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

Christopher J. Ellison,[1, *] John R. Mahoney,[1, †] and James P. Crutchfield[1, 2, ‡]

[1]*Complexity Sciences Center and Physics Department,
University of California at Davis, One Shields Avenue, Davis, CA 95616*
[2]*Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501*
(Dated: May 19, 2009)

We introduce an ambidextrous view of stochastic dynamical systems, comparing their forward-time and reverse-time representations and then integrating them into a single time-symmetric representation. The perspective is useful theoretically, computationally, and conceptually. Mathematically, we prove that the excess entropy—a familiar measure of organization in complex systems—is the mutual information not only between the past and future, but also between the predictive and retrodictive causal states. Practically, we exploit the connection between prediction and retrodiction to directly calculate the excess entropy. Conceptually, these lead one to discover new system invariants for stochastic dynamical systems: crypticity (information accessibility) and causal irreversibility. Ultimately, we introduce a time-symmetric representation that unifies all these quantities, compressing the two directional representations into one. The resulting compression offers a new conception of the amount of information stored in the present.

PACS numbers: 02.50.-r 89.70.+c 05.45.Tp 02.50.Ey

INTRODUCTION

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

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

We first review minimal causal representations of stochastic processes, as developed by *computational mechanics* [1, 2]. We extend its (implied) forward-time representation to reverse-time. Then, we prove that the mutual information between a process's past and future—the *excess entropy*—is the mutual information between its forward- and reverse-time representations.

Excess entropy, and related mutual information quantities, are widely used diagnostics for complex systems. They have been applied to detect the presence of organization in dynamical systems [3, 4, 5, 6], in spin systems [7, 8, 9], in neurobiological systems [10, 11], 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 nonergodic organization necessary for human communication [12, 13].

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 [1]—and the observed information—the excess entropy. Another consequence is that the forward and reverse representations are two projections of a unified time-symmetric representation. From the latter it becomes clear there are important system invariants 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 refer the reader to Ref. [14] for complementary results, that we do not cover here, on the measure-theoretic relationships between the above information quantities. The announcement of those results and those in the present work appeared in Ref. [15]. Here we lay out the theory in detail, giving step-by-step proofs of the main results and the calculational methods.

OPTIMAL CAUSAL MODELS

Our approach starts with a simple analogy. Any process $\Pr(\overleftarrow{X}, \overrightarrow{X})$ is also a *communication channel* with a specified input distribution $\Pr(\overleftarrow{X})$ [32]: It transmits information from the *past* $\overleftarrow{X} = \ldots X_{-3}X_{-2}X_{-1}$ to the

future $\overrightarrow{X} = X_0 X_1 X_2 \ldots$ by storing it in the present. X_t is the random variable for the measurement outcome at time t. Our goal is also simply stated: We wish to predict the future using information from the past. At root, a prediction is probabilistic, specified by a distribution of possible futures $\overrightarrow{X}$ given a particular past $\overleftarrow{x}$: $\Pr(\overrightarrow{X}|\overleftarrow{x})$. At a minimum, a good predictor needs to capture *all* of the information I shared between past and future: $\mathbf{E} = I[\overleftarrow{X}; \overrightarrow{X}]$—the process's *excess entropy* [16, and references therein].

Consider now the goal of modeling—building a representation that allows not only good prediction but also expresses the mechanisms producing a system's behavior. To build a model of a structured process (a memoryful channel), computational mechanics [1] introduced an equivalence relation $\overleftarrow{x} \sim \overleftarrow{x}'$ that groups all histories which give rise to the same prediction:

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

In other words, for the purpose of forecasting the future, two different pasts are equivalent if they result in the same prediction. The result of applying this equivalence gives the process's *causal states* $\boldsymbol{\mathcal{S}} = \Pr(\overleftarrow{X}, \overrightarrow{X})/\sim$, which partition the space $\overleftarrow{\mathbf{X}}$ of pasts into sets that are predictively equivalent. The set of causal states [33] can be discrete, fractal, or continuous; see, e.g., Figs. 7, 8, 10, and 17 in Ref. [17].

State-to-state transitions are denoted by matrices $T_{\mathcal{S}\mathcal{S}'}^{(x)}$ whose elements give the probability $\Pr(X = x, \mathcal{S}'|\mathcal{S})$ of transitioning from one state $\mathcal{S}$ to the next $\mathcal{S}'$ on seeing measurement x. The resulting model, consisting of the causal states and transitions, is called the process's ϵ-*machine*. Given a process $\mathcal{P}$, we denote its ϵ-machine by $M(\mathcal{P})$.

Causal states have a Markovian property that they render the past and future statistically independent; they *shield* the future from the past [2]:

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

Moreover, they are optimally predictive [1] 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 [2] that the causal states capture all of the information shared between past and future: $I[\mathcal{S}; \overrightarrow{X}] = \mathbf{E}$.

ϵ-Machines have an important, if subtle, structural property called *unifilarity* [1, 18]: From the start state, each observed sequence $\ldots x_{-3}x_{-2}x_{-1}\ldots$ corresponds to one and only one sequence of causal states [34]. ϵ-Machine unifiliarity underlies many of the results here. Its importance is reflected in the fact that representations without unifilarity, such as general hidden Markov models, *cannot* be used to directly calculate important system properties—including the most basic, such as, how

random a process is. Nonetheless, unifilarity is easy to verify: For each state, each measurement symbol appears on at most one outgoing transition [35]. The signature of unifilarity is that on knowing the current state and measurement, the uncertainty in the next state vanishes: $H[\mathcal{S}_{t+1}|\mathcal{S}_t, X_t] = 0$. In summary, a process's ϵ-machine is its unique minimal unifilar model.

INFORMATION PROCESSING INVARIANTS

Out of all optimally predictive models $\widehat{\boldsymbol{\mathcal{R}}}$—for which $I[\widehat{\mathcal{R}}; \overrightarrow{X}] = \mathbf{E}$—the ϵ-machine captures the minimal amount of information that a process must store in order to communicate all of the excess entropy from the past to the future. This is the Shannon information contained in the causal states—the *statistical complexity* [2]: $C_\mu \equiv H[\mathcal{S}] \leq H[\widehat{\mathcal{R}}]$. In short, $\mathbf{E}$ is the effective information transmission rate of the process, viewed as a channel, and C_μ is the sophistication of that channel.

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

In addition to $\mathbf{E}$ and C_μ, another key (and historically prior) invariant for dynamical systems and stochastic processes is the entropy rate:

$$h_\mu = \lim_{L \to \infty} \frac{H(L)}{L} , \qquad (3)$$

where $H(L)$ is Shannon entropy of length-L sequences X^L. This is the per-measurement rate at which the process generates information—its degree of intrinsic randomness [19, 20].

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

$$\begin{aligned} h_\mu &= H[X|\mathcal{S}] \\ &= -\sum_{\{\mathcal{S}\}} \Pr(\mathcal{S}) \sum_{\{x\}} T_{\mathcal{S}\mathcal{S}'}^{(x)} \log_2 T_{\mathcal{S}\mathcal{S}'}^{(x)} . \end{aligned} \qquad (4)$$

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

$$\begin{aligned} C_\mu &= H[\mathcal{S}] \\ &= -\sum_{\{\mathcal{S}\}} \Pr(\mathcal{S}) \log_2 \Pr(\mathcal{S}) . \end{aligned} \qquad (5)$$

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

EXCESS ENTROPY

Until recently, $\mathbf{E}$ could not be as directly calculated as the entropy rate and the statistical complexity. This state of affairs was a major roadblock to analyzing the relationships between modeling and predicting and, more concretely, the relationships between (and even the interpretation of) a process's basic invariants—h_μ, C_μ, and $\mathbf{E}$. Ref. [15] announced the solution to this longstanding problem by deriving explicit expressions for $\mathbf{E}$ in terms of the ϵ-machine, providing a unified information-theoretic analysis of general processes. Here we provide a detailed account of the underlying methods and results.

To get started, we should recall what is already known about the relationships between these various quantities. First, some time ago, an explicit expression was developed from the Hamiltonian for one-dimensional spin chains with range-R interactions [8]:

$$\mathbf{E} = C_\mu - R\,h_\mu \ . \qquad (6)$$

It was demonstrated that $\mathbf{E}$ is a generalized order parameter: Compared to structure factors, $\mathbf{E}$ is an assumption-free way to find structure and correlation in spin systems that does not require tuning [9].

Second, it has also been known for some time that the statistical complexity is an upper bound on the excess entropy [18]:

$$\mathbf{E} \leq C_\mu \ . \qquad (7)$$

Nonetheless, other than the special, if useful, case of spin systems, until Ref. [15] there had been no direct way to calculate $\mathbf{E}$. Remedying this limitation required broadening the notion of what a process is.

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 [21].

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 corre-

sponding entropy rates, h_μ^+ and h_μ^-, and statistical complexities, $C_\mu^+ = H[\mathcal{S}^+]$ and $C_\mu^- = H[\mathcal{S}^-]$, respectively, to the process.

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

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

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

Proposition 1. *[2] 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:*

$$H[X_{-L+2},\ldots,X_0] = H[X_{-L+1},\ldots,X_{-1}] \ . \qquad (8)$$

Keeping this in mind, we directly calculate:

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

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

Proposition 2. *[21] There exist stationary processes for which* $C_\mu^- \neq C_\mu^+$.

Proof. *The random-insertion process, analyzed in a later section, establishes this by example.*

Note that $\mathbf{E}$ is mute on this score. Since the mutual information I is symmetric in its variables [22], $\mathbf{E}$ is time symmetric. Proposition 2 puts us on notice that $\mathbf{E}$ necessarily misses many of a process's structural properties.

EXCESS ENTROPY FROM CAUSAL STATES

The relationship between predicting and retrodicting a process, and ultimately $\mathbf{E}$'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 [14] 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. [23] which showed that Shannon entropy $H[\cdot]$ and mutual information $I[\cdot; \cdot]$ form a signed measure over the space of events. Practically, there is a direct correspondence between set theory and these information measures. Using this, Ref. [14] developed an *ϵ-machine information diagram* over four variables, which gives a minimal set of entropies, conditional entropies, mutual informations, and conditional mutual informations necessary to analyze the relationships among h_μ, C_μ, and $\mathbf{E}$ for general stochastic processes.

In a generic four-variable information diagram, there are 15 independent variables. Fortunately, this greatly simplifies in the case of using an ϵ-machine to represent a process; there are only 5 independent variables in the ϵ-machine information diagram [14]. (These results are announced in [15]; see Fig. 1 there.)

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

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

$$\mathbf{E} = I[\mathcal{S}^+; \mathcal{S}^-] \ . \tag{9}$$

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

$$I[\overleftarrow{X}; \overrightarrow{X}; \mathcal{S}^+; \mathcal{S}^-] = I[\overleftarrow{X}; \overrightarrow{X}]$$
$$= \mathbf{E} \ , \tag{10}$$

on the one hand, and

$$I[\overleftarrow{X}; \overrightarrow{X}; \mathcal{S}^+; \mathcal{S}^-] = I[\mathcal{S}^+; \mathcal{S}^-] \ , \tag{11}$$

on the other. □

That is, the process's effective channel capacity $\mathbf{E} = I[\overleftarrow{X}; \overrightarrow{X}]$ 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:*

$$C_\mu^+ = \mathbf{E} + H[\mathcal{S}^+|\mathcal{S}^-] \text{ and} \tag{12}$$
$$C_\mu^- = \mathbf{E} + H[\mathcal{S}^-|\mathcal{S}^+] \ . \tag{13}$$

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

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

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

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

THE BIDIRECTIONAL MACHINE

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

Definition. *Let $M^\pm$ denote the* bidirectional machine *given by the equivalence relation $\sim^\pm$ [36]:*

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

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 [2].

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 I:

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 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. $\square$

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

$$C_{\mu}^{\pm} \equiv H[\mathcal{S}^{\pm}] = H[\mathcal{S}^+, \mathcal{S}^-] . \tag{14}$$

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

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

Thus, we are led to a wholly new interpretation of the excess entropy—in addition to the original three discussed in Ref. [16]: $\mathbf{E}$ 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_{μ}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 nonnegativity of mutual information.* $\square$

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.

Upper Bounds

Here we give new, tighter bounds for $\mathbf{E}$ than Eq. (7) and greatly simplified proofs than those provided in Refs. [2] and [18].

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 $\mathbf{E} \leq C_{\mu}^+$.* $\square$

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 $\Pr(\overleftarrow{X}, \overrightarrow{X})$. A microscopically reversible process scanned in both directions yields the same word distribution; 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. And these are M^+ and M^-, respectively.* $\square$

Corollary 4. *For a microscopically reversible process, $C_{\mu}^- = C_{\mu}^+$.*

Proof. *For a microscopically reversible process $M^- = M^+$. And so, in particular, $\mathcal{S}^- = \mathcal{S}^+$, their transition matrices are the same, and so $\Pr(\mathcal{S}^-) = \Pr(\mathcal{S}^+)$. Thus, $C_\mu^- = C_\mu^+$.* $\square$

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

Definition. *A process's* causal irreversibility *[21] is:*

$$\Xi(\mathcal{P}) = C_\mu^+ - C_\mu^- \ . \tag{15}$$

Corollary 5. $\Xi(\mathcal{P}) = H[\mathcal{S}^+|\mathcal{S}^-] - H[\mathcal{S}^-|\mathcal{S}^+]$.

Note that $\Xi = 0$ does not imply that $M^+ = M^-$. For example, the periodic process $\ldots 123123123 \ldots$ is not microscopically reversible, since $\Pr(123) \neq \Pr(321)$. However, $\Xi = 0$, as $C_\mu^- = C_\mu^+ = \log_2 3$.

It turns out, though, that we are more interested in the following situation.

Proposition 8. *If $\Xi(\mathcal{P}) \neq 0$, then the process is not microscopically reversible.*

Proof. $C_\mu^+ \neq C_\mu^-$ *implies that $M^+ \neq M^-$. And so, $\mathcal{P}^+ \neq \mathcal{P}^-$.* $\square$

So, a vanishing Ξ will indicate "reversibility" for some classes of processes that are not microscopically reversible. The periodic process just described is one such example. In fact, this includes any process whose left- and right-scan processes are isomorphic under a simultaneous measurement-alphabet and causal-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 irreversibility. Interestingly, it appears, based on several case studies, that causal reversibility captures *exactly* that notion. That is, it would seem there are no processes for which $\Xi = 0$, yet $\mathcal{P}^+ \not\approx \mathcal{P}^-$. 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.

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 [24]. 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. [25].

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:*

$$\chi(M^+, M^-) = H[\mathcal{S}^+|\mathcal{S}^-] + H[\mathcal{S}^-|\mathcal{S}^+] \ . \tag{16}$$

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

Proof. *$\chi(M^+, M^-)$ is nonnegative, symmetric, and satisfies a triangle inequality. These follow from the solution of exercise 2.9 of Ref. [22]. See also, Ref. [26].* $\square$

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

$$C_\mu^\pm = \mathbf{E} + \chi \ . \tag{17}$$

Proof. *This follows directly from the corollary and the predictive and retrodictive statistical complexity relations, Prop. (12) and (13).* $\square$

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

Note that if crypticity is low $\chi \approx 0$, 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, $\chi \approx C_\mu^\pm$, then little of it's 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 6. *$M^\pm$'s statistical complexity bounds the process's crypticity:*

$$C_\mu^\pm \geq \chi \ . \tag{18}$$

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

Thus, a truly cryptic process has $C_\mu^\pm = \chi$ 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 χ 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. [15] and Ref. [14]. They also lead to a new classification scheme for stationary processes; see Ref. [27]. In the following, we concentrate instead on how to calculate the preceding quantities, giving a complete informational and structural analysis of general processes.

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.

Time-Reversed Presentation

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

$$T_{\mathcal{R}\mathcal{R}'}^{(x)} \equiv \Pr(X = x, \mathcal{R}'|\mathcal{R}) . \tag{19}$$

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

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

$$\begin{aligned}\widetilde{T}_{\mathcal{R}'\mathcal{R}}^{(x)} &\equiv \Pr(X = x, \mathcal{R}|\mathcal{R}') \\ &= T_{\mathcal{R}\mathcal{R}'}^{(x)} \frac{\Pr(\mathcal{R})}{\Pr(\mathcal{R}')} .\end{aligned} \tag{20}$$

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

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

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

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

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

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

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

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. [28], 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:

$$\mu(w) \equiv \Pr(\mathcal{R}_L|X_0^L = w) \tag{21}$$

$$= \frac{\Pr(X_0^L = w, \mathcal{R}_L)}{\Pr(X_0^L = w)} \tag{22}$$

$$= \frac{\pi T^{(w)}}{\pi T^{(w)} \mathbf{1}}, \tag{23}$$

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:

$$\Pr(x, \mu(wx)|\mu(w)) \equiv \frac{\Pr(wx)}{\Pr(w)} \tag{24}$$

$$= \mu(w) T^{(x)} \mathbf{1} . \tag{25}$$

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

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:

$$
\begin{array}{ccc}
M^+ & \xleftarrow{\;\mathcal{U}\;} & \widetilde{M}^- \\
\tau \downarrow & & \uparrow \tau \\
\widetilde{M}^+ & \xrightarrow[\mathcal{U}]{} & M^-
\end{array} \tag{26}
$$

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

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

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

Returning to the two-step procedure, one must construct the mixed-state presentation of $\widetilde{M}^+$. It is helpful to keep the hidden process lattice of Table I 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 $\nu(w)$:

$$
\nu(w) \equiv \Pr(\mathcal{R}_0|X_0^L = w) \tag{27}
$$
$$
= \frac{\Pr(\mathcal{R}_0, X_0^L = w)}{\Pr(X_0^L = w)} \tag{28}
$$
$$
= \frac{\pi T^{(\widetilde{w})}}{\pi T^{(\widetilde{w})}\mathbf{1}} , \tag{29}
$$

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

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

$$
\Pr(x, \nu(xw)|\nu(w)) \equiv \frac{\Pr(xw)}{\Pr(w)} \tag{30}
$$
$$
= \nu(w)T^{(x)}\mathbf{1}. \tag{31}
$$

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

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

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

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

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

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 [22].

The forward ϵ-machine has three recurrent causal states $\mathcal{S}^+ = \{A, B, C\}$ and is shown in Fig. 1(a). The transition matrices $T^{(x)}$ specify $\Pr(X_0 = x, \mathcal{S}_1^+|\mathcal{S}_0^+)$ and are given by:

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

and

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

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

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

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

Using the $T^{(x)}$ and π, we create the time-reversed presentation $\widetilde{M}^+ = \mathcal{T}(M^+)$. This is shown in Fig. 1(b). Notice that the machine is not unifilar, and so it is clearly not an ϵ-machine. The transition matrices for the time-reversed presentation are given by:

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

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

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(c). Generally, causal states can be categorized into types [28]. Of these, the calculation of **E** depends only on the reachable recurrent causal states. The construction of the mixed-state presentation will generate other types of causal states, such as transient causal states, but we eventually remove them.

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

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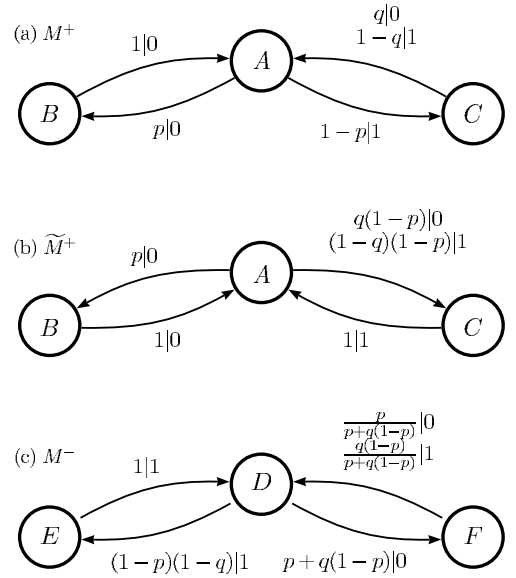$$= \frac{(p, p, q(1-p))}{2p + q(1-p)} .$$

FIG. 1: The presentations used to calculate the excess entropy for the RnC Process: (a) M^+, (b) $\widetilde{M}^+ = \mathcal{T}(M^+)$, and (c) $M^- = \mathcal{U}(\widetilde{M}^+)$. Edge labels $t|x$ give the probability $t = T_{\mathcal{R}\mathcal{R}'}^{(x)}$ of making a transition and seeing symbol x.

and

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

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

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

Note that

$$\nu(10) = \frac{\nu(0)\widetilde{T}^{(1)}}{\nu(0)\widetilde{T}^{(1)}\mathbf{1}} . \tag{32}$$

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

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

To illustrate calculating the transition probabilities, consider the transition from $\nu(00)$ to $\nu(100)$ [37]. By Eq. (31), we have

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

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

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

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

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

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

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

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.

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.

Even Process

The Even Process is a stochastic generalization of the Even System: the canonical example of a *sofic subshift*—a symbolic dynamical system that cannot be expressed as a subshift of finite type [16, 29]. Although it has only two recurrent causal states, the Even Process cannot be expressed as any finite Markov chain over measurement sequences. Somewhat surprisingly, it 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 $\chi = 0$. All of its internal state information is present in measurements; we call it an *explicit*, or *non-cryptic* process.

Its forward ϵ-machine has two recurrent causal states $\boldsymbol{\mathcal{S}}^+ = \{A, B\}$ and transition matrices [16]:

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

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

Figure 2(a) gives M^+, while 2(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 invariants as a function of the probability $p = \Pr(0|A)$ of the self-loop. A simple calculation shows that

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

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

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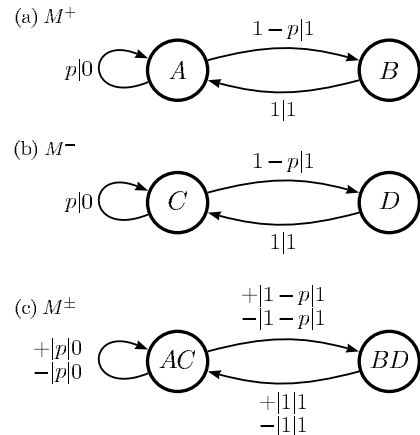

FIG. 2: Forward and reverse ϵ-machines for the Even Process: (a) M^+ and (b) M^-. (c) The bidirectional machine $M^\pm$. Edge labels are prefixed by the scan direction $\{-, +\}$.

Now, let's analyze its bidirectional machine, which is shown in Fig. 2(c). The reverse and forward maps are

given by:

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

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

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.

Direct inspection of M^+ and M^- shows that both ϵ-machines are reverse unifilar. And this is reflected in the fact that $C_\mu^+ = C_\mu^- = \mathbf{E}$; verifying a proposition of Ref. [27].
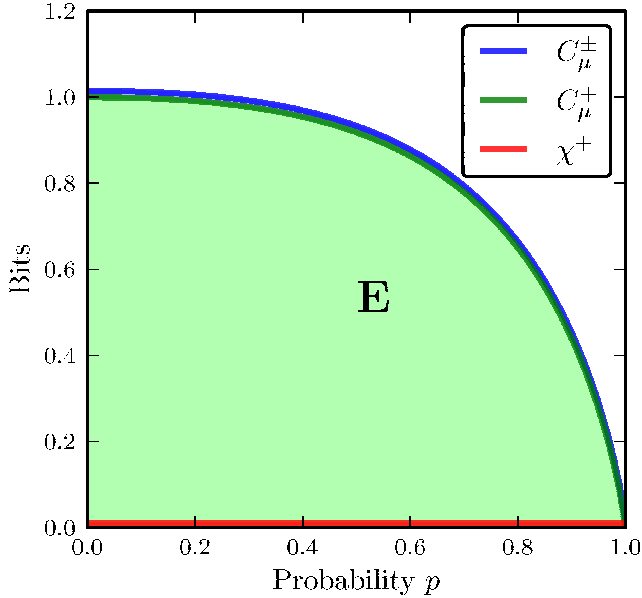

FIG. 3: The Even Process's information processing properties—$C_\mu^\pm$, C_μ^+, and χ^+—as its self-loop probability p varies. The colored area bounded by the curves show the magnitude of $\mathbf{E}$.

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. 3) and to ask structural questions not previously capable of being asked (or answered, for that matter). To see exactly why the Even Process is so simple, let's look at its causal states.

Its histories can be divided into two classes: those that end with an even number of 1s and those that end with an odd number of 1s. Similarly, its futures divide into two classes: those that 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. 2.

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 ($\chi = 0$). 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. 3 rather directly illustrates these properties and shows that they are maintained across the entire process family as the self-loop probability p is varied.

Golden Mean Process

The Golden Mean Process generates all binary sequences except for those with two contiguous 0s. Like the Even Process, it has two recurrent causal states. Unlike the Even Process, its support is a subshift of finite type; 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, $\chi > 0$ 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 [16]:

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

and

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

Figure 4(a) gives M^+, while (b) gives M^-. We see that the ϵ-machines are the same and so the Golden Mean Process is causally reversible ($\Xi = 0$).

Again, we can give general expressions for the information processing invariants 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.

(a) M^+

(b) M^-

(c) $M^\pm$
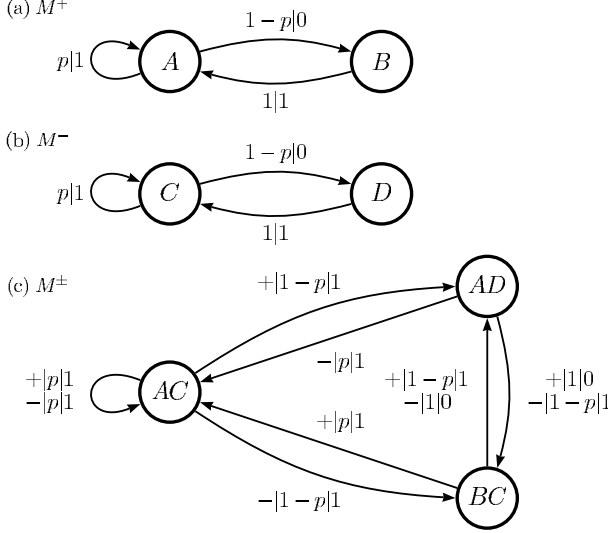
FIG. 4: Forward and reverse ϵ-machines for the Golden Mean Process: (a) M^+ and (b) M^-. (c) The bidirectional machine $M^\pm$.

However, since $\chi \neq 0$ 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. (6).) Thus,

$$\mathbf{E} = H\left(\frac{1}{2-p}\right) - \frac{H(p)}{2-p} \; . \tag{33}$$

Putting these closed-form expressions together gives us a graphical view of how the various information invariants change as the process's parameter is varied. This is shown in Fig. 5.

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

The origin of its crypticity is found by analyzing the bidirectional machine, which is shown in Fig. 4(c). The reverse and forward maps are given by:

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

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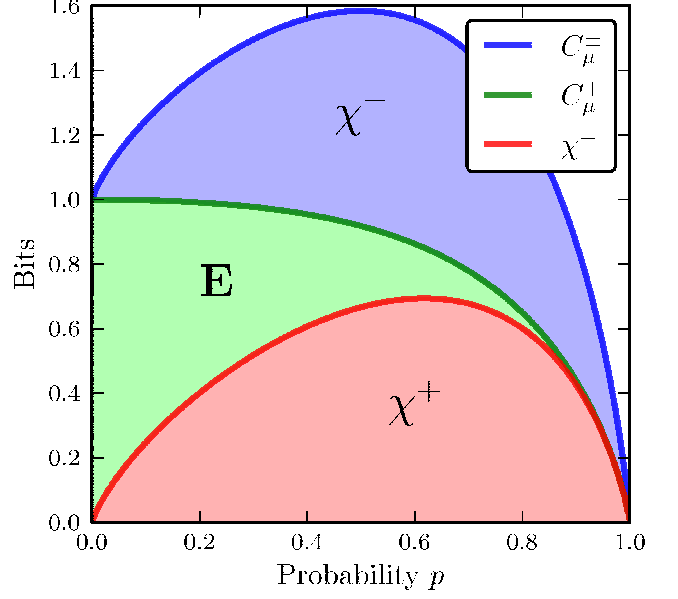

FIG. 5: The Golden Mean Process's information processing invariants—$C_\mu^\pm$, C_μ^+, and χ^+—as its self-loop probability p varies. Colored areas bounded by the curves give the magnitude at each p of χ^-, $\mathbf{E}$, and χ^+.

From $M^\pm$, one can calculate the stationary distribution over the bidirectional causal states: $\Pr(\mathcal{S}^\pm) = \Pr(AC, AD, BC) = (p, 1-p, 1-p)/(2-p)$. For $p = 1/2$, we obtain $C_\mu^\pm = H[\mathcal{S}^\pm] = \log_2 3 \approx 1.5850$ bits, but an $\mathbf{E} = I[\mathcal{S}^+; \mathcal{S}^-] = 0.2516$ bits. Thus, $\mathbf{E}$ is substantially less that the C_μs, a cryptic process: $\chi \approx 1.3334$ 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 5 shows that these properties are maintained across the entire Golden Mean Process family, except at extremes. When $p = 0$, it degenerates to a simple period-2 process, with $\mathbf{E} = C_\mu^+ = C_\mu^- = C_\mu^\pm = 1$ bit of memory. When $p = 1$, it is even simpler, the period-1 process, with no memory. As it approaches this extreme, $\mathbf{E}$ vanishes rapidly, leaving processes with internal state memory dominated by crypticity: $C_\mu^\pm \approx \chi^+ + \chi^-$.

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 $\mathcal{S}^+ = \{A, B, C\}$ and transition matrices:

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

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

Figure 6(b) shows M^- which has four recurrent causal states $\mathcal{S}^- = \{D, E, F, G\}$. We see that the ϵ-machines are not the same and so the RIP is causally irreversible. A direct calculation gives:

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

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

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.

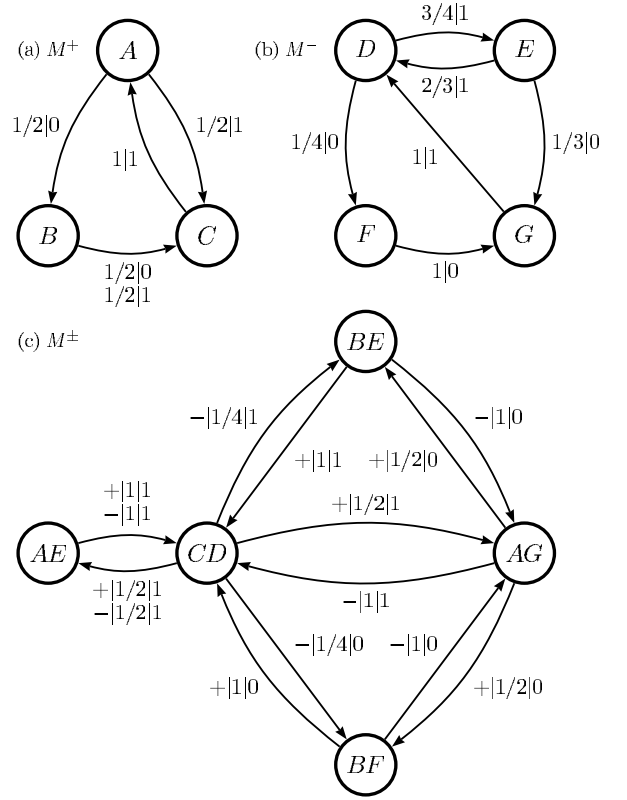Let's analyze the RIP bidirectional machine, which is shown in Fig. 6(c) for $p = q = 1/2$. The reverse and

FIG. 6: Forward and reverse ϵ-machines for the RIP with $p = q = 1/2$: (a) M^+ and (b) M^-. (c) The bidirectional machine $M^\pm$ also for $p = q = 1/2$. (Reprinted with permission from Ref. [15].)

forward maps are given by:

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

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

Or, for general p and q, we have

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

By way of demonstrating the exact analysis now possible, $\mathbf{E}$'s closed-form expression for the RIP family is

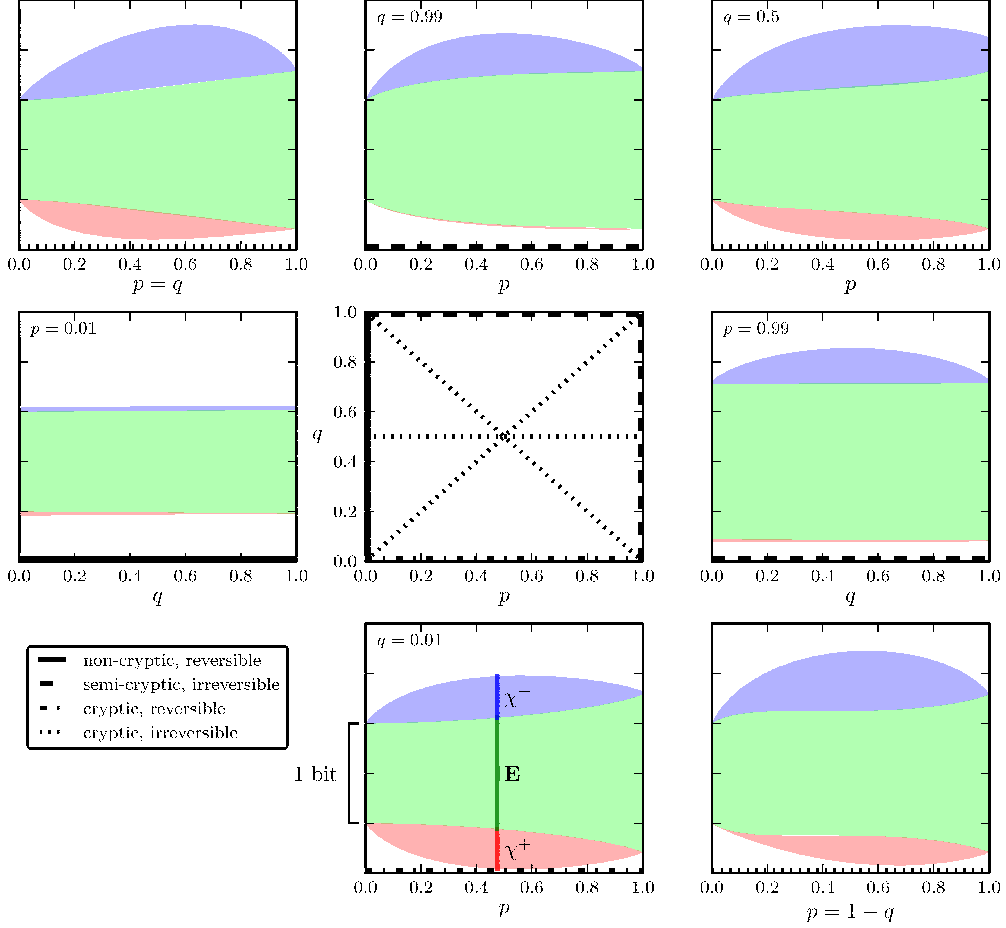$$\mathbf{E} = \log_2(p+2) - \frac{p \log_2 p}{p+2} - \frac{1-pq}{p+2} H\left(\frac{1-p}{1-pq} \right).$$

FIG. 7: The Random Insertion Process's information processing invariants 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^+.

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 $\mathbf{E} = I[\mathcal{S}^+; \mathcal{S}^-] = 1.2464$ bits. That is, the excess entropy (the apparent information) is substantially less than the statistical complexities (stored information)—a moderately cryptic process: $\chi \approx 0.8755$ bits.

Figure 7 shows how the RIP's informational character varies along one-dimensional paths in its parameter space: $(p, q) \in [0, 1]^2$. The four extreme-p and -q paths illustrate that the RIP borders on (i) noncryptic, 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.

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 informational invariants that emerged and captured novel properties of general processes.

Excess entropy, a familiar quantity in a long-applied family of mutual informations, is often estimated [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13] 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 invariants than others—such as, histograms of word counts or general hidden Markov models. The differences between prediction and modeling unearthed new informational quantities—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 χ. In this light, microscopic reversibility appears to be too strong a criterion, missing important structural properties.

The time-agnostic perspective hinged on expanding the space of representations. First, we described parallel predictive and retrodictive causal models joined by the switching maps. We then introduced a bidirectional machine that compressed C_μ^+ and C_μ^- into $C_\mu^\pm$. The associated joint causal-state space allowed us to make rather nonintuitive statements about prediction (retrodiction) conditioned on these joint states. The operational meaning of the bidirectional machine certainly warrants further attention. It also 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.

APPENDIX: THE MIXED-STATE PRESENTATION IS SUFFICIENT TO CALCULATE THE SWITCHING MAPS

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

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

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

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

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

We now show that B is deterministic.

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

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

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

So, by construction, B is deterministic. □

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

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

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

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

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

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

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

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

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

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

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

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

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

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 $\mathcal{R}_B$, we determine if $\mathcal{R}_B \in \mathcal{S}_B$ for each $\mathcal{R}_B$. Lastly, we compute the stationary distribution over the states of B and divide by the stationary probability of the corresponding causal state.

In effect, this establishes a general method for computing the conditional probability of states from the "input" machine given a state of the "resultant" machine.

We can now recall the specific context of forward and reverse ϵ-machines and apply this technique to calculate **E** in the case where the resultant machine $\mathcal{T}(M^+)$ is not an ϵ-machine.

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

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

now becomes:

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

or, more specifically,

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

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

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

[1] J. P. Crutchfield and K. Young. Inferring statistical complexity. *Phys. Rev. Let.*, 63:105–108, 1989.
[2] J. P. Crutchfield and C. R. Shalizi. Thermodynamic depth of causal states: Objective complexity via minimal representations. *Phys. Rev. E*, 59(1):275–283, 1999.
[3] A. Fraser. Chaotic data and model building. In H. Atmanspacher and H. Scheingraber, editors, *Information Dynamics*, volume Series B: Physics Vol. 256 of *NATO ASI Series*, page 125, New York, 1991. Plenum.
[4] M. Casdagli and S. Eubank, editors. *Nonlinear Modeling*, SFI Studies in the Sciences of Complexity, Reading, Massachusetts, 1992. Addison-Wesley.
[5] J. C. Sprott. *Chaos and Time-Series Analysis*. Oxford University Press, Oxford, UK, second edition, 2003.
[6] H. Kantz and T. Schreiber. *Nonlinear Time Series Analysis*. Cambridge University Press, Cambridge, UK, second edition, 2006.
[7] D. Arnold. Information-theoretic analysis of phase transitions. *Complex Systems*, 10:143–155, 1996.
[8] J. P. Crutchfield and D. P. Feldman. Statistical complexity of simple one-dimensional spin systems. *Phys. Rev. E*, 55(2):1239R–1243R, 1997.
[9] D. P. Feldman and J. P. Crutchfield. Discovering non-critical organization: Statistical mechanical, information theoretic, and computational views of patterns in simple one-dimensional spin systems. 1998. Santa Fe Institute Working Paper 98-04-026.

[10] G. Tononi, O. Sporns, and G. M. Edelman. A measure for brain complexity: Relating functional segregation and integration in the nervous system. *Proc. Nat. Acad. Sci. USA*, 91:5033–5037, 1994.

[11] W. Bialek, I. Nemenman, and N. Tishby. Predictability, complexity, and learning. *Neural Computation*, 13:2409–2463, 2001.

[12] W. Ebeling and T. Poschel. Entropy and long-range correlations in literary english. *Europhys. Lett.*, 26:241–246, 1994.

[13] L. Debowski. On the vocabulary of grammar-based codes and the logical consistency of texts. *IEEE Trans. Info. Th.*, 2008.

[14] J. P. Crutchfield, C. J. Ellison, and J. Mahoney. ϵ-Machine information measures. *in preparation*, 2008.

[15] J. P. Crutchfield, C. J. Ellison, and J. Mahoney. Time's barbed arrow: Irreversibility, crypticity, and stored information. *submitted*, 2008. arxiv.org:0902.1209 [cond-mat].

[16] J. P. Crutchfield and D. P. Feldman. Regularities unseen, randomness observed: Levels of entropy convergence. *CHAOS*, 13(1):25–54, 2003.

[17] J. P. Crutchfield. The calculi of emergence: Computation, dynamics, and induction. *Physica D*, 75:11–54, 1994.

[18] C. R. Shalizi and J. P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *J. Stat. Phys.*, 104:817–879, 2001.

[19] C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, Champaign-Urbana, 1962.

[20] A. N. Kolmogorov. A new metric invariant of transient dynamical systems and automorphisms in Lebesgue spaces. *Dokl. Akad. Nauk. SSSR*, 119:861, 1958. (Russian) Math. Rev. vol. 21, no. 2035a.

[21] J. P. Crutchfield. Semantics and thermodynamics. In M. Casdagli and S. Eubank, editors, *Nonlinear Modeling and Forecasting*, volume XII of *Santa Fe Institute Studies in the Sciences of Complexity*, pages 317 – 359, Reading, Massachusetts, 1992. Addison-Wesley.

[22] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, second edition, 2006.

[23] R. W. Yeung. A new outlook on Shannon's information measures. *IEEE Trans. Info. Th.*, 37(3):466–474, 1991.

[24] C. E. Shannon. Communication theory of secrecy systems. *Bell Sys. Tech. J.*, 28:656–715, 1949.

[25] N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw. Geometry from a time series. *Phys. Rev. Let.*, 45:712, 1980.

[26] J. P. Crutchfield. Information and its metric. In L. Lam and H. C. Morris, editors, *Nonlinear Structures in Physical Systems - Pattern Formation, Chaos and Waves*, page 119, New York, 1990. Springer-Verlag.

[27] J. P. Crutchfield, C. J. Ellison, and J. Mahoney. Classes of irreversibility and crypticity in finitary processes. *in preparation*, 2008.

[28] D. R. Upper. *Theory and Algorithms for Hidden Markov Models and Generalized Hidden Markov Models*. PhD thesis, University of California, Berkeley, 1997. Published by University Microfilms Intl, Ann Arbor, Michigan.

[29] B. Weiss. Subshifts of finite type and sofic systems. *Monastsh. Math.*, 77:462, 1973.

[30] J. E. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley, Reading, 1979.

[31] Y. Ephraim and N. Merhav. Hidden markov processes. *IEEE Trans. Info. Theory*, 48:1518–1569, 2002.

[32] Throughout, we follow the notation and definitions of Refs. [2, 22]. In addition, when we say $\overrightarrow{X}$, for example, this should be interpreted as a shorthand for using $\overrightarrow{X}^L$ and then taking an appropriate limit, such as $\lim_{L\to\infty}$ or $\lim_{L\to\infty} 1/L$.

[33] 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.

[34] Following terminology in computation theory this is referred to as *determinism* [30]. However, to reduce confusion, here we adopt the practice in information theory to call it the *unifilarity* of a process's representation [31].

[35] Specifically, the transition matrices have at most one nonzero component in each row.

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

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