has two recurrent causal states $\boldsymbol{\mathcal{S}}^+ = \{A, B\}$ and transition 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 1.5(a) gives M^+, while 1.5(b) gives M^-. We see that the ϵ-machines are the same and so the Even Process is causally reversible ($\Xi = 0$). Note that $\widetilde{M}^+$ is unifilar.

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

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

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

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

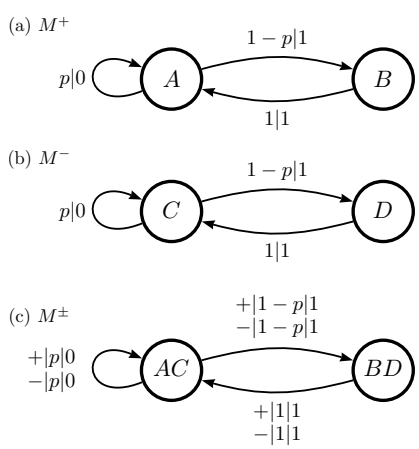

Figure 1.5: Forward and reverse ϵ-machines for the Even Process: (a) M^+ and (b) M^-. (c) The bidirectional machine $M^\pm$. Edge labels are prefixed by the scan direction $\{-,+\}$.

Now, let's analyze its bidirectional machine, which is shown in Fig. 1.5(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.

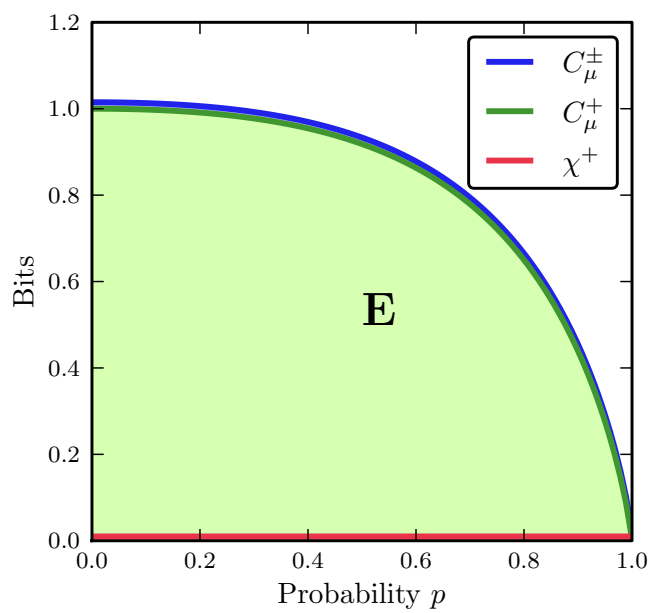

Figure 1.6: The Even Process's information processing properties—$C_\mu^\pm$, C_μ^+, and χ^+—as its self-loop probability p varies. The colored area bounded by the curves show the magnitude of $\mathbf{E}$.

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_μ^+, C_μ^-, χ^+, χ^-, and Ξ, though, it is quite simple. This illustrates one strength of the time-symmetric analysis. The latter's new and independent set of informational measures lead one to explore new regions of process space

(see Fig. 1.6) 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. 1.5.

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

Plotting C_{μ}^{+}, $C_{\mu}^{\pm}$, and χ^{+}, Fig. 1.6 rather directly illustrates these properties and shows that they are maintained across the entire process family as the self-loop probability p is varied.

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

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

Its forward ϵ-machine has two recurrent causal states $\mathcal{S}^+ = \{A, B\}$ 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 1.7(a) gives M^+, while (b) gives M^-. We see that the ϵ-machines are the same and so the Golden Mean Process is causally reversible ($\Xi = 0$).

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

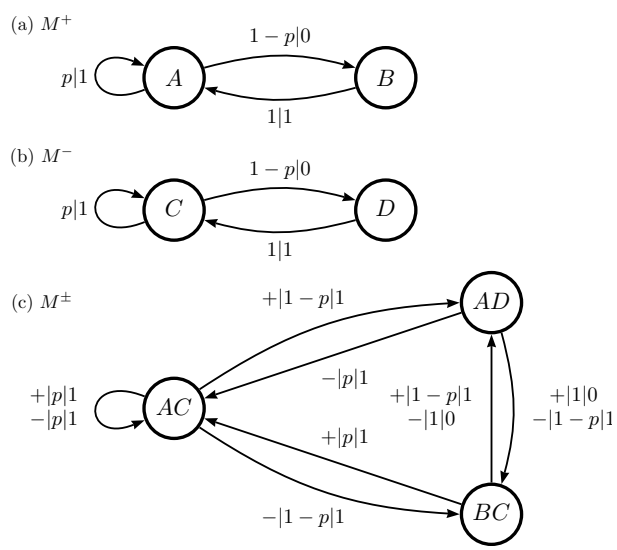

Figure 1.7: Forward and reverse ϵ-machines for the Golden Mean Process: (a) M^+ and (b) M^-. (c) The bidirectional machine $M^\pm$.

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. (1.2).) Thus,

$$\mathbf{E} = H\left(\frac{1}{2-p}\right) - \frac{H(p)}{2-p}\,. \tag{1.28}$$

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. 1.8.

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

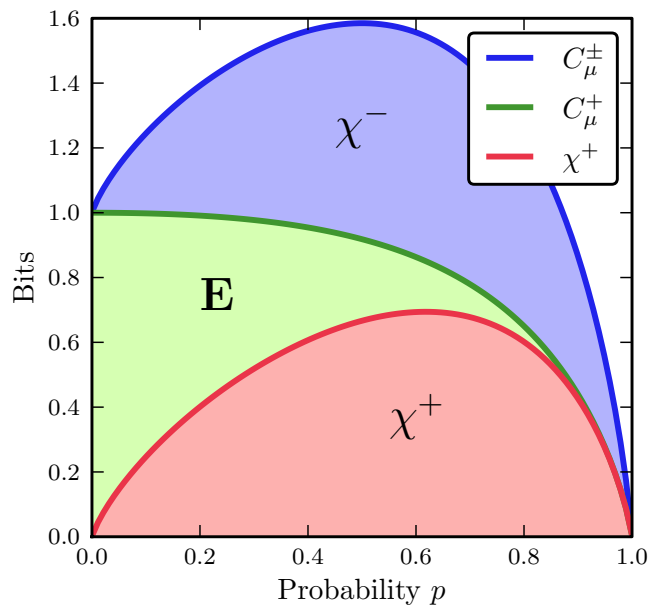

Figure 1.8: The Golden Mean Process's information processing measures—$C_\mu^{\pm}$, C_μ^{+}, and χ^{+}—as its self-loop probability p varies. Colored areas bounded by the curves give the magnitude at each p of χ^{-}, $\mathbf{E}$, and χ^{+}.

The origin of its crypticity is found by analyzing the bidirectional machine, which is shown in Fig. 1.7(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_{μ}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 χ^{+} bits of memory and a generation of χ^{-} bits of uncertainty.

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

Figure 1.8 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: $C_{\mu}^{\pm} \approx \chi^{+} + \chi^{-}$.

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

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

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

Figure 1.9(b) shows M^- which has four recurrent causal states $\boldsymbol{\mathcal{S}}^- = \{D, E, F, G\}$. We see that the ϵ-machines are not the same and so the RIP is causally irreversible. A direct calculation gives:

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

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

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

Let's analyze the RIP bidirectional machine, which is shown in Fig. 1.9(c) for $p = q = 1/2$.

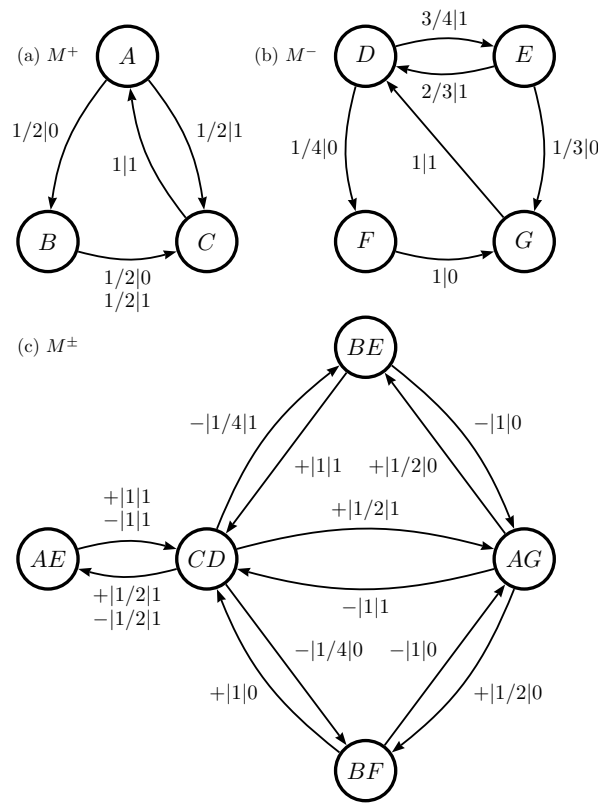

Figure 1.9: Forward and reverse ϵ-machines for the RIP with $p = q = 1/2$: (a) M^+ and (b) M^-. (c) The bidirectional machine $M^\pm$ also for $p = q = 1/2$. (Reprinted with permission from Refs. [**?**].)

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_μ^+ and the last is χ^+.

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

Figure 1.10 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.

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

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

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

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

structural properties.

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

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

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

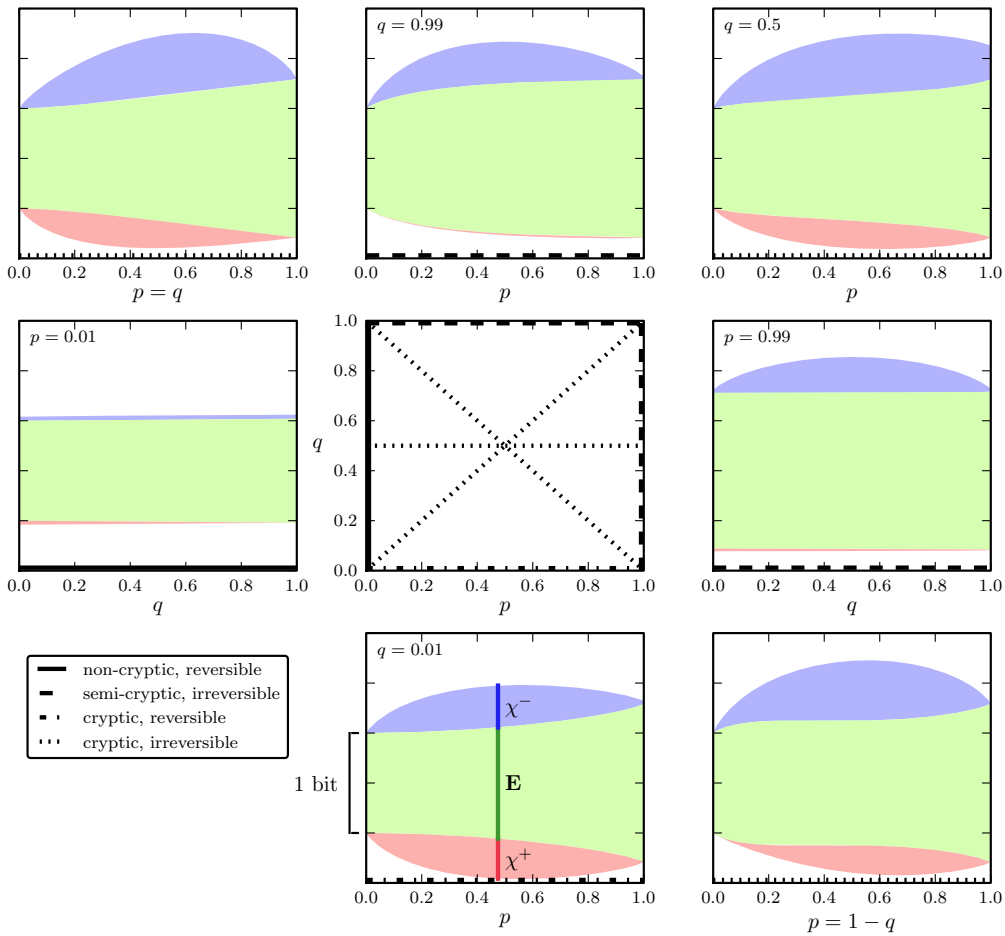

Figure 1.10: The Random Insertion Process's information processing measures as its two probability parameters p and q vary. The central square shows the (p, q) parameter space, with solid and dashed lines indicating the paths in parameter space for each of the other information versus parameter plots. 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 χ^-, $\mathbf{E}$, and χ^+.