Exploring Predictive States via Cantor Embeddings and Wasserstein Distance

Samuel P. Loomis* and James P. Crutchfield†
Complexity Sciences Center and Department of Physics and Astronomy, University of California at Davis, One Shields Avenue, Davis, CA 95616

(Dated: September 27, 2022)

Predictive states for stochastic processes are a nonparametric and interpretable construct with relevance across a multitude of modeling paradigms. Recent progress on the self-supervised reconstruction of predictive states from time-series data focused on the use of reproducing kernel Hilbert spaces. Here, we examine how Wasserstein distances may be used to detect predictive equivalences in symbolic data. We compute Wasserstein distances between distributions over sequences ("predictions"), using a finite-dimensional embedding of sequences based on the Cantor set for the underlying geometry. We show that exploratory data analysis using the resulting geometry via hierarchical clustering and dimension reduction provides insight into the temporal structure of processes ranging from the relatively simple (*e.g.*, finite-state hidden Markov models) to the very complex (*e.g.*, infinite-state indexed grammars).

Keywords: time series, predictive states, Wasserstein distance, hierarchical clustering

LEAD:

Discovering the hidden structure of complex systems has been a widely-recognized goal of nonlinear science for decades, originally starting with extracting "geometry" from a time series. The present results follow directly in this long line of inquiry, especially that focusing on the causal states or effective theories underlying time series generated by complex nonlinear systems. It provides workable statistical estimation methods and dynamical interpretations of hidden mechanisms. It expands upon a new mathematical foundation for causal inference of complex systems, partly by making connections to modern machine learning. In this, its use of Cantor embeddings and Wasserstein distances complements recent work on reproducing-kernel Hilbert space representations.

I. INTRODUCTION

Suppose that we have a finite sequence $x_1 \dots x_L$ of categorical observations drawn from a temporal process. We may suppose that the process is stationary (time-translation invariant) and ergodic (explores all possible behaviors) [1, 2]. We may wish to forecast from the observed information the behavior of the next n observations. If we suspect that the process' temporal correlations do not matter much beyond k symbols, then we take our data to be all subsequences of length $n + k$, splitting each subsequence into words of length k and n respectively. From these "past/future" pairs we can construct an empirical conditional distribution

$$\widehat{P}_{x_1 \dots x_n | x_{-k+1} \dots x_0} = \frac{C_{x_{-k+1} \dots x_n}}{C_{x_{-k+1} \dots x_0}} \ ,$$

where C_w is the number of times the word w appears in our sequence $x_1 \dots x_L$.

Suppose, though, that we do not know how long-range the process' temporal dependencies can stretch. Even very simple stochastic processes can have infinite Markov order, indicating potential long-term dependence of future observations on the past [3]. Given sufficient data, it would be desirable to take pasts of arbitrary length and converge towards a prediction conditioned on the *infinite* past:

$$P_{x_1 \dots x_n | \overleftarrow{x}} = \lim_{k \to \infty} \widehat{P}_{x_1 \dots x_n | x_{-k} \dots x_0} \tag{1}$$

with the infinite sequence $\overleftarrow{x} = (\dots, x_{-1}, x_0)$ of observations stretching into the past. This mathematical ideal is known as the *causal* or *predictive state* [4, 5]. Formally, the conditional predictions $P_{x_1 \dots x_n | \overleftarrow{x}}$ for all forecast lengths n together describe a *probability measure* over future sequences $\overrightarrow{x} = (x_1, x_2, \dots)$; the predictive state is this measure. We denote it simply $P_{\overleftarrow{x}}$.

Predictive states are employed for inference and modeling in dynamical systems [6], renewal processes and neural spike-trains [7, 8], condensed matter physics [9], and spatiotemporal systems [10]. A deep mathematical theory of predictive-state inference has been correspondingly developed [3, 5, 11–16]. If the dataset $x_1 \dots x_L$ is drawn from any stationary process and if L is sufficiently large, then Eq. (1) converges for any n and a probability-1 subset

* sloomis@ucdavis.edu
† chaos@ucdavis.edu

of pasts $\overleftarrow{x}$ [3]. Recently, this result has been further clarified: in the language of measures, $P_{\overleftarrow{x}}$ converges *in distribution*, with respect to the *product topology* of the space of sequences $\mathcal{X}^{\mathbb{N}}$ [17].

The broad goals of predictive-state analysis are threefold [18]. The first is to understand the overall structure of how the predictive states relate to one another geometrically and, possibly, use this geometry to classify pasts based on equivalence of their predictive states. The second is to actually reproduce the prediction to a specified accuracy. The third is to understand the dynamics of how predictions evolve under a stream of new observations.

The following focuses on the first, as it is a crucial building block to achieving the other two. Recent attempts to reconstruct the geometry of predictive states embedded them in reproducing kernel Hilbert spaces (RKHS) [16, 17, 19–21]. This was achieved to great effect largely since convergence in Hilbert spaces generated by universal kernels is equivalent to convergence in distribution [22]. That is, these embeddings allow accurate representations of predictive states because they respect the product topology of sequences in the same way that predictive states themselves do [17]. Here, we achieve the same results using a procedure that can be easily visualized and interpreted. This makes it suitable as a method for exploratory data analysis [23, 24], while moving further toward interpretable machine learning of structured processes.

Understanding the source of the power of RKHS methods in predictive-state analysis frees us to consider other options. The following embeds symbolic sequences in a one-dimensional space that has the same topology as the product space. This embedding is inspired by the fractal Cantor set [25]. Predictive states can then be thought of as distributions in this one-dimensional space. We then use the Wasserstein distance to compute the geometry between predictive states, which is determined by a closed-form integral for one-dimensional distributions. This operates as an alternative to RKHS-based distances since the Wasserstein distance also reproduces the topology of convergence in distribution [26]. The resulting distance matrix then is used to find low-dimensional embeddings [27] of the geometry or hierarchical clusterings [28] of the predictive states. When combined with the fractal embedding, the latter, in particular, provides a highly interpretable visualization of the predictive-state space.

II. EXAMPLE PROCESSES

The methods here are intended to be applied to stationary and ergodic stochastic processes that generate categorical time-series data. For these purposes we consider a stochastic process to be a collection of probability distributions $\Pr_\mu(\, x_1 \ldots x_L \,)$ over any finite, contiguous sequence, taking values in a finite set $\mathcal{X}$. Formally, this describes a measure μ over the set of all bi-infinite sequences $(\ldots, x_{-1}, x_0, x_1, \ldots) \in \mathcal{X}^{\mathbb{Z}}$.

These processes are generated by a number of systems with widely varying complexity. Most popularly studied are those often characterized as having a degree of "finite memory": *Markov chains*, *hidden Markov chains*, and *observable operator models* (also termed *generalized hidden Markov chains*) [3, 5]. Beyond these, one can also generate processes using probabilistic grammars, such as probabilistic context-free and indexed grammars [29]. Additionally, coarse-grained data from chaotic dynamical systems—such as the *logistic map*—display behavior varying widely in complexity [6].

We refer back frequently to the following example processes of increasing computational complexity:

1. The *fair coin process* is simply generated by flipping a coin repeatedly and writing down a 0 for every tail and 1 for every head. There is no memory in this process, and all pasts will map to the same predictive state. We will discuss it occasionally as a counterpoint to the more complex examples below. We also provide code and figures for this example in our GitHub repository [30].
2. The *even process* can be generated by repeatedly tossing a coin and writing down a 0 for every tail and 11 for every head. The process is essentially random except that 1s only appear in contiguous blocks of even size bounded by 0s. The even process has infinite Markov order but can be generated by a two-state hidden Markov chain [31]. A typical example might look like 01100111101100011.
3. The $\mathtt{a}^n\mathtt{b}^n$ *process* can be generated by choosing a random integer $n \geq 1$ (we suppose via a Poisson process) and writing n `a`s followed by an equal number of `b`s, and then repeating this procedure indefinitely. This results in sequences where any contiguous block of `a`s is followed by a block of `b`s of equal size. The $\mathtt{a}^n\mathtt{b}^n$ process cannot be generated by any finite hidden Markov chain, though it is a simple example of a probabilistic context-free language [32]. A typical example might look like `abaaabbbabaabb`.
4. The $x + f(x)$ *process* is a probabilistic context-free language modeling the syntactic structure of simple mathematical expressions. It has terminal symbols {`(` , `)` , `;` , `+` , `f` , `x`} and nonterminals $\{A, B, C\}$, and starts with a finite sequence of As (such as AAA). Sequences are generated by applying the

production rules:

$$\begin{aligned} &A \mapsto B + C \ ; \mid C \ ; \\ &B \mapsto B + C \mid C \\ &C \mapsto \mathtt{f}(B) \mid \mathtt{x} \ . \end{aligned}$$

The sequence generation process does not "end" until all non-terminals (that is, A's, B's and C's) have been replaced by terminal symbols. A typical example might look like $\mathtt{x}+\mathtt{x};\mathtt{f}(\mathtt{x}+\mathtt{f}(\mathtt{x}));\mathtt{f}(\mathtt{f}(\mathtt{x}))$; which results from the substitutions

$$\begin{aligned} AAA &\mapsto B + C \ ; C \ ; C \ ; \\ &\mapsto C{+}\mathtt{x};\mathtt{f}(B);\mathtt{f}(B); \\ &\mapsto \mathtt{x}+\mathtt{x};\mathtt{f}(B+C);\mathtt{f}(C); \\ &\mapsto \mathtt{x}+\mathtt{x};\mathtt{f}(C+\mathtt{f}(C));\mathtt{f}(\mathtt{f}(C)); \\ &\mapsto \mathtt{x}+\mathtt{x};\mathtt{f}(\mathtt{x}+\mathtt{f}(\mathtt{x}));\mathtt{f}(\mathtt{f}(\mathtt{x})); \end{aligned}$$

5. The $\mathtt{a}^n\mathtt{b}^n\mathtt{c}^n$ *process* is a probabilistic indexed language [32] that is analogous to $\mathtt{a}^n\mathtt{b}^n$ except after writing the blocks of a's and b's, we also write a block of c's of length n. A typical example might look like abcaaabbbcccabcaabbcc. Though it looks very similar to the $\mathtt{a}^n\mathtt{b}^n$ process, it is in a formally distinct complexity class, as $\mathtt{a}^n\mathtt{b}^n$ may be generated with a simple stack automaton while $\mathtt{a}^n\mathtt{b}^n\mathtt{c}^n$ requires nested stacks.
6. The *Morse-Thue process* is generated by sampling from the time series of the logistic map at critical "onset of chaos" parameter $r_c \approx 3.56995$:

$$y_{t+1} = r y_t (1 - y_t)$$

and then coarse-graining the data by taking $x_t = 0$ if $0 < y_t \leq \frac{1}{2}$ and $x_t = 1$ if $\frac{1}{2} < y_t < 1$ [25]. Alternatively, we can generate this process by starting with a single 0 and executing the replacements $0 \mapsto 11$ and $1 \mapsto 01$ consecutively. The resulting process is an indexed-context free language [6]. A typical example might look like 1101110101011101110110101011101—the fifth generation of the replacement rule starting from 0.

Though our examples are abstract, they represent real structures which may arise in practice.

Examples 3 through 5, in particular, involve long-range correlations in their sequences which arise from structures which, upon "opening," must be "closed" an arbitrary time later. Such structures are present in human language as well as in genetic sequences. In the latter case, this long-range correlation exists due to the folding of proteins which allows distant codons to interact with one another. Detecting and modeling these folding structures requires more complex models than hidden Markov models, such as the context-free grammars in examples 3 and 4 [33]. Example 6 represents the general problem of recovering knowledge of a dynamical system of which only coarse-grained data is available.

III. CANTOR-EMBEDDING SEQUENCES

The geometry of sequences is inherently self-similar. Given an infinite sequence $\overrightarrow{x} = (x_1, x_2, \dots)$, we can split it into its leading word $x_1 x_2 \dots x_L$ and a following sequence $\overrightarrow{x}_L = (x_{L+1}, x_{L+2}, \dots)$. That is, the space of sequences $\mathcal{X}^{\mathbb{N}}$ can be factored into $\mathcal{X}^L \times \mathcal{X}^{\mathbb{N}}$ for any L; much like a fractal, then, $\mathcal{X}^{\mathbb{N}}$ is comprised of copies of itself. The fractal nature of sequence-space is encoded in the structure of its *product topology*, in which the closeness of two sequences is measured by the number of sites at which their values match. In other words, the neighborhoods of the product topology are generated by these smaller copies of $\mathcal{X}^{\mathbb{N}}$.

We can visualize this self-similarity in an interesting way by constructing a mapping between sequence space and the celebrated *Cantor set* (or one of its generalizations). Suppose a symbolic sequence $(x_1, x_2, \dots)$ takes values in an alphabet $\mathcal{X}$ of size $|\mathcal{X}|$. To each $x \in \mathcal{X}$ we associate a unique integer between 0 and $|\mathcal{X}| - 1$ inclusive; call this $J(x)$. Then, there is a function $C : \mathcal{X}^{\mathbb{N}} \to [0, 1]$ that maps every sequence to a positive real number:

$$C(x_1, x_2, \dots) = \sum_{k=1}^{\infty} \frac{2J(x_k)}{(2|\mathcal{X}| - 1)^k} \ .$$

For instance, suppose that $|\mathcal{X}| = 2$ has two elements; then the mapping C looks like

$$C(x_1, x_2, \dots) = \sum_{k=1}^{\infty} \frac{2J(x_k)}{3^k} \ .$$

The range of this mapping, taken over all binary sequences, matches the traditional Cantor set fractal, which is attained by removing the middle third from the unit interval, and repeating this process on all remaining intervals *ad infinitum*. When $\mathcal{X}$ has more than two elements, the resulting range is a generalized form of the Cantor set, where instead of removing the middle (second) third, we remove the second and fourth quintiles ($|\mathcal{X}| = 3$), or second, fourth and sixth heptiles ($|\mathcal{X}| = 4$), and so on. For a finite sequence of length L, we truncate the sum at $k = L$, resulting in a finite-depth approximation to the Cantor set.

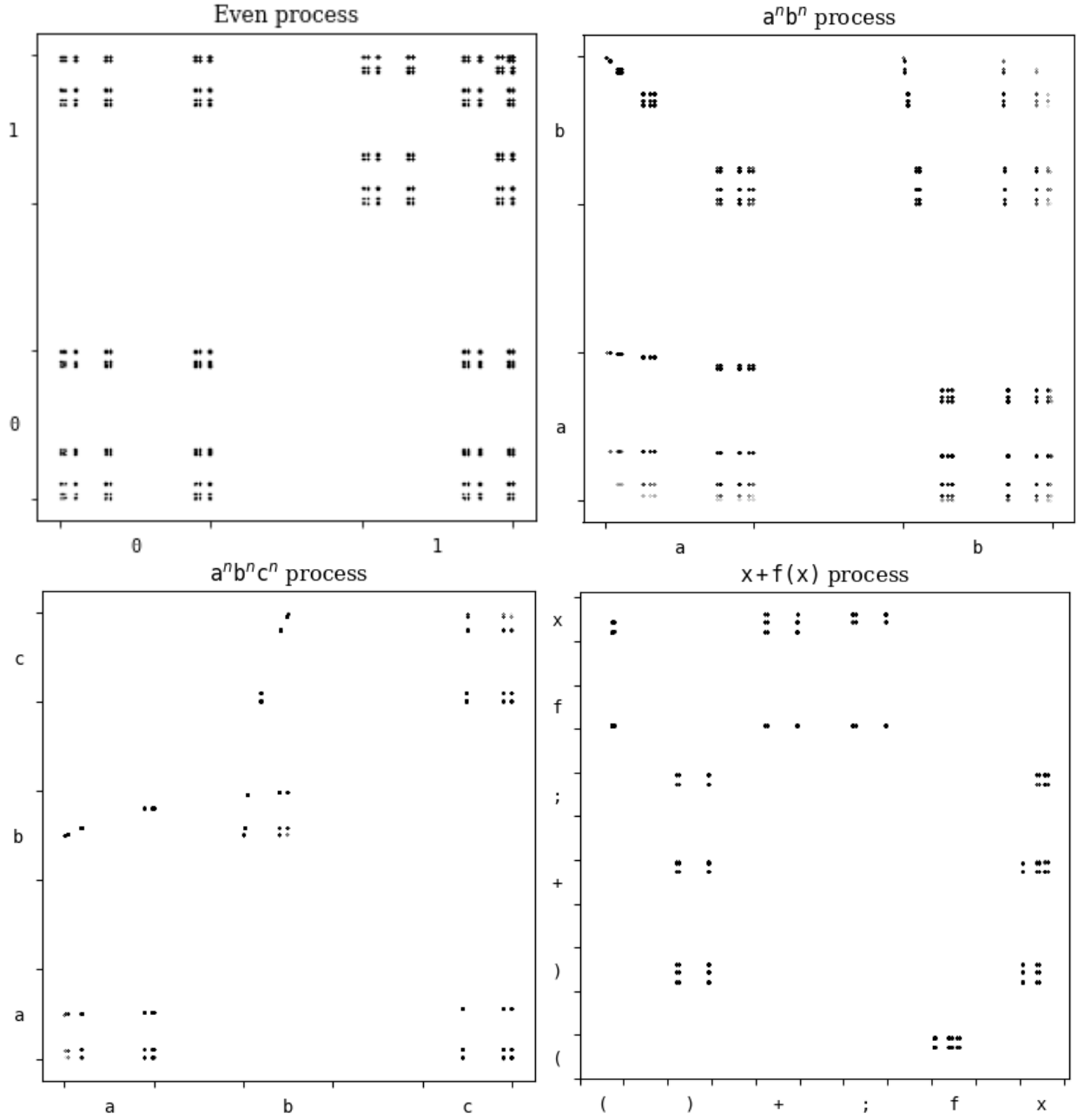

FIG. 1: Cantor plots for the even, $\mathtt{a}^n\mathtt{b}^n$, $\mathtt{a}^n\mathtt{b}^n\mathtt{c}^n$, and $\mathtt{x}+\mathtt{f}(\mathtt{x})$ processes. Each point (x, y) corresponds to a pair of sequences corresponding to the past x and future y, respectively. The symbol on the x (y) axis indicates that all points above (to the right of) that symbol have a past (future) whose most recent observation is that symbol. Though not marked, further proportional subdivisions of each segment of the axes indicate the value of the second, third, and so on symbols. For instance, one can read from the $\mathtt{x}+\mathtt{f}(\mathtt{x})$ fractal that any past ending in `f` must be paired with a future beginning in (`f` or (`x`.

Why use this somewhat convoluted construction for an embedding? The reason is that the embedding C preserves the basic structure of the product topology on sequences [25]. Imagine, for instance, that we chose a more direct map, such as mapping a binary sequence to the corresponding binary number (*e.g.* $010110\cdots \mapsto 0.010110\ldots$): then dissimilar sequences such as $0111\ldots$ and $1000\ldots$ would in fact map to the same real number. The padding added by removed intervals in the Cantor set means we can map distinct sequences to distinct real numbers, while keeping similar sequences close together.

Stationary processes, due to their time-translation invariance, inherit the fractal temporality of sequence space. This can be easily visualized: Given a length-L sample

$x_1 \dots x_L$, and $n, k > 0$, take a sliding window of pasts and futures, $(x_{t-k+1} \dots x_t, x_{t+1} \dots x_n)$ for $t = k, \dots, L-n$. For each past-future pair, compute the truncated Cantor embeddings on the *reversed* past and (unreversed) future: $(C(x_t \dots x_{t-k+1}), C(x_{t+1} \dots x_n))$. The resulting pairs of real numbers can be plotted as (x, y)-values on a scatter plot. The fractal that emerges contains, in essence, all information necessary to understand a process' temporal structures. See Fig. 1 for examples and guidance on how to interpret the visualization.

Note that for $|\mathcal{X}| > 2$ the embedding C introduces additional structure that may or may not be desired. Associating each symbol x with an integer $J(x)$ endows an ordinal structure on the set $\mathcal{X}$. That is, the specific mapping from symbols to integers is arbitrary, and the way the symbols are ordered in the Cantor embedding—for instance, which symbol gets assigned to the middle segment instead of the end segments when $|\mathcal{X}| = 3$—can affect the results of the visualization and downstream processing. Higher-dimension embeddings may be contrived which do not impose an ordering on the symbols. However, they come at the cost of increased computational complexity when determining the Wasserstein distance. So, for now, we assume that ordinal artifacts are either desired or sufficiently tolerable as to not outweigh the computational benefits of working in one dimension.

IV. WASSERSTEIN DISTANCE ON PREDICTIVE STATES

Figure 1's Cantor fractals represent probability distributions: We interpret a vertical slice of the fractal, located at horizontal position $C(\overleftarrow{x})$, as visualizing the predictive state $P_{\overleftarrow{x}}$ as a distribution over Cantor-embedded futures $C(\overrightarrow{x})$.

For example, by examining the even process' Cantor fractal, one notices that there are effectively only 2 distinct predictive states—every vertical column is just one of two types. This corresponds with the 2 states of the hidden Markov model that generates the even process. By comparison, if we had plotted the fair coin's Cantor fractal, we would have found it to be perfectly symmetric, with every vertical column identical.

This allows us to see how predictive states distribute their probability over the intrinsic geometry of potential futures. We compare predictive states not only on how much their supports overlap, but on how geometrically close their supports are to one another. For the $\mathtt{a}^n\mathtt{b}^n$ process, for example, we see that the first few columns (corresponding to pasts of the form $\dots \mathtt{ba}^n$ for some n) are inherently similar to one another, though they are shifted upwards the closer to the axis they are. (The latter corresponds to the increasing number of $\mathtt{b}$s in the predicted future as n increases.)

This provides some intuition for how to effectively choose a distance metric for predictive states. Metrics such as the total variation distance or divergence-based distance metrics are too coarse; a small shift in the support of the distributions can result in dramatic changes in the distances between the distributions. Instead, we want a distance metric for distributions that quantifies the differences between the supports. The Wasserstein metric, also known as the earth-mover's distance, provides this feature [26]. As the name "earth-mover's" suggests, we may imagine the probability distributions as masses of dirt, and quantify the difference between two distribution

Algorithm 1: Convert a sequence of categorical time-series data into a labeled collection of empirical distributions of Cantor-embedded futures and a matrix of Wasserstein distances between said distributions.

function `CantorWasserstein` $(k, n, x_1 \dots x_L)$;
input : Integers k, n of past and future lengths
input : Length-L sequence $x_1 \dots x_L$ of observations
output : List UnqPasts of unique pasts
output : List of lists Cantors of Cantor-embedded futures
output : Matrix Wass of Wasserstein distances
UnqPasts $\leftarrow []$;
Cantors $\leftarrow []$;
for $t \leftarrow n$ **to** $L - k$ **do**
  $\overleftarrow{\mathbf{x}} \leftarrow [x_{t-k+1}, \dots, x_t]$;
  $\overrightarrow{\mathbf{x}} \leftarrow [x_{t+1}, \dots, x_{t+n}]$;
  $p \leftarrow \sum_{\ell=1}^{n} \frac{2J(\overrightarrow{x}_\ell)}{(2|\mathcal{X}|-1)^\ell}$;
  if $\overleftarrow{\mathbf{x}} \in$ UnqPasts **then**
    `append` $\overleftarrow{\mathbf{x}}$ **to** UnqPasts;
    `append` $[p]$ **to** Cantors;
  else
    $j \leftarrow$ `index` $(\overleftarrow{\mathbf{x}}, \text{UnqPasts})$;
    `append` p **to** Cantors$_j$;
  end
end
$K \leftarrow$ `length`(UnqPasts);
Wass $\leftarrow$ `Matrix`(K, K);
for $i \leftarrow 1$ **to** K **do**
  for $j \leftarrow 1$ **to** K **do**
    Wass$_{ij} \leftarrow$ `Wasserstein`(Cantors$_i$, Cantors$_j$);
    Wass$_{ji} \leftarrow$ Wass$_{ij}$;
  end
end
Result: UnqPasts,Cantors,Wass

as the cost of shifting the dirt around to transform one distribution to another. This way, if two distributions have similar but non-overlapping supports, the Wasserstein metric will be small.

Formally, we can define the Wasserstein metric as follows. Given two measures μ and ν defined on a metric space $\mathcal{M}$ with metric d, the Wasserstein distance between μ and ν is given by:

$$W(\mu, \nu) = \min_{\pi \in \Gamma(\mu,\nu)} \int_{\mathcal{M}\times\mathcal{M}} d(x, y) d\pi(x, y) \ ,$$

where $\Gamma(\mu, \nu)$ is the set of all measures on $\mathcal{M} \times \mathcal{M}$ whose left and right marginals are μ and ν, respectively. It is the minimal cost to "shift" the probability mass from one distribution to match the other's shape.

$W(\mu, \nu)$ is the solution to a constrained linear optimization (since the objective function and constraints are linear functions of π). This can become computationally costly, scaling as $O(n^3 \log n)$ in the number of samples n [34]. However, when $\mathcal{M} \subseteq \mathbb{R}$, there is in fact a closed-form solution to the Wasserstein optimization problem [35]. Let F and G respectively be the cumulative distributions functions of μ and ν. Then:

$$W(\mu, \nu) = \int_{-\infty}^{\infty} |F(t) - G(t)| \, dt \ .$$

This closed-form solution is considerably faster to compute than the linear optimization required for arbitrary metric spaces. Since the Cantor embedding embeds the space of sequences directly into $[0, 1]$, we can directly employ this formula.

Topologically, the Wasserstein distance replicates the topology of convergence in distribution, which means that algorithms which construct predictive states using the Wasserstein distance will converge.

Combining the Cantor embedding and Wasserstein distance leads to a straightforward program for analyzing categorical time series:

1. Apply Algorithm 1 to the data stream $x_1 \ldots x_L$ for a specified past length k and future length n, retrieving the (i) set of unique observed pasts, (ii) empirical Cantor distributions corresponding to each past, and (iii) matrix of Wasserstein distances computed from these distributions.

2. To elucidate the relative geometry of the predictive states, use the Wasserstein distance matrix to perform additional methods of geometric data analysis, such as hierarchical clustering [28] and multidimensional scaling [27].

The next two sections examine the results of this approach.

V. INTERPRETABLE PREDICTIVE STATES WITH HIERARCHICAL CLUSTERING

Figure 2 displays the result of collecting the Cantor-embedded empirical predictions for all pasts of a given length for four processes—even, $\mathtt{a}^n\mathtt{b}^n$, $\mathtt{a}^n\mathtt{b}^n\mathtt{c}^n$, and $\mathtt{x} + \mathtt{f}(\mathtt{x})$. For each, the Wasserstein distance between every pair of predictions was computed and used to hierarchically cluster the pasts with others that produced similar predictions, using the Ward method [28].

The resulting clustered Cantor plots offer a highly interpretable visualization of the relationship between pasts and futures and of the predictive states' geometry. Each plot, in a certain sense, sorts the columns in the Cantor fractals of Fig. 1 with the white space between columns removed. For instance, the even process's clustered Cantor plot clearly contains the two major states, with a third "transient" state visible. (The latter corresponds to the increasingly unlikely event of never seeing a 0 in a block of length n.) This third state was previously hidden mostly out of view on the far-right side of the 2-dimensional Cantor plot of the even process in Fig. 1.

Other features are worth calling out. Close observation shows that hierarchical clustering reveals the (mostly) scale-free distinctions between pasts with subtle differences. For the $\mathtt{a}^n\mathtt{b}^n$ process, pasts of the form $\ldots\mathtt{b}\mathtt{a}^n$ are distinguished for different n, as each involves a distinct number of $\mathtt{b}$'s appearing in the near future. Meanwhile, the clustering scheme carefully distinguishes pasts of the form $\ldots\mathtt{b}\mathtt{a}^n\mathtt{b}^{n-k}$ for different k but *not* for different n, as k is the essential variable for predicting the remaining number of $\mathtt{b}$'s. (The scale-free discernment of the algorithm breaks down past $n = 5$—the scale at which sampling error becomes relevant for our chosen sample size.)

Similar discernment is seen for the $\mathtt{a}^n\mathtt{b}^n\mathtt{c}^n$ and $\mathtt{x} + \mathtt{f}(\mathtt{x})$ processes as well. We draw attention to the manner in which the presence of a semicolon in pasts from the $\mathtt{x} + \mathtt{f}(\mathtt{x})$ process affects the comparison of predictions.

By analyzing clustered Cantor plots, one gains insight into the properties of pasts that make them similar in terms of future predictions, even if they are superficially quite distinct. Furthermore, the horizontal axis allows for continued use of the Cantor set's natural geometry for visualizing the future forecasts associated with each cluster of predictions.

VI. PREDICTIVE STATE GEOMETRY WITH MULTIDIMENSIONAL SCALING

Sacrificing direct visualization of future predictions leads to a more intuitive picture of predictive-state space geom-

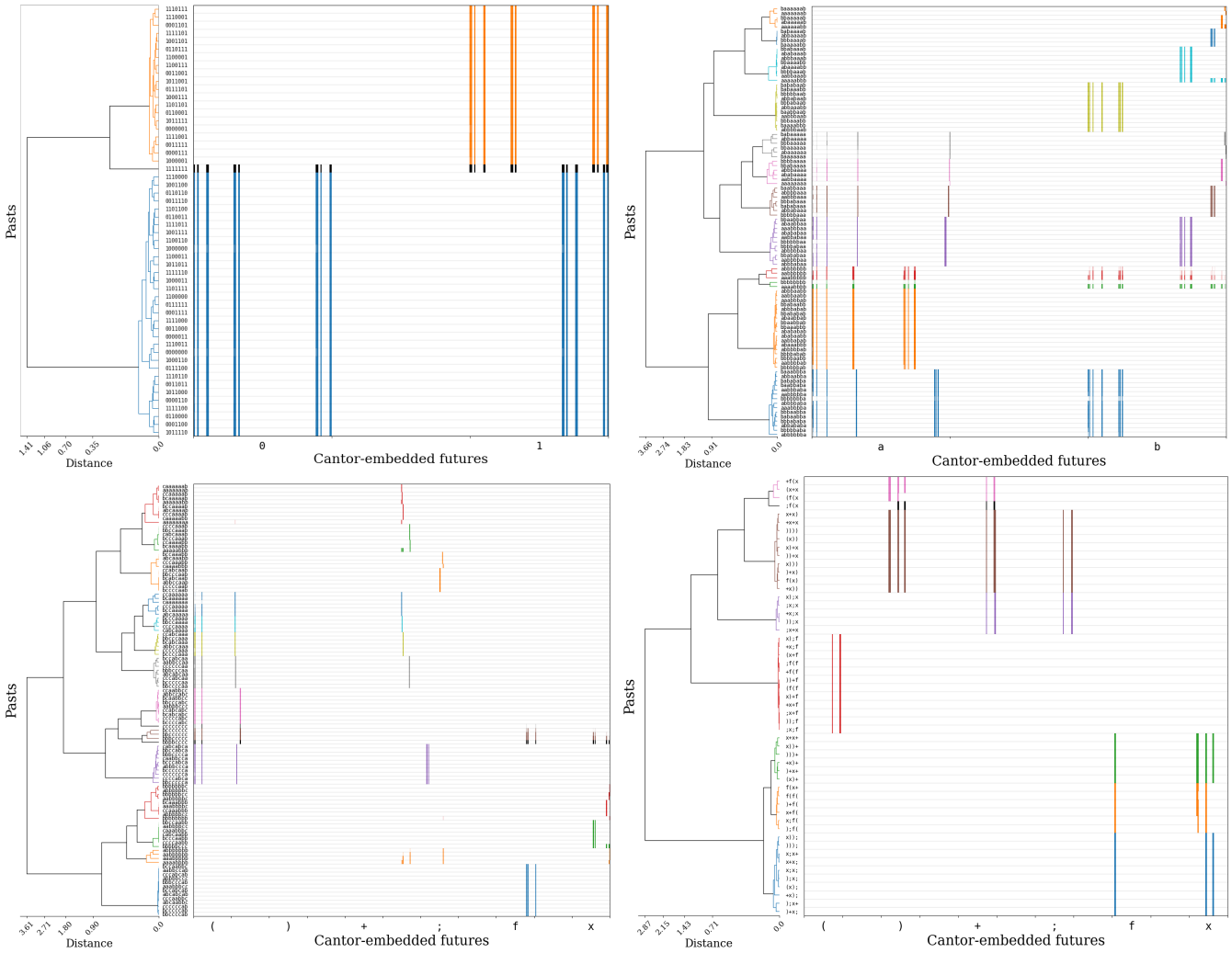

FIG. 2: (Upper left to lower right) Clustered Cantor diagrams of the even, $\mathtt{a}^n\mathtt{b}^n$, $\mathtt{a}^n\mathtt{b}^n\mathtt{c}^n$, and $\mathtt{x}+\mathtt{f}(\mathtt{x})$ processes. Zoom for detail. For each, the vertical axis shows all pasts of a given length k along with their hierarchically clustered dendrogram. $k = 8$ for the even, $\mathtt{a}^n\mathtt{b}^n$, and $\mathtt{a}^n\mathtt{b}^n\mathtt{c}^n$ processes and $k = 4$ for the $\mathtt{x}+\mathtt{f}(\mathtt{x})$ process. For present purposes, the coloring threshold was chosen to aid visual interpretation. The lines in each row show the empirical distribution of Cantor-embedded futures observed following each past. As such, the horizontal axis corresponds exactly to the vertical axis of Fig. 1.

etry. Applying any desired dimension reduction algorithm to the matrix of Wasserstein distances between predictions yields a coordinate representation of the similarities between predictive states.

Figure 3 plots the first two dimensions of a multidimensional scaling (MDS) decomposition [27] for the even and $\mathtt{a}^n\mathtt{b}^n$ processes. MDS extracts coordinates so that the Euclidean distance between points in MDS coordinates matches the Wasserstein distance between the distributions being embedded. Clusters are colored in the same manner as in Fig. 2 and labeled by the specific pattern that distinguishes the pasts in some of the clusters. Note that the clusters and labels are directly drawn from Fig. 2 for reference. They are not the result of the MDS algorithm itself. However, interactive plotting approaches may allow for similar exploration from these decompositions without the need for prior clustering.

The even process, as in all other cases seen thus far, has two dominant prediction clusters. These correspond to the predictive states that result from seeing an even-sized block of 1s (or, equivalently, no 1s) and that result from seeing an odd-sized block of 1s. The lone cluster in the middle corresponds to a transient state induced by seeing all 1s and then a 0. The latter then synchronizes to right cluster. By comparison, the fair coin's MDS embedding would have a single cluster since every past predicts the same future.

The $\mathtt{a}^n\mathtt{b}^n$ plot is much more sophisticated. Intriguingly,

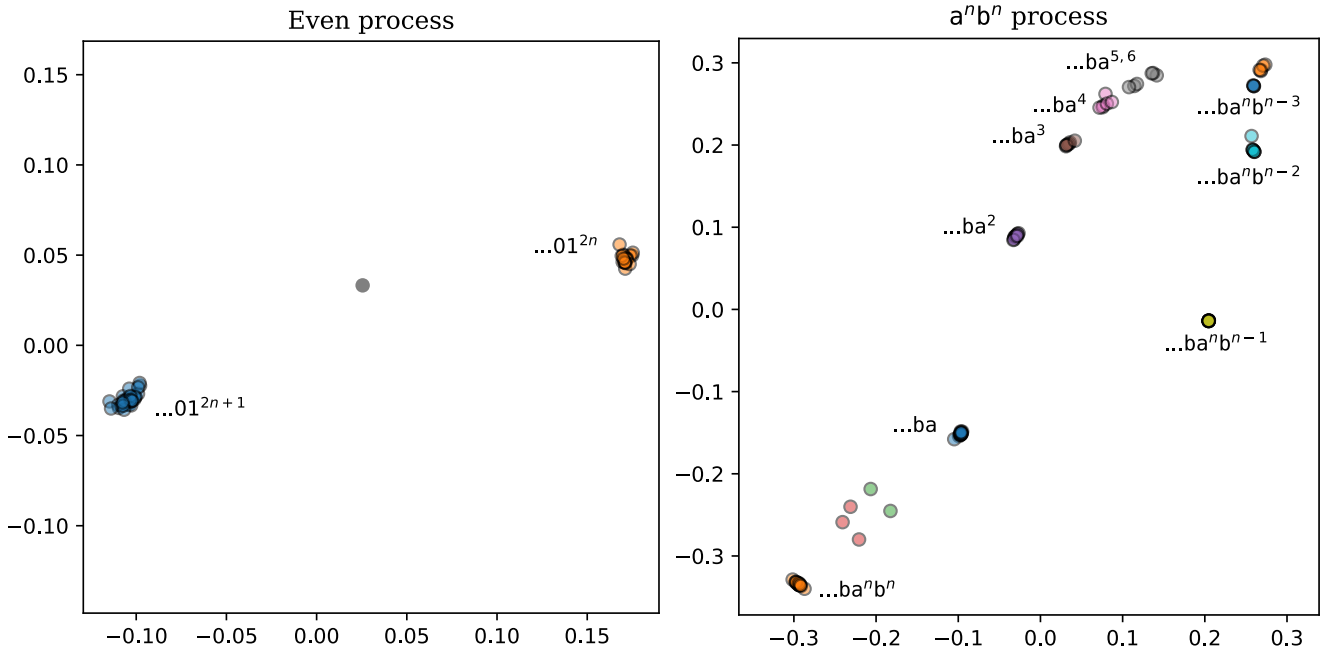

FIG. 3: Scatterplots of the first two MDS coordinates of the reconstructed predictive states: (Left) Even process. (Right) $\mathtt{a}^n\mathtt{b}^n$ process. Clusters colored according to the scheme determined by the dendrogram in Fig. 2 and the label on each cluster describes the pattern that uniquely characterizes the pasts in that cluster.

its geometry not only clearly distinguishes predictively distinct states, but organizes them in a manner highly suggestive of an *pushdown stack*. The latter is particularly appropriate given that stack automata are the natural analog of hidden Markov chains but for context-free languages. Observing more a's pushes more symbols onto the stack, with the predictive states moving further up towards the plot's upper-right corner. And, as more b's are observed the top symbol is popped off the stack, and the predictive states move back towards the lower left. The latter represents equality between a's and b's.

The geometric approach is particularly insightful when computing the Wasserstein matrix between predictions estimated from Morse-Thue process data. Recall that the Morse-Thue process is a coarse-graining of the iterated logistic map $y_{t+1} = ry_t(1 - y_t)$, $y_0 = 0$, at the critical chaos parameter $r_c \approx 3.56995$. The resulting stream of 0s and 1s is a well-known instance of high complexity at the "order-disorder border". Specifically, setting parameter r on either side of r_c results in sequences that can be generated by finite hidden Markov chains. However, at r_c itself the resulting Morse-Thue process is context-sensitive and therefore requires infinite predictive states. That is, when it comes to capturing its behavior, the process is several orders higher in model complexity. It is further up the Chomsky language hierarchy.

Despite this high order of structural complexity, the predictive state geometry reconstructed from a sufficiently large sample of the Morse-Thue process recovers the neighborhoods of $[0, 1]$ that are relevant to the dynamics of the original logistic map. Said differently, there is a correspondence between each past $x_{-k+1} \ldots x_0$ and a subset $V_{x_{-k+1} \ldots x_0}$, such that $V_{x_{-k+1} \ldots x_0}$ is the set of all points y for which $x(f^{-t}(y)) = x_{-t}$ for $0 \leq t < n$. (Here, $f(y) = ry(1 - y)$ and $x(y)$ is the encoding $y \mapsto \{0, 1\}$.) As it happens, pasts $x_{-k+1} \ldots x_0$ whose predictive states are close under the Wasserstein distance are also pasts for which the sets $f(V_{x_{-k+1} \ldots x_0})$ are close. That is, they correspond to predictively similar ranges of the logistic map variable.

Figure 4 directly visualizes the relationship between the reconstructed predictive states of the Morse-Thue process, neighborhoods of the logistic variable y, and the logistic map dynamics. In short, despite the fact that the Morse-Thue process is a highly coarse-grained form of the logistic map, the essential geometry of that map can be recovered by reconstructing predictive state geometry with the Wasserstein metric and the Cantor embedding.

Note that, due to the deterministic nature of the Morse-Thue process, the combination of the Wasserstein metric and the Cantor embedding is particularly important to achieving this result. Asymptotically, each past corresponds to a unique future. And so, there is asymptotically no overlap between predictions. The choice of the Cantor map facilitates placing together forecasts that match up to a certain time in the future. And, the Wasserstein

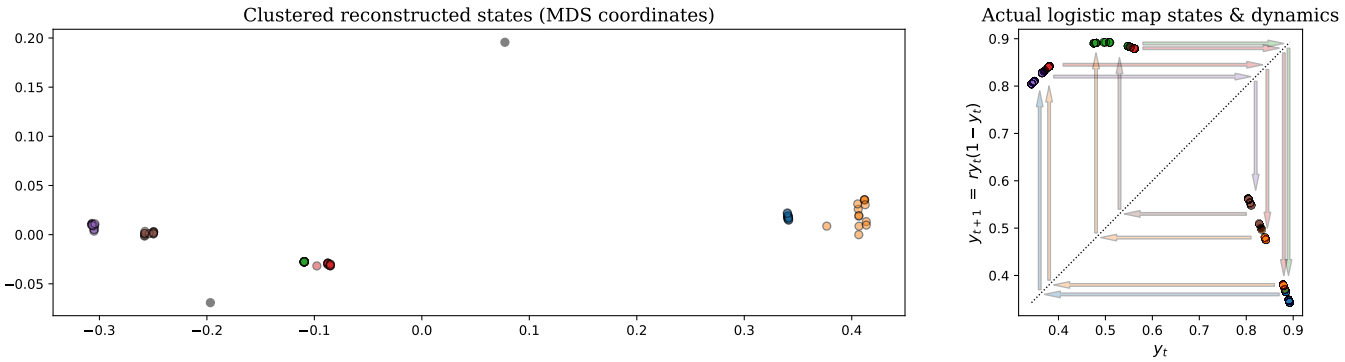

FIG. 4: (Left) Scatterplot of the first two MDS coordinates of the reconstructed predictive states for the Morse-Thue process, color-coded by cluster. (Right) Scatterplot of the corresponding points in the domain of the logistic map, plotting for each point both the present value y_t and the next value y_{t+1}, with the $x = y$ line for reference. Each pair of color-coded arrows shows where each cluster maps to under the action of the logistic map. The predictively reconstructed clusters thus correspond to dynamically similar neighborhoods of the logistic map domain.

distance allows directly comparing predictions whose supports are geometrically close. In this way, the combination of the two approaches enables the straightforward recovery of the underlying dynamical system's (logistic map's) geometry.

We also note that the direct reconstruction of the logistic map in this case is possible due to the choice of partition to transform the continuous variable y into the discrete variable x. The partition we chose is a *generating partition* of the dynamical system [25], which is a special kind of partition which preserves many of the measure-theoretic and topological features of the original system. Non-generating partitions would not yield the original logistic map if our methods are applied to them; however, nor would they if any other method were applied, as a non-generating partition represents an irreversible loss of information. In this case predictive state reconstruction would still recover the minimal sufficient statistics of the past sequence for future prediction (see, for instance, the discussion in [36]).

VII. CONCLUDING REMARKS

We presented a general approach for predictive state analysis—Cantor fractal embedding sequences and Wasserstein distance comparison of predictions. We offered two approaches to visualizing the results of this method—one a direct application of multidimensional scaling and the other being a clustered Cantor diagram built from combining hierarchical clustering with the introduced Cantor embedding.

Compared to using reproducing kernel Hilbert spaces—a dominant approach to predictive states at present [16, 17, 19–21]—our combining the Cantor set with the Wasserstein distance may appear idiosyncratic. However, as the results demonstrated, there are strong benefits to both and together the two methods synergize their benefits in a unique way. The topology of convergence in distribution can be replicated with both the Wasserstein distance and the RKHS inner product. However, the Wasserstein distance depends on far fewer parameters—such as, the choice of the eponymous kernel in RKHS approaches. Moreover, its value is directly interpretable in terms of the shapes of the distributions it compares.

Similarly, there are many ways to metrize the product topology on sequences, but the Cantor embedding offers a direct way to connect the product topology with a visualizable geometry. And, embedding in a single dimension enables efficient computation of the Wasserstein metric. The benefits of the Cantor and Wasserstein approaches adds interpretability to the resulting predictive-state geometry along two distinct axes, most clearly seen in Fig. 2's clustered Cantor diagrams. We hope that the success of this approach in providing clear insights will complement existing thrusts in the direction of abstract embeddings and mathematical formalism by motivating further development on interpretable approaches to predictive state analysis.

ACKNOWLEDGMENTS

We thank Nicolas Brodu, Adam Rupe, Alex Jurgens, David Gier, and Mikhael Semaan. JPC acknowledges the kind hospitality of the Telluride Science Research Center, Santa Fe Institute, Institute for Advanced Study at the University of Amsterdam, and California Institute of Technology for their hospitality during visits. This material is based upon work supported by, or in part by, Templeton World Charity Foundation Diverse Intelligences grants

TWCF0440 to the SETI Institute and TWCF0570 to UC Davis and by Foundational Questions Institute and Fetzer Franklin Fund grant number FQXI-RFP-CPW-2007 and grants W911NF-18-1-0028 and W911NF-21-1-0048 from the U.S. Army Research Laboratory and the U.S. Army Research Office.

REPRODUCIBILITY STATEMENT

For the purposes of reproducibility, we provide a GitHub repository [30] containing the code necessary to generate this manuscript and its figures, including a notebook for generating the data used for the examples and for building both static and interactive figures for further exploration.

[1] W. Feller. *An Introduction to Probability Theory and its Applications*. Wiley, New York, third, revised edition, 1970. 1

[2] O. Kallenberg. *Foundations of Modern Probability*. Springer, New York, 2 edition, 2001. 1

[3] 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. 1, 2

[4] J. P. Crutchfield and K. Young. Inferring statistical complexity. *Phys. Rev. Let.*, 63:105–108, 1989. 1

[5] H. Jaeger. Observable operator models for discrete stochastic time series. *Neural Computation*, 12(6):1371–1398, 2000. 1, 2

[6] J. P. Crutchfield. The calculi of emergence: Computation, dynamics, and induction. *Physica D*, 75:11–54, 1994. 1, 2, 3

[7] S. Marzen, M. R. DeWeese, and J. P. Crutchfield. Time resolution dependence of information measures for spiking neurons: Scaling and universality. *Front. Comput. Neurosci.*, 9:109, 2015. 1

[8] S. Marzen and J. P. Crutchfield. Statistical signatures of structural organization: The case of long memory in renewal processes. *Phys. Lett. A*, 380(17):1517–1525, 2016. 1

[9] D. P. Varn and J. P. Crutchfield. Chaotic crystallography: How the physics of information reveals structural order in materials. *Curr. Opin. Chem. Eng.*, 7:47–56, 2015. 1

[10] A. Rupe, N. Kumar, V. Epifanov, K. Kashinath, O. Pavlyk, F. Schimbach, M. Patwary, S. Maidanov, V. Lee, Prabhat, and J. P. Crutchfield. Disco: Physics-based unsupervised discovery of coherent structures in spatiotemporal systems. In *2019 IEEE/ACM Workshop on Machine Learning in High Performance Computing Environments (MLHPC)*, pages 75–87, 2019. 1

[11] C. R. Shalizi, K. L. Shalizi, and J. P. Crutchfield. Pattern discovery in time series, Part I: Theory, algorithm, analysis, and convergence. *arXiv.org/abs/cs.LG/0210025*. 1

[12] L. Song, B. Boots, S. Siddiqi, G. Gordon, and A. Smola. Learning and discovery of predictive state representations in dynamical systems with reset. In *Proceedings of the Twenty-first International Conference on Machine Learning*, page 53. ACM, 2004.

[13] S. Still, J. P. Crutchfield, and C. J. Ellison. Optimal causal inference: Estimating stored information and approximating causal architecture. *CHAOS*, 20(3):037111, 2010.

[14] C. C. Strelioff and J. P. Crutchfield. Bayesian structural inference for hidden processes. *Phys. Rev. E*, 89:042119, 2014.

[15] M. Thon and H. Jaeger. Links between multiplicity automata, observable operator models and predictive state representations – a unified learning framework. *J. Mach. Learn. Res.*, 16:103–147, 2015.

[16] N. Brodu and J. P. Crutchfield. Discovering causal structure with reproducing-kernel hilbert space ϵ-machines. *Chaos*, 32:023103, 2022. 1, 2, 9

[17] S. P. Loomis and J. P. Crutchfield. Topology, convergence, and reconstruction of predictive states. 2021. arXiv:2109.09203. 2, 9

[18] C. R. Shalizi and J. P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *J. Stat. Phys.*, 104:817–879, 2001. 2

[19] L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proc. 26th Intl. Conf. Machine Learning*, page 961–968. ACM, 2009. 2, 9

[20] L. Song, B. Boots, S. Siddiqi, G. Gordon, and A. Smola. Hilbert space embeddings of hidden markov models. In *Proc. 27th Intl. Conf. Machine Learning*, pages 991–998. Omnipress, 2010.

[21] B. Boots, A. Gretton, and G. Gordon. Hilbert space embeddings of predictive state representations. In *Proceedings of the 29th International Conference on Uncertainty in Artificial Intelligence*, pages 92–101, 2013. 2, 9

[22] B. Sriperembudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *J. Machine Learn. Res.*, 11:1517–1561, 2010. 2

[23] J. W. Tukey. The future of data analysis. *Ann. Math. Stat.*, 33, 1962. 2

[24] J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, Reading, Massachusetts, 1977. 2

[25] P. Kůrka. *Topological and Symbolic Dynamics*. Société Mathématique de France, Paris, 2003. 2, 3, 4, 9

[26] V. M. Panaretos and Y. Zemel. Statistical aspects of Wasserstein distances. *Annu. Rev. Stat. Appl.*, 6:405–431, 2019. 2, 5

[27] I. Borg and P. J. F. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer Series in Statistics. Springer, New York, NY, second edition, 2005. 2, 6, 7

[28] D. Müllner. Modern hierarchical, agglomerative clustering algorithms. 2011. 1109.2378v1. 2, 6

[29] S. Geman and M. Johnson. Probabilistic grammars and their applications. In *In International Encyclopedia of the Social & Behavioral Sciences. N.J. Smelser and P.B*, pages 12075–12082, 2000. 2

[30] Predictive state geometry via Cantor embeddings and Wasserstein distance. `https://github.com/samlikesphysics/PredStateCantorWass`, 2022. 2, 10

[31] J. P. Crutchfield and D. P. Feldman. Regularities unseen, randomness observed: Levels of entropy convergence. *CHAOS*, 13(1):25–54, 2003. 2

[32] J. E. Hopcroft, R. Motwani, and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Prentice-Hall, New York, third edition, 2006. 2, 3

[33] Y. Sakakibara, M. Brown, R. Hughley, I. S. Mian, K. Sjölander, R. C. Underwood, and D. Haussler. Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Research*, 22(23):5112–5220, 1994. 3

[34] O. Pele and M. Werman. Fast and robust earth-mover's distances. In *2009 IEEE 12th International Conference on Computer Vision*, pages 460–467. IEEE, 2009. 6

[35] O. Thas. *Comparing Distributions*. Springer Series in Statistics. Springer, New York, NY, 2010. 6

[36] J. P. Crutchfield and C. R. Shalizi. Thermodynamic depth of causal states: Objective complexity via minimal representations. *Physical Review E*, 59(1):275–283, 1999. 9