Taxonomy of Prediction

Alexandra Jurgens[1, *] and James P. Crutchfield[2, †]

[1] *GEOSTAT Team, INRIA – Bordeaux Sud Ouest*
33405 Talence Cedex, France

[2] *Complexity Sciences Center and Physics Department*
University of California at Davis, One Shields Avenue, Davis, CA 95616

(Dated: April 20, 2025)

A prediction makes a claim about a system's future given knowledge of its past. A retrodiction makes a claim about its past given knowledge of its future. We introduce the ambidextrous hidden Markov chain that does both optimally—the bidirectional machine whose state structure makes explicit all statistical correlations in a stochastic process. We introduce an informational taxonomy to profile these correlations via a suite of multivariate information measures. While prior results laid out the different kinds of information contained in isolated measurements, in addition to being limited to single measurements the associated informations were challenging to calculate explicitly. Overcoming these via bidirectional machine states, we expand that analysis to information embedded across sequential measurements. The result highlights fourteen new interpretable and calculable information measures that fully characterize a process' informational structure. Additionally, we introduce a labeling and indexing scheme that systematizes information-theoretic analyses of highly complex multivariate systems. Operationalizing this, we provide algorithms to directly calculate all of these quantities in closed form for finitely-modeled processes.

Keywords: entropy rate, multivariate mutual information, information diagram, information atoms

CONTENTS

I. INTRODUCTION

How much information can be learned from a single measurement? Shannon information theory tells us that, on average, information learned by observing a single realization of a random variable is equivalent to the reduction in our uncertainty over the outcome. So, more information is learned from a fair coin flip than from the outcome of a highly biased one. If the coin flip is one in a sequence, each successive measurement gives the same amount of information. That said, for stochastic processes we are typically not interested solely in the analysis coin flip sequences. Processes of interest typically have a stochastic component mixed in with structure in the form of

* alexandra.jurgens@inria.fr; http://csc.ucdavis.edu/~ajurgens/
† chaos@ucdavis.edu; http://csc.ucdavis.edu/~chaos/

correlations across time—single observations are far from independent events.

The puzzle over the question of statistical dependence has quite a long and illustrious history, going back at least to the 1700s with Jacob Bernoulli [1] and the 1800s with Simeon Poisson [2] and Pafnuty Chebyshev [3]. Its more modern form, though, was initially developed by Andrei Andreevich Markov [4] at the turn of the 20^{th} century. These culminated in the weak Law of Large Numbers, the Central Limit Theorem, and Markov chains—transition probabilities, irreducibility, and stationarity—to mention only a few of the concepts we use today in exploring statistical dependence.

In the 1940s, with the development of information theory Shannon introduced for stationary, discrete-symbol, discrete-time processes the entropy rate, which quantifies how much *new* information we learn upon successive observations of a process, given knowledge of its infinite past [5]. Or, to change the question around, how *predictable* the new measurement is given knowledge of the past. This is the amount of *intrinsic randomness* in each new measurement, maximized for a coin flip and minimized by a constant, or periodic, signal.

This said, the full answer to what we can learn from a single measurement is rather more elaborate than this. Indeed, a full taxonomy of information rates has been developed for the information in an isolated measurement [6]. This taxonomy includes five distinct information measures—conditional entropies and mutual informations—that fully map out the information contained in a single bit, not only the new information.

The following expands the taxonomy beyond isolated measurements to the task of prediction itself—statistical dependence over time. An optimally predictive model of a stochastic process is one whose error rate is bounded below by the process's Shannon entropy rate [7]. When also constrained to be minimal, the optimally predictive model is unique and is a hidden Markov chain (HMC) called the ϵ-machine [8]. The ϵ-machine necessarily captures in its state structure all information in the process required for optimal prediction—which is to say, the long-range historical correlations that impact the future.

There is an equivalent, but complementary task—retrodiction or making claims about the past given the present. Although it is well known that the Shannon entropy rate is time-symmetric for stationary processes, the tasks of optimal prediction and optimal retrodiction are not. Prediction and retrodiction generically require different modeling architectures, even for relatively simple processes. To characterize the informational structure of prediction, one needs to consider not only the architecture of the predictive "forward time" ϵ-machine but also the architecture of the retrodictive "reverse time" ϵ-machine. These architectures capture correlations in the process that impact the present but are not accessible through isolated measurement.

To this end, the following develops the *bidirectional machine*, an ambidextrous hidden Markov chain capable of simultaneous optimal prediction and retrodiction [9]. We show that knowledge of the bidirectional machine allows one to fully characterize a prediction—which we take to be the observation *and* all inaccessible but relevant information in the process—using a taxonomy of fourteen information quantities. Furthermore, and importantly, we show that these are exactly calculable in closed form and do not need to be approximated as the limits of information rates, as the previous "anatomy of a bit" analysis assumed [6].

Given that this setting involves highly multivariate information (n-way correlations across arbitrary times), we first review information theory and introduce a systematic method for generating the set of "irreducible" information atoms for an arbitrary set of random variables. We then apply this to a single time-step of the bidirectional machine, generating fourteen informational atoms. These express the full taxonomy of prediction (and retrodiction). We then relate them to previously-defined information measures and give several worked examples for binary stochastic processes of increasing complexity, along with the algorithms needed.

II. INFORMATION THEORY

To study and characterize processes and their associated models we make use of *Shannon's information theory* [5, 7], a widely-used foundational framework that provides tools to describe how stochastic processes generate, store, and transmit information. First, though, we deviate from our development to briefly recall several basic concepts it requires. The reader familiar with information theory may be comfortable skipping this section, although the notation given in Section II C for finding sets of information atoms of arbitrary random variables will be useful later on.

A. Information Measures

Let X be a discrete-valued random variable defined on a *probability space* $(\mathcal{X}, \Sigma, \mu)$ [10, 11]. We call $\mathcal{X}$ the *event space* or *measurement alphabet* of X and take it to be a finite set. The probability of random variable X taking value x is determined by the *measure* μ: $\Pr(X = x) = \mu(\{x\} \in \mathcal{X})$. That is, we denote instances of random variables by capital Latin letters and specific realizations by lower case.

The most basic quantity in information theory is the *Shannon entropy*—the average amount of information learned upon a single measurement of a random variable. (It is, modulo sign, also the amount of uncertainty one faces when predicting the outcome of the measurement.) The Shannon entropy $\mathrm{H}[X]$ of the random variable X is

defined:

$$\mathrm{H}\left[X\right] = -\sum_{x\in\mathcal{X}} \Pr(X=x)\log_2 \Pr(X=x)\ . \qquad (1)$$

We can also characterize the relationship between a pair of jointly-distributed random variables, say, X and Y. The *joint entropy* $\mathrm{H}\left[X,Y\right]$ is of the same functional form as Eq. (1), applied to the joint distribution $\Pr\left(X,Y\right)$. This can, in principle, be straightforwardly extended to a set of N variables $\mathfrak{X} = \{X_i \mid i \in (1,\ldots,N)\}$.

The *conditional entropy* $\mathrm{H}\left[X \mid Y\right]$ gives the additional information learned from observation of one random variable X given knowledge of another random variable Y. The conditional entropy is given by:

$$\mathrm{H}\left[X \mid Y\right] = \mathrm{H}\left[X,Y\right] - \mathrm{H}\left[Y\right]\ . \qquad (2)$$

The fundamental measure of information shared between random variables is the *mutual information*:

$$\mathrm{I}\left[X;Y\right] = \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} \Pr\left(X=x,Y=y\right)\times \log_2\left(\frac{\Pr\left(X=x,Y=y\right)}{\Pr\left(X=x\right)\Pr\left(Y=y\right)}\right)\ . \qquad (3)$$

The probabilities of both variables are taken over the joint probability distribution, while the single probabilities are taken according to the marginals. The mutual information can also be written in terms of Shannon entropies and conditional entropies:

$$\mathrm{I}\left[X;Y\right] = \mathrm{H}\left[X,Y\right] - \mathrm{H}\left[X \mid Y\right] - \mathrm{H}\left[Y \mid X\right]\ . \qquad (4)$$

Direct inspection shows that the mutual information between two variables is symmetric. When X and Y are statistically independent, the mutual information between them vanishes.

As with entropy, we may condition the mutual information on another random variable Z, giving the *conditional mutual information*:

$$\mathrm{I}\left[X;Y \mid Z\right] = \mathrm{H}\left[X \mid Z\right] + \mathrm{H}\left[Y \mid Z\right] - \mathrm{H}\left[X,Y \mid Z\right]\ . \qquad (5)$$

The conditional mutual information is the amount of information shared by X and Y, given we know the third Z.

Similar to the joint entropy, the mutual information between all three variables—also known as the *interaction information* or the *multivariate mutual information*—is given by the difference between mutual information and conditional mutual information:

$$\mathrm{I}\left[X;Y;Z\right] = \mathrm{I}\left[X;Y\right] - \mathrm{I}\left[X;Y \mid Z\right]\ . \qquad (6)$$

There are two cases worth pointing out here. Two variables X and Y can have positive mutual information but be conditionally independent in the presence of Z, in which case the interaction information is positive. It is also possible, though, for two independent variables to become correlated in the presence of Z, making the conditional mutual information positive and the interaction information negative. In other words, conditioning on a third variable Z can either increase or decrease mutual information and X and Y variables can appear more or less dependent given additional data [7]. That is, there can be *conditional independence* or *conditional dependence* between a pair of random variables. Note that the interaction information is symmetric, so this intuition holds regardless of the conditioning variable selected.

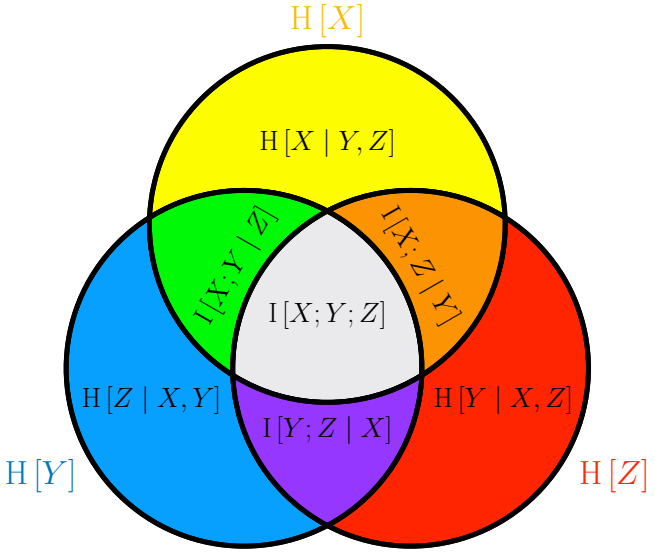

FIG. 1. Information diagram with three random variables, X, Y, and Z.

B. Information Diagrams

We will now make the relationship between information quantities defined in the last section and the algebra of events clear [12]. First, consider only two random variables X and Y. The set of the associated event spaces $\mathfrak{X} = \{\mathcal{X},\mathcal{Y}\}$ induces an algebra $\mathcal{F}$ over $\mathfrak{X}$ closed under complements, unions, and intersections. $\mathcal{F}$ is generated by the partition:

$$F = \left\{\mathcal{X}\setminus\mathcal{Y}, \mathcal{Y}\setminus\mathcal{X}, \mathcal{X}\bigcup\mathcal{Y}, \Omega\setminus\left(\mathcal{X}\bigcup\mathcal{Y}\right)\right\}\ .$$

Note that these elements correspond to the unique areas of an Euler diagram of two overlapping but non-identical and non-empty sets. The algebra $\mathcal{F}$ over $\mathfrak{X}$ is generated by unions over F and so has $2^{|F|} = 2^4 = 16$ elements. We will discuss the case of arbitrarily many variables in the next section, but in general for N variables $|F| = 2^N$ and $|\mathcal{F}| = 2^{2^N}$.

Now, specify a real-valued measure μ^* for each element in F such that:

1. $\mu^*(\mathcal{X} \setminus \mathcal{Y}) = \mathrm{H}[X \mid Y]$,
2. $\mu^*(\mathcal{Y} \setminus \mathcal{X}) = \mathrm{H}[Y \mid X]$,
3. $\mu^*(\mathcal{X} \bigcap \mathcal{Y}) = \mathrm{I}[X;Y]$,
4. $\mu^*(\Omega \setminus (\mathcal{X} \bigcup \mathcal{Y})) = \mu^*(\emptyset) = 0$.

It has been shown that μ^* exists and corresponds uniquely to the joint probability measure on X and Y [13]. In other words, information can be reframed as an additive set function, revealing that there is no semantic difference between "types" of information—entropy, mutual information, and so on—but rather a single underlying quantity being referenced. We call the elements of $\mathcal{F}$ *information atoms*. The elements of F can cannot be decomposed into a sum of other information atoms and are so called the *irreducible* atoms. They circumscribe the range of possible correlations between random variables.

The correspondence between information and the event algebra allows us to represent information quantities via an *information diagram*—an Euler diagram representing the informational relationships between variables. The entropies of some number of random variables—$\mathrm{H}[X]$, $\mathrm{H}[Y]$, $\mathrm{H}[Z]$, and so on—are represented by the area contained in their respective circle. A three-variable example is shown in Fig. 1. When two variables are independent, their respective circles do not overlap. Conditioning corresponds to area subtraction, and shared information to area intersection.

Information diagrams are useful graphical tools but note that they may be misleading—μ^* is a signed measure, but all nonzero atoms are visually portrayed by the i-diagram as having positive area. It is also possible for the informational quantity depicted by an i-diagram to diverge—for instance, the joint entropy of infinitely many random variables—such as in the stochastic processes we will encounter. Furthermore, it is difficult to practically use i-diagrams beyond five or six random variables (unless those random variables have helpful relational structure that limits the size of $\mathcal{F}$). Despite these limitations, i-diagrams remain the tool of choice for visualizing information-theoretic structure in collections of random variables.

C. Information in Collections of Variables

Now, we will show how to find F, $\mathcal{F}$, and μ^* for an arbitrary collection of random variables $\mathfrak{X} = \{X_0, X_1, \ldots, X_k, \ldots, X_{N-1}\}$. To be explicit when taking functions of sets, we borrow the iterable unpacking notation common in modern programming languages. So, we write:

$$f(*A) = f\left(X_0, X_1, \ldots, X_k, \ldots, X_{N-1}\right)$$

where $A = \{X_0, X_1, \ldots, X_k, \ldots, X_{N-1}\}$. We also abuse notation and take all power sets to exclude the empty set by default; i.e., $\mathcal{P}(\mathfrak{X}) = \mathcal{P}(\mathfrak{X}) \setminus \emptyset$. With this notation we concisely write down the interaction information for arbitrary variables as:

$$\mathrm{I}[*\mathfrak{X}] = \sum_{A \in \mathcal{P}(\mathfrak{X})} (-1)^{|A|-1} \mathrm{H}[*A] \ . \tag{7}$$

(Compare to Eq. (6), Eq. (5), and Eq. (4).)

The challenge is to construct the set of irreducible information atoms for a finite random variable set $\mathfrak{X}$ of size N. This set consists of, maximally, N conditional informations, one multivariate mutual information, $2^N - 2 - N$ conditional mutual informations, and the empty set.

First, there is the arbitrary conditional entropy, which breaks down into two entropies:

$$\mathrm{H}[*A \mid *(\mathfrak{X} \setminus A)] = \mathrm{H}[*\mathfrak{X}] - \mathrm{H}[*(\mathfrak{X} \setminus A)] \ , \tag{8}$$

where $A \in \mathcal{P}(\mathfrak{X})$. Then, the arbitrary conditional mutual information is:

$$\mathrm{I}[*A \mid *(\mathfrak{X} \setminus A)] = \sum_{a \in \mathcal{P}(A)} (-1)^{|a|+1} \Big(\mathrm{H}[*(a \cup \mathfrak{X} \setminus A)] - \mathrm{H}[*(\mathfrak{X} \setminus A)] \Big) \ . \tag{9}$$

Notice that when $|A| = 1$, Eq. (9) reduces to Eq. (8) and when $A = \mathfrak{X}$ it reduces to Eq. (7). So, we only need to apply Eq. (9) to each subset $A \in \mathcal{P}(\mathfrak{X})$ find every irreducible information atom—this is equivalent to finding $\mu^*(F)$.

D. Labeling Irreducible Information Atoms

Working with information atoms for arbitrarily many variables very quickly becomes unwieldy due to the exponential growth of the number of atoms. Fortunately, there is a natural ordering for the set of irreducible information atoms. The atoms are labeled by indexing the power set of $\mathfrak{X}$ with an isomorphism to the binary representation of numbers from 1 to $2^N - 1$. We simply indicate the presence of the X_k variable in a subset by the kth digit of the binary sequence—1 if the variable is in the joint distribution and 0 if it is being conditioned on. Recall we exclude the empty set by default.

Notice that this ordering of binary digits is *reversed* compared to the typical representation—compare the Lexicographic column in Table I to the Decimal column. This is due to our primarily working with time-indexed variables and our choosing (rather arbitrarily) to imagine time flowing from left-to-right. Ordering the lexicographic labels from left to right allows easily identifying the semantic meaning of binary strings at a glance.

Given $i \in [1, \ldots, 2^N - 1]$, let A_i be the ith set in $\mathcal{P}(\mathfrak{X})$. The associated irreducible information atom is:

$$\alpha_i = \mathrm{I}[*A_i \mid *(\mathfrak{X} \setminus A_i)] \ , \tag{10}$$

So, the set of irreducible information atoms for $\mathfrak{X}$ is given by:

$$F_{\mathfrak{X}} = \left\{\alpha_i \mid i \in [1, \ldots, 2^N - 1]\right\} \ . \tag{11}$$

The explicit listing of F is given for the $N = 3$ case by Table I, which also gives the *indicial* label of each information atom. This is simply the indices k of the random variables present in the joint distribution. This label is shorter than the lexicographic and often easier to identify at a glance. It is also useful when the index of the random variable carries relational meaning, as it will in our specific use case. The associated i-diagram is depicted in Fig. 1.

This completes our review of basic information theory—a toolset to initiate a full information-theoretic analysis of any set of random variables if we so chose. In principle, one only need construct F as detailed above and then generate the full set of information atoms $\mathcal{F}$ through unions. In practice—even assuming one already has access to the full joint probability distribution over all variables, a non-trivial assumption to say the least—the growth rate of these sets and the difficulty of mechanistic interpretation once one begins to consider more than three variables has historically stymied these approaches. Moreover, the literature has long debated the semantic meaning of various information atoms—the negativity of interaction information, to pick one example, has been a hotly-debated topic [14].

We sidestep these concerns to a degree by narrowing our focus from a totally arbitrary collection of random variables to a collection of random variables that are measurements of a stochastic process over time. This introduces a significant amount of structure into the informational relationships between the variables, as we will show in the next section.

III. INFORMATION IN STOCHASTIC PROCESSES

As noted at the end of Section II C, we are interested here not in truly arbitrary collections of random variables but rather stochastic processes, which can be understood as a sequence of random variables related to each other through time by a particular dynamic. Specifically, we investigate the relationship between the informational quantities of random variable blocks belonging to the process and to the process dynamic.

A. Discrete Discrete Processes

We take a *stochastic process* $\mathcal{P}$ to consist of a $\mathbb{Z}$-indexed random variable X defined on the measure space $\left(\mathcal{X}^{\mathbb{Z}}, \Sigma, \mu\right)$. This indexing is temporal and is done by the use of subscripts. For example, we write $X_t = x_t$ to say that $x \in \mathcal{X}$ is the specific value of X at time t.

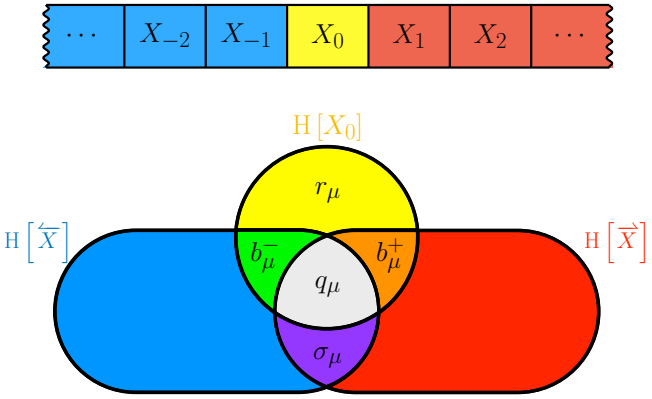

FIG. 2. Above: a tape representing a measurements of a discrete-time stochastic process. Below: information diagram representing the informational relationships between the future, the present X_0, and the past measurements of a in a generic discrete symbol, discrete-time stochastic process. The i-diagram is labeled with the ephemeral information r_μ, the binding informations b_μ, the enigmatic information q_μ, and the elusive information σ_μ.

The dynamic of the stochastic process is given by the *shift operator*, also called the translation operator, which is an operator $\tau : \mathcal{X}^{\mathbb{Z}} \to \mathcal{X}^{\mathbb{Z}}$ that maps t to $t+1$: $\tau x_t = x_{t+1}$. It also acts on the measure: $(\tau\mu)(E) = \mu(\tau^{-1}E)$ for $E \in \Sigma$. This addition extends the measure space to a dynamical system $(\mathcal{X}^{\mathbb{Z}}, \Sigma, \mu, \tau)$.

Blocks of the process, called *words*, are denoted by $X_{a:b} = \{X_t : a < t \leq b; a, b \in \mathbb{Z}\}$ with the left index inclusive and the right exclusive. A word could also refer to a particular realization of a given length. For instance, one might write $X_{0:3} = X_0X_1X_2$ or $x_{0:3} = x_0x_1x_2$.

To simplify our mathematical development, we restrict to stationary, ergodic processes: those for which $\Pr(X_{t:t+\ell}) = \Pr(X_{0:\ell})$ for all $t \in \mathbb{Z}$, $\ell \in \mathbb{Z}^+$, and for which individual realizations obey all of those statistics.

We refer to the observation at $t = 0$ as the *present* X_0. We call the infinite sequence $X_{-\infty:0}$ the *past*, which we also (more frequently) denote with an arrow pointing left: $\overleftarrow{X}$. Accordingly, the infinite sequence $X_{1:\infty}$ is called the *future* and denoted $\overrightarrow{X}$. Note that due to process stationarity, the index denoting the present nominally can be set to any value without altering any subsequent analysis.

Our strategy for developing the information theoretics of stochastic processes primarily is concerned with profiling the relationships between the past, present, and future. Given this, a useful perspective on processes is to picture them as an communication channel transmitting information from the past $\overleftarrow{X} = \ldots X_{-3}X_{-2}X_{-1}$ to the future $\overrightarrow{X} = X_1X_2X_3\ldots$ through the medium of the present X_0. This perspective motivates deviating from three-way symmetry in our i-diagrams of processes, as in Fig. 2. The

Label Type					Partition		Information Atom
Decimal	Lexicographic			Indicial	Joint Dist.	Conditioned	
i	X	Y	Z	k	A_i	$\overline{A_i}$	α_i
1	1	0	0	0	$\{X\}$	$\{Y,Z\}$	$\mathrm{H}[X \mid Y,Z]$
2	0	1	0	1	$\{Y\}$	$\{X,Z\}$	$\mathrm{H}[Y \mid X,Z]$
3	1	1	0	01	$\{X,Y\}$	$\{Z\}$	$\mathrm{I}[X;Y \mid Z]$
4	0	0	1	2	$\{Z\}$	$\{X,Y\}$	$\mathrm{H}[Z \mid X,Y]$
5	1	0	1	02	$\{X,Z\}$	$\{Y\}$	$\mathrm{I}[X;Z \mid Y]$
6	0	1	1	12	$\{Y,Z\}$	$\{X\}$	$\mathrm{I}[Y;Z \mid X]$
7	1	1	1	012	$\{X,Y,Z\}$	$\emptyset$	$\mathrm{I}[X;Y;Z]$

TABLE I. The irreducible information atoms for a set of three random variables $\mathfrak{X} = \{X, Y, Z\}$. Compare the list of α_i to the areas of the information diagram depicted in Fig. 1.

past and the future are depicted here as extending to the left and the right, respectively, to mirror visualizing the bi-infinite chain of random variables.

B. Process Information Atoms

Although one might expect increasing difficulty when moving to a dynamical system, on the surface profiling a process' information atoms in terms of its past $\overleftarrow{X}$, present X_0, and future $\overrightarrow{X}$ requires no more tools than already developed in Section II A. We need only apply Eq. (11):

$$\begin{aligned} F_{\mathcal{P}} = \Big\{ & \mathrm{H}\left[\overleftarrow{X} \mid X_0, \overrightarrow{X}\right], \mathrm{H}\left[X_0 \mid \overleftarrow{X}, \overrightarrow{X}\right], \\ & \mathrm{I}\left[\overleftarrow{X}; X_0 \mid \overrightarrow{X}\right], \mathrm{H}\left[\overrightarrow{X} \mid X_0, \overleftarrow{X}\right], \\ & \mathrm{I}\left[\overleftarrow{X}; \overrightarrow{X} \mid X_0\right], \mathrm{I}\left[X_0; \overrightarrow{X} \mid \overleftarrow{X}\right], \\ & \mathrm{I}\left[\overleftarrow{X}; X_0; \overrightarrow{X}\right] \Big\} . \end{aligned}$$

As there are only three (admittedly aggregate) random variables in play, applying Eq. (11) gives the expected set of seven quantities. The atoms are shown in information diagram form in Fig. 2, alongside an infinite length chain depicting the measurements of the associated process. The shape of the i-diagram has been distorted from the symmetrical one in Fig. 1 to emphasize the empirically known relationships between the variables (i.e., their temporal ordering) but it is worth confirming that each atom in Fig. 2 is identifiable as one of the atoms depicted in Fig. 1.

Five out of the seven atoms in $F_{\mathcal{P}}$ have been named and can be explained intuitively [6]:

1. *Ephemeral* r_μ: The information localized to single measurement of $\mathcal{P}$ at one time and not correlated to its peers:

$$r_\mu = \mathrm{H}\left[X_0 \mid \overleftarrow{X}, \overrightarrow{X}\right] . \tag{12}$$

2. *Binding* b_μ: Two equivalent quantities, *forward binding information* b_μ^+ and *reverse binding information* b_μ^-:

$$\begin{aligned} b_\mu^+ &= \mathrm{I}\left[X_0; \overrightarrow{X} \mid \overleftarrow{X}\right] \text{ and} \\ b_\mu^- &= \mathrm{I}\left[X_0; \overleftarrow{X} \mid \overrightarrow{X}\right] . \end{aligned} \tag{13}$$

   For stationary processes we always have $b_\mu^+ = b_\mu^-$. The forward and reverse binding informations can be interpreted as how correlated any given measurement of a process is with the future and the past, respectively.

3. *Enigmatic* q_μ: Aptly named, this is the interaction information between any given measurement of a process and the infinite past and future:

$$q_\mu = \mathrm{I}\left[X_0; \overleftarrow{X}; \overrightarrow{X}\right] . \tag{14}$$

   As this is a multivariate mutual information, it can be negative.

4. *Elusive* σ_μ: The amount of information shared between the past and future that is not communicated through the present:

$$\sigma_\mu = \mathrm{I}\left[\overleftarrow{X}; \overrightarrow{X} \mid X_0\right] . \tag{15}$$

Note that the μ in these refers to the process measure defined in Section III A and is historical notation.

The Shannon entropy rate h_μ is is not an irreducible information atom. It is given by $h_\mu = \mathrm{H}[X_0 \mid \overleftarrow{X}] = b_\mu^+ + r_\mu$. As long as the process is finitary, which is to say

its *excess entropy* $\boldsymbol{E} = \mathrm{I}[\overleftarrow{X}; \overrightarrow{X}] = b_\mu^+ + q_\mu + \sigma_\mu$ is finite, the atoms above will be finite.

The other two atoms, $\mathrm{H}[\overleftarrow{X} \mid X_0, \overrightarrow{X}]$ and $\mathrm{H}[\overrightarrow{X} \mid X_0, \overleftarrow{X}]$ are typically infinite, although they scale linearly with the length (ℓ) of a window stretching into the future and past:

$$\mathrm{H}\left[\overleftarrow{X}^\ell \mid X_0, X\right] \sim \ell h_\mu, \quad \text{and}$$
$$\mathrm{H}\left[\overrightarrow{X}^\ell \mid X_0, X\right] \sim \ell h_\mu \ .$$

IV. OPTIMAL MODELS OF DISCRETE PROCESSES

Directly working with processes—nominally, infinite sets of infinite sequences and their probabilities—is cumbersome. Practically, we do not want to determine entropies over distributions of infinite pasts and futures. Rather, we wish to build a minimal (finitely-specified) model that captures all correlations in stochastic process $\mathcal{P}$ relevant to the present X_0, allowing access to a process' complete informational profile.

A. Computational Mechanics

The framework of *computational mechanics* [8] provides an exact solution to the problem of optimal minimal predictive modeling in the form of the *ϵ-machine*—a model whose states are the classes defined by an equivalence relation $\overleftarrow{x} \sim \overleftarrow{x}'$ that groups all pasts giving rise to the same prediction. These classes are called the *causal states*.

Definition 1. A process' *causal states* are the members of the range of the function:

$$\epsilon[\overleftarrow{x}] = \Big\{ \overleftarrow{x}' \mid \Pr\left(\overrightarrow{X} = \overrightarrow{x} \,|\, \overleftarrow{X} = \overleftarrow{x}\right) = \Pr\left(\overrightarrow{X} = \overrightarrow{x} \,|\, \overleftarrow{X} = \overleftarrow{x}'\right) \text{ for all } \overleftarrow{x} \in \overleftarrow{X}, \overleftarrow{x}' \in \overleftarrow{X} \Big\}$$

that maps from pasts to sets of pasts: $\epsilon : \overleftarrow{\boldsymbol{X}} \to \boldsymbol{\mathcal{S}}$. The latter is the set of causal states, with corresponding random variable $\mathcal{S}$ and realizations σ.

The causal states partition the space $\overleftarrow{\boldsymbol{X}}$ of all pasts into sets (causal states $\sigma \in \boldsymbol{\mathcal{S}}$) of pasts that are predictively equivalent. The set of causal states $\boldsymbol{\mathcal{S}}$ may be finite, fractal, or continuous, depending on the properties of the underlying process [15]. In the following, we focus on processes with finite causal state sets.

The dynamic over the casual states is inherited from the shift operator τ on the process. State-to-state transitions occur on measurement of a new symbol $X_0 = x$, which is appended to the observed history to give a new history: $\overleftarrow{x} \to \overleftarrow{x}x$. Therefore, the causal state transition is $\epsilon[\overleftarrow{x}] = \sigma_i \to \epsilon[\overleftarrow{x}x] = \sigma_j$ and occurs with probability $\Pr(X_0 = x \mid \mathcal{S}_0 = \sigma_i)$. Note that the subscripts on the realizations σ indicate a specific element of $\boldsymbol{\mathcal{S}}$, while the subscripts on the random variables X and $\mathcal{S}$ indicate time. Section IV B 3 discusses the temporal indexing of causal states in more detail.

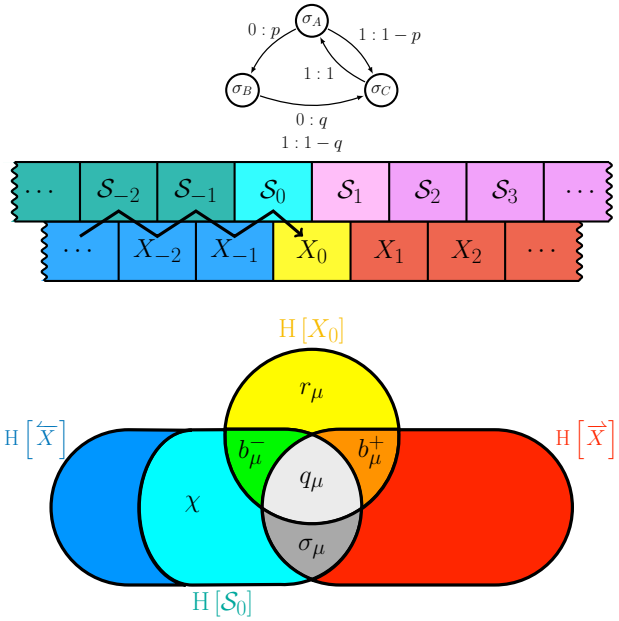

FIG. 3. (Top) A process' ϵ-machine as a state-transition diagram—a stochastic state machine. (Middle) Time indexing of causal states and measurements represented on an bi-infinite chain. The arrow depicts the trajectory (random variable sequence) through time. (Bottom) Process information diagram with the casual state $\mathcal{S}_0$ at time $t = 0$; cf. Fig. 2. The causal state is a function of the infinite past—which is to say its atom $H[\mathcal{S}_0]$ in the i-diagram is contained entirely within the past $H[\overleftarrow{X}]$. The model complexity measure χ is shown alongside the process-defined quantities in Section III B.

The causal state set together with this dynamic is the *ϵ-machine* $M_\epsilon = \{\boldsymbol{\mathcal{S}}, \mathcal{X}, \{T^{(x)} : x \in \mathcal{X}\}\}$, where $T_{ij}^{(x)} = \Pr(\sigma_j, x \mid \sigma_i)$. The ϵ-machine is guaranteed to be optimally predictive because knowledge of what causal state a process is in at any time is equivalent to knowledge of the entire past: $\Pr(\overrightarrow{X} \mid \mathcal{S}) = \Pr(\overrightarrow{X} \mid \overleftarrow{X})$. The dynamic over causal states is Markovian in that they render the past and future statistically independent: $\Pr(\overleftarrow{X}, \overrightarrow{X} \mid \mathcal{S}) = \Pr(\overleftarrow{X} \mid \mathcal{S}) \Pr(\overrightarrow{X} \mid \mathcal{S})$. We call these properties together *causal shielding*. ϵ-Machines also have a property called *unifilarity*, which means that knowledge of the current causal state and the next symbol is sufficient to determine the next state: $\mathrm{H}[\mathcal{S}_{t+1} \mid X_t, \mathcal{S}_t] = 0$.

These properties are visually represented in Fig. 3, where the information $H[\mathcal{S}_0]$ contained in causal state $\mathcal{S}_0$ is entirely encapsulated by the information $H[\overleftarrow{X}]$ in the past $\overleftarrow{X}$. The casual state also must encompass the entirety of

the excess entropy $\boldsymbol{E} = \mathrm{I}\left[\overleftarrow{X}; \overrightarrow{X}\right]$. These two constraints result in an i-diagram that contains strictly fewer atoms than four random variables would maximally allow. In this case, an i-diagram has a maximum of nine random variables. This constraint makes i-diagrams a useful tool to study ϵ-machines beyond the point they would normally become intractable for sets of random variables.

The ϵ-machine is the minimal model in the sense that the amount of information stored by the states is smaller than any other optimal rival model. We quantify this by taking the Shannon entropy over the causal states $C_\mu = \mathrm{H}[\mathcal{S}]$, which we call the *statistical complexity* [8]. The difference between model information and the excess entropy is called the *crypticity* [9]:

$$\chi = C_\mu - \boldsymbol{E} \ .$$

χ is an additional measure of model complexity that quantifies how much internal-state information is not directly available through measurement sequences.

B. Directional Computational Mechanics

While computational mechanics is built under the assumption of optimizing over prediction, it can also be applied to the goal of *retrodiction*—finding a distribution over pasts given knowledge of the future. We can think of this, equivalently, as predicting the *reverse process*—the process in a world where time runs in the opposite direction.

1. Reverse ϵ-Machine

Informationally speaking, the time-reversal of a stationary process is not particularly interesting. As noted in Section III, the forward and reverse binding informations b_μ are equal, and the excess entropy $\boldsymbol{E}$, the ephemeral information r_μ, the enigmatic information q_μ, and the elusive information q_μ are all time symmetric by definition.

However, it is not generally the case that the predictive causal states are the same as the retrodictive ones. And so, for a full analysis of a process' informational structure we must consider the directional casual states. This is their construction is straightforward but requires new notation. We rename the objects defined in Definition 1 to the *forward causal states* $\sigma^+ \in \boldsymbol{\mathcal{S}}^+$ and denote the equivalence function as $\epsilon^+[\overleftarrow{x}]$. Similarly, the associated ϵ-machine will now be called the *forward ϵ-machine* and be denoted M_ϵ^+. The definitions do not change. In contrast, we have:

Definition 2. A process' *reverse causal states* are the members of the range of the function:

$$\begin{aligned}\epsilon^-[\overrightarrow{x}] = \Big\{ \overrightarrow{x}' \,|\, &\Pr\left(\overleftarrow{X} = \overleftarrow{x} \,|\, \overrightarrow{X} = \overrightarrow{x}\right) \\ &= \Pr\left(\overleftarrow{X} = \overleftarrow{x} \,|\, \overrightarrow{X} = \overrightarrow{x}'\right) \\ &\quad \text{for all } \ \overrightarrow{x} \in \overrightarrow{X}, \overrightarrow{x}' \in \overrightarrow{X} \Big\}\end{aligned}$$

that maps from futures to sets of futures. The set of reverse causal states is denoted $\boldsymbol{\mathcal{S}}^-$, with corresponding random variable $\mathcal{S}^-$ and realizations σ^-.

The *reverse ϵ-machine* M_ϵ^- is defined in the expected way, running the shift operator τ in reverse time. It is worth noting that the reverse ϵ-machine is not guaranteed to be finite when the forward ϵ-machine is finite, and vice versa. However, the following will consider processes for which both machines are finite.

As noted above, the statistical complexity C_μ typically differs in the forward and reverse directions. Accordingly, we also have directional crypticities with more concise expressions than those given above:

$$\chi^+ = \mathrm{H}\left[\mathcal{S}_t^+ \mid \mathcal{S}_t^-\right] \text{ and} \tag{16}$$

$$\chi^- = \mathrm{H}\left[\mathcal{S}_t^- \mid \mathcal{S}_t^+\right] \ . \tag{17}$$

The crypticities χ^+ and χ^- have compelling interpretations. χ^+ is the amount of information in the forward ϵ-machine that is not contained in the excess entropy—which, recall, is the total amount of information the process communicates through time. It may seem odd that the causal states could contain more information than this, but consider the classic example of a "nearly"-IID process. Such a process looks arbitrarily close to random, and so the amount of information communicated through time is vanishingly small. However, in fact, there exist very long-range correlations that can marginally improve on optimal prediction, which must therefore be stored in the causal states. Indeed, it is not only possible, but even typical for processes generated by hidden Markov models for the excess entropy to be finite while the statistical complexity and therefore the crypticity, diverge [15].

2. Bidirectional Machine

With both the forward ϵ-machine and the reverse ϵ-machine in hand, we can consider the *bidirectional machine* $M_\epsilon^\pm$, which simultaneously optimally predicts and retrodicts.

Definition 3. The *bidirectional causal states* of a process are the members of the range of the function:

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

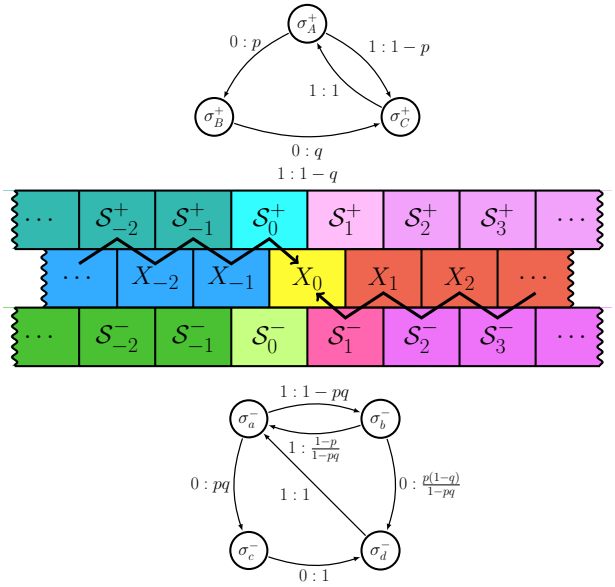

FIG. 4. The forward (Top) and reverse (Bottom) ϵ-machines of a stochastic process. The time indexing of the causal states and the emitted measurements are laid out on three parallel horizontal chains. The variables on the chain are color coded to match Fig. 5, which depicts the accompanying information diagram. The arrows depict the path through time in the forward (Top) and reverse (Bottom) directions, respectively, cf. Fig. 3.

that maps histories to set of histories. The set of bidirectional causal states is denoted $\boldsymbol{\mathcal{S}}^{\pm}$, with corresponding random variable $\mathcal{S}^{\pm}$ and realizations $\sigma^{\pm}$.

The bidirectional causal states are a subset of the Cartesian product of forward and reverse casual states: $\boldsymbol{\mathcal{S}}^{\pm} \subseteq \boldsymbol{\mathcal{S}}^{+} \times \boldsymbol{\mathcal{S}}^{-}$. Our convention in the following is to label causal states with Latin letters, using upper case for the forward direction and lower case for the reverse direction: i.e., $\boldsymbol{\mathcal{S}}^{+} = \{BC\}$ and $\boldsymbol{\mathcal{S}}^{-} = \{a, b, c, d\}$ as in Fig. 4. The bidirectional states are labeled by their corresponding forward and reverse states: i.e., $\boldsymbol{\mathcal{S}}^{\pm} = \{Aa, Ba, \dots\}$. See Fig. 6 for examples.

We primarily use the bidirectional machine in the algorithm that calculates our new informational properties, as discussed in Section VII.

3. *Temporal Indexing of Causal States*

For process random variables—e.g., X and $\mathcal{S}$—subscripts indicate the variable's time index. While subscripts on realizations indicate their set index. In other words, $\mathcal{S}_t^+ = \sigma_i$ refers to the forward causal state at time t, which takes on the value of the ith element of $\boldsymbol{\mathcal{S}}^{+}$. The ϵ-machine is drawn as a state-transition diagram with transition probabilities $\Pr\left(X_t = x \mid \mathcal{S}_t, \mathcal{S}_{t+1}\right)$ from $\mathcal{S}_t$ to $\mathcal{S}_{t+1}$ written as $x : \Pr$ ($x : \Pr$: direction for the bidirectional machine) on the appropriate transitions.

Figure 4 depicts the forward ϵ-machine (Top) and the reverse ϵ-machine (Bottom) of a given process. The time-indexed states of the ϵ-machines are depicted on state chains $\dots \mathcal{S}_1 \mathcal{S}_2 \dots$ sandwiching the chain of process measurements $\dots X_1 X_2 \dots$. Although we index the causal states with integers, we imagine them as occurring on "half time steps" in between the measurement time indices. The arrows trace the path through time along the causal states and observed measurements. Note that in the forward direction, the causal state at time t emits the measurement at time t, but in the reverse direction the causal state at time t is said to emit the measurement at time $t-1$. This offset is a consequence of using integer indices for the states. The mismatch in the reverse time direction (rather than the forward direction) is a matter of convention.

Consider how there are four states that symmetrically "surround" each measurement. For the present X_0, these states are $\mathcal{S}_0^+$, $\mathcal{S}_0^-$, $\mathcal{S}_1^+$, and $\mathcal{S}_1^-$. The informational relationship the forward and reverse states have with the measurement they surround is asymmetrical. We might say that two of the states—$\mathcal{S}_1^+$ and $\mathcal{S}_0^-$—have already "seen" the measurement X_0, as it was emitted on the transition *to* that state. From the perspective of these states, X_0 is included in the past or future, respectively. We say that $\mathcal{S}_0^-$ and $\mathcal{S}_1^+$ are "interior" to the measurement, drawing on the visual depiction in the i-diagram in Fig. 5, where these states (kidney bean in shape) are positioned as closer to the center of the diagram. The other states $\mathcal{S}_0^+$ and $\mathcal{S}_1^-$ are then "exterior"—they trail on either end of the i-diagram due to their access to information furthest in the past or future, respectively.

V. ATOMIC TAXONOMY

With the causal states in place, we can develop a full information-theoretic analysis of prediction—that is, prediction and retrodiction.

A. Information Atoms from Causal States

Naively, our new information atom set is formed by simply adding the four causal states "surrounding" the present measurement to our random variable set:

$$\mathfrak{X}_\epsilon = \left\{ \overleftarrow{X}, \mathcal{S}_0^+, \mathcal{S}_0^-, X_0, \mathcal{S}_1^+, \mathcal{S}_1^-, \overrightarrow{X} \right\} .$$

However, thanks to causal shielding, we can drop the infinite past and future, as they are redundant with the causal states.

1. *Casual Shielding*

Our secondary purpose of introducing computational mechanics was to reap the benefits of the causal shielding

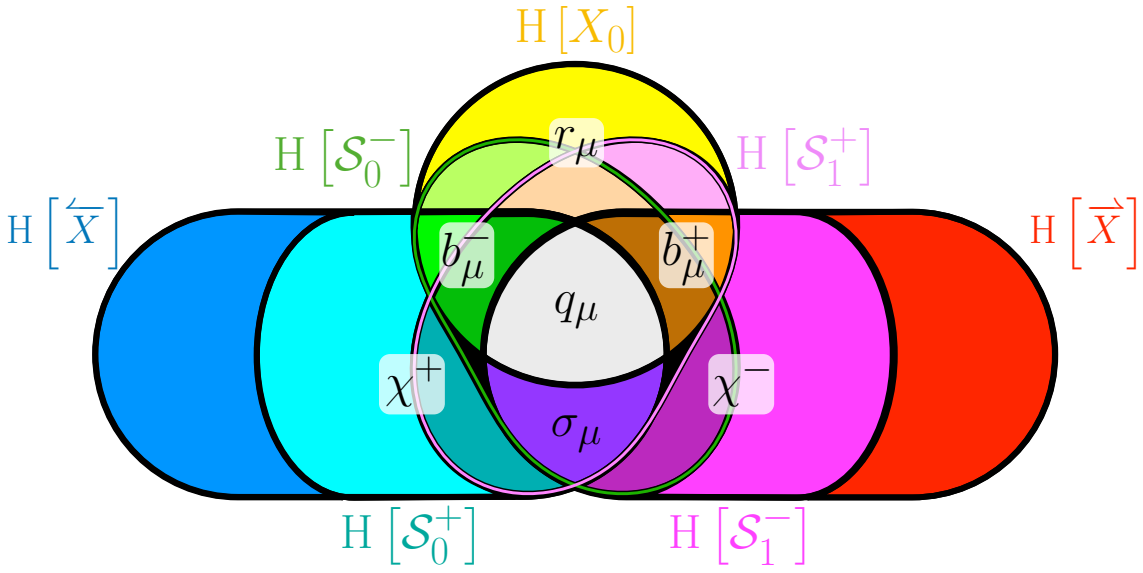

FIG. 5. Information diagram representing all possible positive atoms of a single transition of a bidirectional machine, including the states at $t = 0$ and the states at $t = 1$. The majority of the information atoms theoretically possible go to zero due to the causal shielding of the causal states. The information atoms defined in Section III B, five of which are no longer irreducible, are overlaid over their corresponding atoms.

properties of the causal states. By using the ϵ-machines as our model, we can study all temporal correlations that impact the present without including the infinite-length past and future in our informational analysis, as we did in Section III B. In practice, this means replacing the conditioning variables according to shielding order.

Given a set of variables $\mathfrak{X}$ and subsets $A, B \in \mathfrak{X}$ we say that A *shields* the set $\mathfrak{X}$ from B if:

$$(C \perp B \mid A) \ , \ \text{for all } C \in (\mathfrak{X} \setminus B) \ .$$

For example, $\mathcal{S}_0^+$ shields $\mathfrak{X}_\epsilon$ from the past $\overleftarrow{X}$ and $\mathcal{S}_1^-$ shields $\mathfrak{X}_\epsilon$ from the future $\overrightarrow{X}$. This means that wherever $\overleftarrow{X}$ appears in an informational quantity, it is possible to replace the variable with $\mathcal{S}_0^+$ with no loss of information and to replace $\overrightarrow{X}$ with $\mathcal{S}_1^-$. So, in fact, our relevant random variable set is:

$$\mathfrak{X}_\epsilon = \{\mathcal{S}_0^+, \mathcal{S}_0^-, X_0, \mathcal{S}_1^+, \mathcal{S}_1^-\} \ .$$

Five random variables maximally produces an irreducible atom set of $2^5 = 32$ atoms, but $F_{M_\epsilon^\pm}$ consists of only fourteen nonzero irreducible atoms. This reduction is due to the particular properties of the causal states—namely unifilarity and causal shielding. The structured nature of Fig. 5 indicates the influence of these properties, which we discuss in further depth in Section V B. First, to get there we introduce the nonzero information atoms of an optimally modeled process.

2. *Anatomy of a Bit Revisited*

Ten of our new information atoms are related to the original five atoms given in Section III B. First, rewrite those atoms in terms of the causal states, replacing infinite futures and pasts with the appropriate shielding causal states:

- $r_\mu = \mathrm{H}\left[X_0 \mid \mathcal{S}_0^+, \mathcal{S}_1^-\right]$,
- $b_\mu^+ = \mathrm{I}\left[X_0; \mathcal{S}_1^- \mid \mathcal{S}_0^+\right]$,
- $b_\mu^- = \mathrm{I}\left[X_0; \mathcal{S}_0^+ \mid \mathcal{S}_1^-\right]$,
- $q_\mu = \mathrm{I}\left[\mathcal{S}_0^+; X_0; \mathcal{S}_1^-\right]$, and
- $\sigma_\mu = \mathrm{I}\left[\mathcal{S}_0^+; \mathcal{S}_1^- \mid X_0\right]$.

The increase in number of atoms from five to ten is due to the "splitting" of the binding informations b_μ and the ephemeral information r_μ into transient and persistent pieces. By *transient information* we mean information that will be "forgotten" by the the ϵ-machines within a single time step, either into the future (for the forward ϵ-machine) or into the past (for the reverse ϵ-machine). By *persistent information* we mean information that is "stored" in the model, and remains correlated with new causal states. Figure 5 depicts this by overlaying the taxonomy of a process' informational quantities over their new constituent atoms. The persistent informations are colored darker in shade. The full list of atoms is given by Table II, organized by their parent "anatomy of a bit" quantity.

Consider first the reverse binding information b_μ^-. This splits into two terms:

$$b_\mu^- = \underbrace{\mathrm{I}\left[X_0; \mathcal{S}_0^+; \mathcal{S}_0^- \mid \mathcal{S}_1^+\right]}_{\text{transient}} + \underbrace{\mathrm{I}\left[X_0; \mathcal{S}_0^+; \mathcal{S}_0^-; \mathcal{S}_1^+ \mid \mathcal{S}_1^-\right]}_{\text{persistent}} \ .$$

The first term is *transient binding information* in the forward causal state at $t = 0$ that is not carried through to the forward causal state at $t = 1$. The second term is called *persistent* as it is that part of the binding information correlated with $\mathcal{S}_1^+$. It therefore influences the future states of the forward ϵ-machine.

We can do the same analysis with the forward binding information and the reverse causal states, recalling that the reverse ϵ-machine runs in reverse time:

$$b_\mu^+ = \underbrace{\mathrm{I}\left[X_0; \mathcal{S}_1^-; \mathcal{S}_1^+ \mid \mathcal{S}_0^-\right]}_{\text{transient}} + \underbrace{\mathrm{I}\left[X_0; \mathcal{S}^-; \mathcal{S}_0^-; \mathcal{S}_1^+ \mid \mathcal{S}_0^+\right]}_{\text{persistent}} \ .$$

The second term is persistent reverse binding information correlated with $\mathcal{S}_0^-$ and it, therefore, influences *past* states of the reverse ϵ-machine.

Analogously, the ephemeral information splits into four terms:

$$\begin{aligned} r_\mu = & \overbrace{\mathrm{H}\left[X_0 \mid \mathcal{S}_1^+, \mathcal{S}_0^-\right]}^{\text{transient}} \\ & + \mathrm{I}\left[X_0; \mathcal{S}_0^- \mid \mathcal{S}_1^+\right] + \mathrm{I}\left[X_0; \mathcal{S}_1^+ \mid \mathcal{S}_0^-\right] \\ & \underbrace{+ \ \mathrm{I}\left[X_0; \mathcal{S}_1^+; \mathcal{S}_0^- \mid \mathcal{S}_0^+, \mathcal{S}_1^-\right]}_{\text{persistent}} \ . \end{aligned}$$

It helps to compare the terms above to the atoms of Fig. 5. The first term is the transient ephemeral information, which is truly ephemeral in that it remains uncorrelated with any causal state at any time. The remaining three are all persistent: the second term is ephemeral information that is correlated with only the states of the reverse ϵ-machine, the third term only with states of the forward ϵ-machine, and the fourth information term is correlated with both.

3. Splitting Causal State Information

The prior accounted for the ten information atoms corresponding to process measurements. There are still four purely causal model information atoms, two of which are new to this analysis. Recall the forward and reverse crypticities Eq. (17). For our system, we have:

$$\begin{aligned} \chi^+ &= \mathrm{H}\left[\mathcal{S}_0^+ \mid \mathcal{S}_0^-\right] \\ \chi^- &= \mathrm{H}\left[\mathcal{S}_1^- \mid \mathcal{S}_1^+\right] \ . \end{aligned}$$

As already noted by Section IV B, the crypticities are a type of modeling information—the amount of information required for the causal states to do optimal prediction or retrodiction above and beyond the excess entropy. As with the binding and ephemeral informations, some of this information is transient and some persistent.

Consider the forward crypticity:

$$\chi^+ = \underbrace{\mathrm{H}\left[\mathcal{S}_0^+ \mid \mathcal{S}_1^+, \mathcal{S}_0^-\right]}_{\text{transient}} + \underbrace{\mathrm{I}\left[\mathcal{S}_0^+; \mathcal{S}_1^+ \mid \mathcal{S}_0^-\right]}_{\text{persistent}} \ .$$

The first term is the transient forward crypticity. This is modeling information that is "forgotten" after one time step—necessary for optimal prediction of X_0 but not of X_1. The second term is the persistent forward crypticity, which is correlated with $\mathcal{S}_1^+$ and continues to be influential in prediction of future observations.

The reverse crypticity splits in the same manner:

$$\chi^- = \underbrace{\mathrm{H}\left[\mathcal{S}_1^- \mid \mathcal{S}_1^+, \mathcal{S}_0^-\right]}_{\text{transient}} + \underbrace{\mathrm{I}\left[\mathcal{S}_1^-; \mathcal{S}_0^- \mid \mathcal{S}_1^+\right]}_{\text{persistent}} \ .$$

Again, the first term is transient and the second is persistent, although in this direction the difference is whether the information is correlated with the reverse causal state $\mathcal{S}_0^-$.

B. Atomic Indicial Structure

As already noted, our informational taxonomy of a prediction results in only fourteen atoms despite a theoretically-possible set of thirty two. This reduction is a result of the structural properties of the causal states. These properties are concisely described using the indicial labeling described in Section II D. Our convention is to order sequences of causal states and measurements starting with a forward-time causal state and continuing in the order: $\mathcal{S}_t^+, \mathcal{S}_t^-, X_t, \mathcal{S}_{t+1}^+, \mathcal{S}_{t+1}^-, X_{t+1}, \ldots$.

This means that in the indicial notation, we have:

$$\begin{aligned} \mathcal{S}_t^+ &\rightarrow k = t \\ \mathcal{S}_t^- &\rightarrow k = t + 1 \\ X_t &\rightarrow k = t + 2 \ . \end{aligned}$$

Using the shorthand notation $\mathrm{H}[k] = \mathrm{H}\left[\mathcal{S}_t^+\right]$, we can then express the structural properties in terms of patterns in the indexes of the random variables, as follows:

1. *Unifilarity*: Given a measurement and the causal state that emitted it, there is no longer any uncertainty in the next state. In the forward and reverse directions, for $k \in \mathbb{N}, k \bmod 3 = 0$, the disallowed atoms are given by:

$$\begin{aligned} \mathrm{H}\left[k+3; \ldots \mid k, k+2, \ldots\right] = 0 \text{ and} \\ \mathrm{H}\left[k+1; \ldots \mid k+2, k+4, \ldots\right] = 0 \ , \end{aligned}$$

where the dots indicate that the remaining two variables may be added to either side of the partition.

Label Type							Partition		Information Atom	
Decimal	Lexicographic					Indicial	Joint Set	Conditioned	Atom	Type
i	$\mathcal{S}_0^+$	$\mathcal{S}_0^-$	X_0	$\mathcal{S}_1^+$	$\mathcal{S}_1^-$	k	A_i	$\overline{A_i}$	α_i	
1	1	0	0	0	0	0	$\{\mathcal{S}_0^+\}$	$\{\mathcal{S}_0^-, X_0, \mathcal{S}_1^+, \mathcal{S}_1^-\}$	$\mathrm{H}\left[\mathcal{S}_0^+ \mid \mathcal{S}_0^-, \mathcal{S}_1^+\right]$	t. χ^+
9	1	0	0	1	0	03	$\{\mathcal{S}_0^+, \mathcal{S}_1^+\}$	$\{\mathcal{S}_0^-, X_0, \mathcal{S}_1^-\}$	$\mathrm{I}\left[\mathcal{S}_0^+; \mathcal{S}_1^+ \mid \mathcal{S}_0^-\right]$	p. χ^+
7	1	1	1	0	0	012	$\{\mathcal{S}_0^+, \mathcal{S}_0^-, X_0\}$	$\{\mathcal{S}_1^+, \mathcal{S}_1^-\}$	$\mathrm{I}\left[\mathcal{S}_0^+; \mathcal{S}_0^-; X_0 \mid \mathcal{S}_1^+\right]$	t. b_μ^-
15	1	1	1	1	0	0123	$\{\mathcal{S}_0^+, \mathcal{S}_0^-, X_0, \mathcal{S}_1^+\}$	$\{\mathcal{S}_1^-\}$	$\mathrm{I}\left[\mathcal{S}_0^+; \mathcal{S}_0^-; X_0; \mathcal{S}_1^+ \mid \mathcal{S}_1^-\right]$	p. b_μ^-
4	0	0	1	0	0	2	$\{X_0\}$	$\{\mathcal{S}_0^+, \mathcal{S}_0^-, \mathcal{S}_1^+, \mathcal{S}_1^-\}$	$\mathrm{H}\left[X_0 \mid \mathcal{S}_0^-, \mathcal{S}_1^+\right]$	t. r_μ
6	0	1	1	0	0	12	$\{\mathcal{S}_0^-, X_0\}$	$\{\mathcal{S}_0^+, \mathcal{S}_1^+, \mathcal{S}_1^-\}$	$\mathrm{I}\left[\mathcal{S}_0^-; X_0 \mid \mathcal{S}_0^+, \mathcal{S}_1^+\right]$	p. r_μ^-
12	0	0	1	1	0	23	$\{X_0, \mathcal{S}_1^+\}$	$\{\mathcal{S}_0^+, \mathcal{S}_0^-, \mathcal{S}_1^-\}$	$\mathrm{I}\left[X_0; \mathcal{S}_1^+ \mid \mathcal{S}_0^-, \mathcal{S}_1^-\right]$	p. r_μ^+
14	0	1	1	1	0	123	$\{\mathcal{S}_0^-, X_0, \mathcal{S}_1^+\}$	$\{\mathcal{S}_0^+, \mathcal{S}_1^-\}$	$\mathrm{I}\left[\mathcal{S}_0^-; X_0; \mathcal{S}_1^+ \mid \mathcal{S}_0^+, \mathcal{S}_1^-\right]$	p. $r_\mu^\pm$
28	0	0	1	1	1	234	$\{X_0, \mathcal{S}_1^+, \mathcal{S}_1^-\}$	$\{\mathcal{S}_0^+, \mathcal{S}_0^-\}$	$\mathrm{I}\left[X_0; \mathcal{S}_1^+; \mathcal{S}_1^- \mid \mathcal{S}_0^-\right]$	t. b_μ^+
30	0	1	1	1	1	1234	$\{\mathcal{S}_0^-, X_0, \mathcal{S}_1^+, \mathcal{S}_1^-\}$	$\{\mathcal{S}_0^+\}$	$\mathrm{I}\left[\mathcal{S}_0^-; X_0; \mathcal{S}_1^+; \mathcal{S}_1^- \mid \mathcal{S}_0^+\right]$	p. b_μ^+
16	0	0	0	0	1	4	$\{\mathcal{S}_1^-\}$	$\{\mathcal{S}_0^+, \mathcal{S}_0^-, X_0, \mathcal{S}_1^+\}$	$\mathrm{H}\left[\mathcal{S}_1^- \mid \mathcal{S}_0^-, \mathcal{S}_1^+\right]$	t. χ^-
18	0	1	0	0	1	14	$\{\mathcal{S}_0^-, \mathcal{S}_1^-\}$	$\{\mathcal{S}_0^+, X_0, \mathcal{S}_1^+\}$	$\mathrm{I}\left[\mathcal{S}_0^-; \mathcal{S}_1^- \mid \mathcal{S}_1^+\right]$	p. χ^-
27	1	1	0	1	1	0134	$\{\mathcal{S}_0^+, \mathcal{S}_0^-, \mathcal{S}_1^+, \mathcal{S}_1^-\}$	$\{X_0\}$	$\mathrm{I}\left[\mathcal{S}_0^+; \mathcal{S}_0^-; \mathcal{S}_1^+; \mathcal{S}_1^- \mid X_0\right]$	σ_μ
31	1	1	1	1	1	01234	$\{\mathcal{S}_0^+, \mathcal{S}_0^-, X_0, \mathcal{S}_1^+, \mathcal{S}_1^-\}$	$\emptyset$	$\mathrm{I}\left[\mathcal{S}_0^+; \mathcal{S}_0^-; X_0; \mathcal{S}_1^+; \mathcal{S}_1^-\right]$	q_μ

TABLE II. For a given process, the irreducible, nonzero information atoms for a set of five random variables $\mathfrak{X} = \{\mathcal{S}_0^+; \mathcal{S}_0^-; X_0; \mathcal{S}_1^+; \mathcal{S}_1^-\}$. The decimal, lexicographic, and indicial labels are given in the left side columns, as laid out in Section II D. The partitioning of the variables is given in the middle two columns, with variables in the left side in the joint distribution and variables on the right side in the conditioning distribution. On the far right, the corresponding information atom is written explicitly (with redundant conditioning variables dropped) alongside the "type" of atom in the taxonomic scheme given in Section III B and whether it is transient (t.) or persistent (p.).

For our analysis of the present, this zeroes out four atoms in each direction. One of these atoms is shared, and so there are seven atoms eliminated in total.

2. *Minimal optimal prediction*: the forward-time causal states are strict functions of the past. They contain no extra information about the future that is not contained within the past, but as optimal predictors they capture *all* of this information, i.e., all of the excess entropy. In information-theoretic terms this means, when conditioning on the future, the forward causal states cannot share information with any other variables except other forward causal states. The same holds in the reverse-time case. For $k, j \in \mathbb{N}, k \mod 3 = 0$, the disallowed atoms are given by:

(i) For $j \mod 3 \neq 0$ and $j > k$:
$$\mathrm{I}\left[k; j; \ldots \mid k+1, \ldots\right] = 0$$
(ii) For $j \mod 3 \neq 1$ and $j < k+4$:
$$\mathrm{I}\left[k+4; j; \ldots \mid k+3, \ldots\right] = 0$$

This accounts for six variables in each direction. However, two atoms are the same in each direction so there are ten atoms eliminated total.

3. *Markov shielding*: This property does not eliminate any atoms when considering only a single time step, but it is worth noting. Since the causal states are Markov order-1, no information may be shared between measurements that is not also contained within the states. For $k \in \mathbb{N}, k \mod 3 = 0$;

$$\mathrm{I}\left[k+2; k+5; \ldots \mid k+3, k+4, \ldots\right] = 0 \ .$$

As a final note on indicial ordering, consider the sixth column in Table II, which lists the informational quantities discussed. Comparing to the fourth and fifth columns, which give the partitioning of $\mathfrak{X}_\epsilon$, it is clear that we are able to write the informational quantities without necessarily including all variables in the conditioning set. (This is sometimes also true for the joint distribution, but we take it as a convention to always explicitly include all variables in the joint distribution.)

We are able to do this because our second property, minimal optimal prediction, is equivalent to saying that the forward (reverse) causal states render future (prior) variables conditionally independent with respect to all prior (future) measurements and prior forward (future reverse) causal states. Figure 5 depicts this property as the forward time causal states covering all space shared between future variables and the prior measurements and prior forward causal states.

When writing conditional informational quantities, our convention is to drop all forward causal states "eclipsed" by forward causal states further along in the future and all reverse causal states "eclipsed" by reverse casual states further in the past. We also drop measurements "eclipsed" by causal states in either direction. To see the result of this, compare the $\overline{A_i}$ column in Table II to the conditioning variables in the information quantities listed in the α_i column.

VI. PROCESSES

With our new information quantities established, we now consider a suite of exactly-solvable taxonomies for example binary discrete stochastic processes.

A. Independent, Identically-Distributed

The first is the simplest possible: an infinite sequence of independent, identically-distributed (IID) coin flips. The ϵ-machines for such a process with a bias of p are given in the first three rows Fig. 6 (a). In this case, since the process has no structure or memory, there is only a single causal state in each direction, which can be identified as equivalent in the bidirectional machine. Accordingly, all transitions are self-loops.

With only a single state, the statistical complexity (causal state or model information) C_μ vanishes, zeroing out all information in a single measurement except the transient ephemeral information r_μ. If $p = 0.5$, $r_\mu = 1$ bit, as shown in the bar chart in the bottom row of Fig. 6 (a). The information in the infinite past and future diverges.

B. Periodic

The second example process is nearly as trivial. An n-periodic process requires exactly n causal states but has only deterministic transitions. As such, knowledge of the current measurement is equivalent to knowledge of the infinite past and infinite future, as well as the forward and reverse causal states. Intuitively, we understand then that the only remaining positive quantity is q_μ. This is the information shared between all model variables. For an n-periodic process, $q_\mu = \log_2 n$ bit.

The $n = 3$ case is depicted in Fig. 6 (b). Notice that each of the forward causal states can be uniquely identified with a reverse causal state and vice versa. We call this a *noncryptic* process in both directions. In the bar chart of information quantities in the bottom row of Fig. 6 (c), $n = 3$ and so $r_\mu = 1.58$.

C. Even

The Even process is a binary process of sequences of 0s of any length interspersed with even-length sequences of 1s. The probability distribution of the length of the sequences of 0s and 1s are controlled by a single parameter $p \in (0, 1)$. The forward direction ϵ-machine is depicted in Fig. 6 (c). There are two forward-time casual states $\{A, B\}$. The self-loop on state A occurs with probability p when the machine is in state A. The topology of the reverse-time ϵ-machine is identical to the topology of the forward-time machine and, in fact, the Even process is also noncryptic in both directions.

Despite the Even process' apparent simplicity, its prediction taxonomy is surprisingly complex. The process is infinite-order Markov, which is to say that the probability of the next symbol depends on the infinite length past and cannot be exactly extrapolated from any finite-length history. As such, there is no finite Markov model that generates the Even process—it can only be finitely modeled with a hidden Markov model.

As noted, both forward and reverse crypticities vanish. The ephemeral informations vanish for similar reasons: since knowledge of the forward states is equivalent to knowledge of the reverse states, for there to be any positive ephemeral information the model would need to have two possible transitions between bidirectional causal states on a single time step, which it does not.

We are left with two transient binding informations, the enigmatic information, and the elusive information. The entropy rate of the Even process is produced entirely by the choice between the self-loop and the transition on state Aa. This, exactly, is the transient portion of the forward binding information: it is not determined by knowledge of the previous state $\mathcal{S}_0^-$ (which again, in this case is equivalent to $\mathcal{S}_0^+$). The reverse argument explains the reverse binding information.

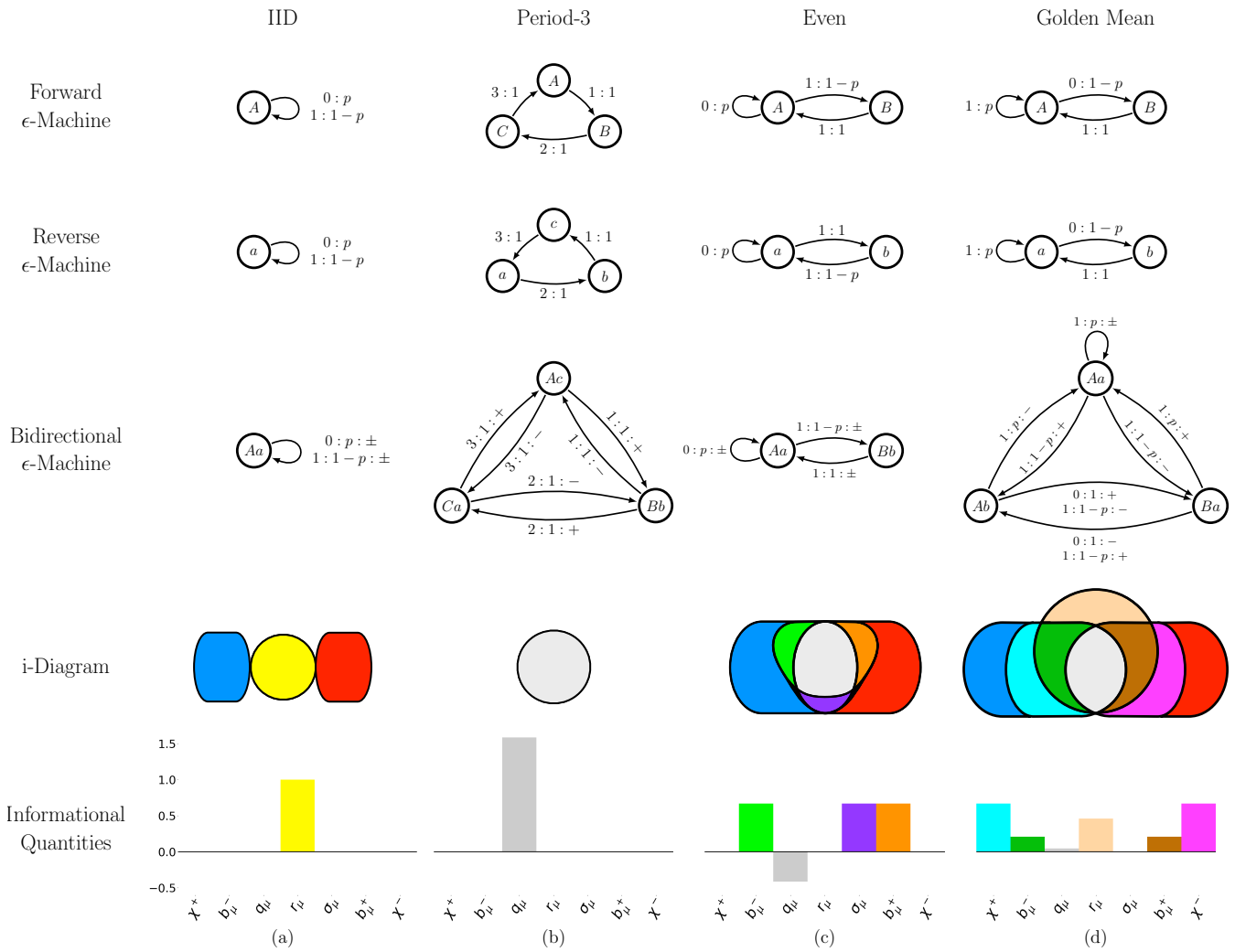

FIG. 6. Example prediction taxonomies: The forward ϵ-machines (top row), reverse ϵ-machines (second row), bidirectional ϵ-machines (third row), i-diagrams (fourth row), and exact informational quantities plotted in a bar chart (bottom row) of four discrete stochastic process: a) an identically independently distributed (IID) sequence of coin flips, b) an n-periodic process, c) the Even process, d) the Golden Mean process. Recall that our convention is to use uppercase Latin letters for the forward causal states and lower case Latin letters for the reverse causal states.

The elusive information σ_μ is positive since the bidirectional causal state is not uniquely determined by the current measurement—a 1 may indicate the machine has just transitioned to Aa or Bb. Finally, we have the enigmatic information q_μ, which is negative. In this case the elusive information is the multivariate mutual information between $\mathcal{S}_0^- = \mathcal{S}_0^+$, X_0, and $\mathcal{S}_1^- = \mathcal{S}_1^+$. Recall that the negativity of multivariate mutual information means that the addition of the third variable (which can be taken to be any of the three, due to symmetry) *increases* the shared information between the other two.

How to understand this? Notice that the Even state machine ties one symbol to self-loops (0) and one symbol to transitions (1). This means that knowledge of the measurement reveals that the temporal ordering of the states is also a structural relationship, increasing the shared information.

D. Golden Mean

Finally, consider a binary stochastic process that is superficially quite similar to the Even process, but has a very different informational structure. The Golden Mean Process is a binary process that can have sequences of 1's of any length, interspersed with only single 0's. The probability of a 1s sequence decreases as the length increases, although the nature of this probability distribution depends on a single parameter $p \in (0, 1)$. The ϵ-machines of this process family is given in Fig. 6 (d). There are two forward-time causal states and p determines the probability split between the self-loop and the state transition on state A, controlling the probability of seeing a 0 after a sequence of 1s.

In contrast to the the Even process, the bidirectional

machine given in Fig. 6 shows that the forward and reverse causal states are not one and the same nor are they independent—there are three bidirectional causal states. The "missing" bidirectional state is Bb, which would represent the forward machine being in state B and the reverse machine being in state b simultaneously. This is impossible as it implies a sequence of two 0s.

Unlike the Even process, the bidirectional machine is *cryptic*: even if one knows the current causal state in one direction, it is possible to be uncertain of the causal state in the opposite direction. The elusive information σ_μ vanishes because the causal state can always be determined by a measurement of the present (1s lead to either A or a, 0s lead to B or b, depending on scan direction).

All other types of information are represented. The entropy rate splits into ephemeral information and forward binding information. We can intuitively think of this split as the information produced by the process each time step—a splitting into a piece that does not explain the future (ephemeral) and a piece that does (forward binding). The ephemeral information appears only in its least ephemeral form, in a sense—we are only uncertain about the observed symbol if we are also uncertain of the previous reverse causal state $\mathcal{S}_0^-$ and the next forward causal state $\mathcal{S}_1^+$. This uncertainty occurs when the machine is in state A, which could transition from $Aa \to Aa$ on a 1 or from $Ab \to Ba$ on a 0. That is only two of the three possible transitions out of state A, however. The machine can also transition from $Aa \to Ab$ on a 1. This transition is informative about the future, in that it determines the value of $\mathcal{S}_1^-$ and so contributes to the forward binding information. As usual, this logic also applies in reverse to the reverse binding information.

Finally, we have the enigmatic information, which is positive for all values of p. To understand this, we recall our discussion of negative enigmatic information in the previous example (Section VI C). There the value of the present symbol improved our ability to guess what kind of transition the machine was undergoing. In this case, the opposite intuition holds.

VII. BIDIRECTIONAL ATOM ALGORITHMS

ϵ-Machine informational properties are useful not only in that they define a suite of interpretable informational quantities, but also because knowledge of the ϵ-machine allows directly and exactly calculating those quantities [16]. With knowledge of a finitely-specified forward ϵ-machine of a discrete stochastic process (which can even be inferred from time series data [17]), we can find the reverse and bidirectional ϵ-machines and from there calculate all the quantities defined in Section V.

Before describing relevant algorithms, we recall and define a few preliminary concepts.

A machine M is given by a list of square *transition matrices* $\left\{T^{(x)} : x \in \mathcal{X}\right\}$ where $T^{(x)}_{ij} = \Pr\left(\sigma_j, x \mid \sigma_i\right)$. Let $N = |\boldsymbol{\mathcal{S}}^+|$ and $M = |\boldsymbol{\mathcal{S}}^-|$ so that the transition matrices of the forward ϵ-machine are $N \times N$ and the transition matrices of the reverse ϵ-machine are $M \times M$.

The *mixed-state algorithm*, fully elucidated in Ref. [15], finds the mixed states η of a hidden Markov model M. Briefly, for a length-ℓ word w generated by M the mixed state $\eta(w)$ is an observer's best guess as to which state the machine is in after observing w:

$$\eta(w) = \left[\Pr\left(\mathcal{S}_i \mid X_{0:\ell} = w\right)\right] \tag{18}$$

given an initial guess of π—the asymptotic stationary distribution of the machine: $\pi = \pi T$, where the state transition matrix is $T = \sum_{x \in \mathcal{X}} T^{(x)}$. The mixed states of a machine are the set:

$$\boldsymbol{\mathcal{H}} = \left\{\eta(w) : w \in \mathcal{X}^+, \Pr(w) > 0\right\} \ . \tag{19}$$

If the process generated by M has a finite ϵ-machine, the mixed-state algorithm finds the recurrent causal-state set by collecting mixed states for an arbitrarily long word. In general, $|\boldsymbol{\mathcal{H}}| \to \infty$, so we typically set a threshold past which if the mixed state set continues to grow, we assume there is no finite representation.

Definition 4. A *flipped* machine $\widetilde{M}$ is a machine where each transition $T^{(x)}_{ij}$ has been replaced with the transition:

$$\widetilde{T}^{(x)}_{ji} = T^{(x)}_{ij} \frac{\pi_j}{\pi_i} \ .$$

This, in effect, flips the direction of the arrows on each transition and renormalizing the probability. This will typically produce a nonunifilar machine.

Definition 5. The *forward switching matrix* S^+ between the forward and reverse ϵ-machines is defined $S^+_{ij} = \Pr(\sigma^+_j | \sigma^-_i)$. The *reverse switching matrix* S^- is similarly defined $S^-_{ij} = \Pr(\sigma^-_i | \sigma^+_j)$.

These pieces allow writing down a simple algorithm for *reversing* an ϵ-machine—i.e., constructing the ϵ-machine in the reverse direction given the forward ϵ-machine.

Algorithm 1 Reverse ϵ-machine

procedure ReverseEM(M_ϵ^+)
  input forward ϵ-machine M_ϵ^+.
  Flip M_ϵ^+.
  Apply the mixed state algorithm to $\widetilde{M_\epsilon^+}$, collecting the unique mixed states in a set $\boldsymbol{\mathcal{H}}_{\widetilde{M_\epsilon^+}}$. If this set converges to a finite set, it consists of the reverse causal states, given in terms of a distribution over forward causal states.
  Stack the mixed states vertically into the forward switching matrix S^+ of shape $M \times N$.
  Initialize empty list T^-.
  for x in $\mathcal{X}$ **do**
    Initialize empty $M \times M$ matrix $T^{-(x)}$.
    for $i = 1, \ldots, M$ **do**
      Calculate probability:

$$\mathbf{e}_i \widetilde{T^+}^{(x)} \mathbf{1} \ .$$

      Calculate next state:

$$\frac{\mathbf{e}_i \widetilde{T^+}^{(x)}}{\mathbf{e}_i \widetilde{T^+}^{(x)} \mathbf{1}} \ .$$

      Initialize empty list.
      for $j = 1, \ldots, N$ **do**
        if next state equals $S^+ e_j$ **then**
          Append probability to list.
        else
          Append a zero to list.
        end if
      end for
      Replace the ith row of $T^{-(x)}$ with list.
    end for
    Append $T^{-(x)}$ to T^-.
  end for
  return M_ϵ^- as list of reverse ϵ-machine transition matrices T^- over symbols $x \in \mathcal{X}$.
end procedure

If one starts from the reverse ϵ-machine, the forward ϵ-machine can be constructed in the expected manner. Indeed, the labeling of the time direction is somewhat arbitrary absent a physical system.

With the forward and reverse ϵ-machines in hand, it is straightforward to construct the bidirectional machine as in Algorithm 2. Since retaining consistent state labeling is important, it is highly recommended to use a data structure capable of containing labeled axes (rows and columns) and to maintain a distinct convention for labeling forward and reverse causal states. As already noted, our convention is to use Latin letters, uppercase for forward states and lowercase for reverse states. This is particularly important when constructing the bidirectional machine.

Let $A^{+(x)}$ be the $N \times N$ forward symbol labeled adjacency matrix of $T^{+(x)}$, which is to say the elements $a_i^+ j$ are one when $T^{+(x)}_{ij} > 0$, indicating a positive probability of transition, and zero otherwise.

Algorithm 2 Bidirectional machine

procedure BidirectionalMachine(A^+, M_ϵ^-)
  input reverse ϵ-machine M_ϵ^-.
  Flip M_ϵ^-.
  Initialize empty list $T^\pm$.
  for x in $\mathcal{X}$ **do**
    From $A^{+(x)}$ construct the block matrix:

$$\begin{bmatrix} a^+_{11}\widetilde{T^-}^{(x)} & \ldots & a^+_{1N}\widetilde{T^-}^{(x)} \\ \vdots & & \vdots \\ a^+_{N1}\widetilde{T^-}^{(x)} & \ldots & a^+_{NN}\widetilde{T^-}^{(x)} \end{bmatrix} , \tag{20}$$

    inheriting state labels as appropriate.
    Drop all rows and columns consisting of only zeroes, leaving a square matrix.
    Append matrix to list of bidirectional machine transition matrices $T^\pm$.
  end for
  return $M_\epsilon^\pm$ as list of bidirectional machine transition matrices $T^\pm$.
end procedure

As with Algorithm 1, the bidirectional machine can be constructed in the "reverse direction", by starting with A^- and M_ϵ^+ and making the appropriate substitutions. Regardless, the same bidirectional machine will be constructed.

Once the bidirectional machine is in hand, calculating a process' prediction taxonomy quantities is conceptually straightforward, if somewhat subtle with regard to tracking indices of the states and observations. See Algorithm 3.

Algorithm 3 Informational anatomy

procedure InfoAnatomyModel($M_\epsilon^\pm$)
  input bidirectional ϵ-machine $M_\epsilon^\pm$.
  Generate list of nonzero measure partitions, according to the indicial rules laid out in Section V B.
  Calculate the probability of all possible transitions of the bidirectional machine from an initial distribution over states. Unless otherwise noted, use the stationary distribution $\pi^\pm$.
  Initialize empty list.
  for A_i in partition **do**
    Apply the information function Eq. (10).
    Append information value to list.
  end for
  return list of information quantities.
end procedure

Once again, data structures capable of retaining labeled axes are recommended, along with a consistent indicial labeling strategy as laid out in Section V B.

VIII. CONCLUSION

This concludes our development of the informational taxonomy of an optimally predicted and retrodicted process. There are several few points of interest to highlight.

Step 3 of Algorithm 3 requires choosing a distribution over the states of the bidirectional machine to determine the probability of paths through the machine (and, of observing words of the process). We have not discussed this aspect of the prediction taxonomy explicitly, implicitly assuming that the process is in the stationary distribution. However, this is a choice, and a potentially interesting one—one can calculate the taxonomy of information measures for any distribution over the states of the bidirectional machine, although the canonical computational mechanics quantities like C_μ are typically defined in terms of the stationary distribution π [6].

As the ϵ-machines are constrained to be ergodic Markov chains over the states, any initial distribution will eventual converge to the stationary distribution when evolved by the state transition matrix T. We conjecture this is true for the bidirectional machine as well, so one can track the convergence of the prediction taxonomy quantities by starting the bidirectional machine away from equilibrium and allowing it to evolve towards the stationary distribution.

Another, alternative analysis is to explore the informational properties of prediction when the machine is constrained to a subset of possible observations. The informational exploration of the ϵ-machine operating away from the stationary state is an intriguing area of exploration that has been considered in related work on thermodynamically coupled ϵ-machines [18]. We reserve the discussion of this avenue for future work.

We also wish to note that this development is closely related to other fine-grained informational analyses of stochastic processes. In particular, we are interested in exploring the relationship between the results here and partial information decomposition [14]. Reference [6] showed that analyzing the quantities described in Section III B with the partial information lattice allows one to relate enigmatic information q_μ to the synergy and redundancy. We are interested in a similar analysis with our new, expanded taxonomy, but this is outside the present scope.

As one may conclude from the indicial rules laid out in Section V B and Algorithm 3, the procedure for generating the informational anatomy of a model can be straightforwardly extended beyond a single time step. Indeed, doing so leads to even more intriguing informational representations of processes and complexity measures. However, this extension too is beyond the present scope, but will be discussed instead in a sequel.

ACKNOWLEDGMENTS

The authors thank the Telluride Science and Innovation Center for hospitality during visits and the participants of the Information Engines Workshops there. This material is based upon work supported by, or in part by, U.S. Army Research Laboratory and the U.S. Army Research Office under grant W911NF-21-1-0048.

[1] J. Bernoulli. *Ars Conjectandi, Opus Posthumum, Accedit Tractatus de Seriebus infinitis, et Epistola Gallice scripta de ludo Pilae recticularis*. Basileae, 1713. Chapters 1–4 translated into English by B. Sung, Ars Conjectandi, Technical Report No. 2, Department of Statistics, Harvard University, 1966.

[2] S.-D. Poisson. *Recherches sur la Probabilité des Jugements en Matiere Criminelle et en Metière Civile, Préédées des Regles Généales du Calcul des Probabilitiés*. Imprimeur–Libraire pour les Mathematiques. Bachelier, Paris, 1837.

[3] P. L. Chebyshev. Des valuers moyennes. *J. Math. Pure Appl.*, 12:177–184, 1867.

[4] A. A. Markov. Rasprostranenie predel'nyh teorem ischisleniya veroyatnostej na summu velichin svyazannyh v cep'. *Zapiski Akademii Nauk po Fiziko-matematicheskomu otdeleniyu*, 3, 1908. Translated into German, Ausdehnung der Satze uber die Grenzwerte in der Wahrscheinlichkeitsrechnung auf eine Summe verketteter Grossen, in: A.A. Markoff (Ed.), Wa- hrscheinlichkeitsrechnung (translated by H. Liebmann), B.G. Teuber, Leipzig, 1912, pp. 272–298. Translated into English, Extension of the limit theorems of probability theory to a sum of variables connected in a chain (translated by S. Petelin) in: R.A. Howard (Ed.), Dynamic Probabilities Systems, vol. 1, Wiley, New York, 1971, pp. 552–576.

[5] C. E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423, 623–656, 1948.

[6] R. G. James, C. J. Ellison, and J. P. Crutchfield. Anatomy of a bit: Information in a time series observation. *CHAOS*, 21(3):037109, 2011. doi:10.1063/1.3637494.

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

[8] C. R. Shalizi and J. P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *J. Stat. Phys.*, 104:817–879, 2001. doi: http://dx.doi.org/10.1023/A:1010388907793.

[9] J. P. Crutchfield, C. J. Ellison, and J. R. Mahoney. Time's barbed arrow: Irreversibility, crypticity, and stored information. *Phys. Rev. Lett.*, 103(9):094101, 2009. doi: 10.1103/PhysRevLett.103.094101.

[10] O. Kallenberg. *Foundations of Modern Probability*. Springer, New York, 2 edition, 2001.

[11] P. Kůrka. *Topological and Symbolic Dynamics*. Société Mathématique de France, Paris, 2003.

[12] R. W. Yeung. *Information Theory and Network Coding*. Springer, New York, 2008.

[13] H. K. Ting. On the amount of information. *Theory of Probability and its Applications*, 7(4):439–447, 1962. doi:10.1137/1107041.

[14] R. D. Beer and P. L. Williams. Information processing and dynamics in minimally cognitive agents. *Cognitive Science*, 39(1):1–38, 2014. doi:10.1111/cogs.12142.

[15] A. Jurgens and J. P. Crutchfield. Divergent predictive states: The statistical complexity dimension of stationary, ergodic hidden Markov processes. *Chaos*, 31(8):0050460, 2021. doi:10.1063/5.0050460.

[16] J. P. Crutchfield, P. Riechers, and C. J. Ellison. Exact complexity: Spectral decomposition of intrinsic computation. *Phys. Lett. A*, 380(9-10):998–1002, 2016.

[17] C. C. Strelioff and J. P. Crutchfield. Bayesian structural inference for hidden processes. *Physical Review E*, 89: 042119, 2014. doi:10.1103/PhysRevE.89.042119.

[18] J. P. Crutchfield and C. Aghamohammdi. Not all fluctuations are created equal: Spontaneous variations in thermodynamic function. *Entropy*, 26(11):894, 2024. doi: 10.3390/e26110894.