Predictive State Geometry via Cantor Embeddings and Wasserstein Distance

Anonymous Author(s)

ABSTRACT

Predictive states for stochastic processes are a non-parametric 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 has 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 construct a finite-dimensional embedding of sequences based on the Cantor set, and use distances in this embedding to compute Wasserstein distances between distributions over sequences ("predictions"). We show that analyzing the resulting geometry (via hierarchical clustering and dimension reduction) provides insight into the temporal structure of processes ranging from the relatively simple (*e.g.* hidden Markov models) to the very complex (*e.g.* indexed grammars).

CCS CONCEPTS

• **Mathematics of computing** → **Probabilistic representations**; **Nonparametric representations**; **Stochastic processes**; **Time series analysis**; *Markov processes*; *Topology*; *Information theory*; • **Theory of computation** → *Formal languages and automata theory*; *Grammars and context-free languages.*

KEYWORDS

time series, predictive states, wasserstein distance, hierarchical clustering

ACM Reference Format:

Anonymous Author(s). 2022. Predictive State Geometry via Cantor Embeddings and Wasserstein Distance. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'22).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 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 (has a tendency to explore all possible behaviors) [11, 24]. We may wish to forecast from the observed information the behavior of the next n observations. If we suspect that temporal correlations of the process do not matter much beyond k symbols, then we can use the data we have to take all subsequences of length $n + k$, splitting each subsequence into words of length k and n respectively. From these "past/future" pairs we could construct an empirical conditional distribution

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

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

But suppose now that we do not know how long-range the temporal dependencies in our process can stretch. Even very simple stochastic processes can have infinite Markov order, indicating potential long-term dependence of future observations on the past [34]. Given sufficient data, it would be desirable to take pasts of arbitrary length, and converge towards some prediction conditioned on the *infinite* past:

$$P_{x_1 \dots x_n | \overleftarrow{x}} = \lim_{k \to \infty} \hat{P}_{x_1 \dots x_n | x_{-k} \dots x_0} \tag{2}$$

where $\overleftarrow{x}$ denotes an infinite sequence $\overleftarrow{x} = (\dots, x_{-1}, x_0)$ of observations stretching into the past. This mathematical ideal is known to previous literature as the *causal* or *predictive state* [6, 10]. 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 will denote it simply by $P_{\overleftarrow{x}}$.

Predictive states have been utilized for inference and modeling in dynamical systems [4, 7, 22], renewal processes and spike-trains [14, 15], condensed matter physics [35, 36] and spatiotemporal systems [19–21], and a deep mathematical theory of predictive state inference has been correspondingly developed [3, 10, 16, 25, 26, 30, 31, 33]. Despite this, universal conditions for predictive state convergence were not given until recently. It is now known that if the dataset $x_1 \dots x_L$ is drawn from any stationary process, and if L is sufficiently large, then Eq. (2) will converge for any n and a probability-1 subset of pasts $\overleftarrow{x}$ [13]. In the language of measures, $P_{\overleftarrow{x}}$ converges *in distribution*, with respect to the *product topology* of the space of sequences $\mathcal{X}^{\mathbb{N}}$.

The general goals of predictive state analysis are typically threefold [23]. 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. We will focus in this paper on the first, as it is a crucial building block to achieving the other two.

Recent attempts at reconstructing the geometry of predictive states have focused on embedding them in reproducing kernel Hilbert spaces [1, 3, 13, 27, 28], to great effect largely because convergence in Hilbert spaces generated by universal kernels is equivalent to convergence in distribution [29]. That is, these embeddings allow accurate representations of predictive states because they

KDD'22, August 2022, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

respect the product topology of sequences in the same way that predictive states themselves do.

Understanding the source of the power of RKHS methods in predictive state analysis frees us to consider other options. In this paper, we will embed 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 [12]. 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 can be computed with a closed-form integral for one-dimensional distributions. This works as an alternative to RKHS-based distances because the Wasserstein distance also reproduces the topology of convergence in distribution [18]. The resulting distance matrix can be used to find low-dimensional embeddings [2] of the geometry or hierarchical clusterings [17] of the predictive states. The latter, in particular, when combined with our fractal embedding, provides a highly interpretable visualization of the predictive state space.

2 EXAMPLE PROCESSES

The methods in this paper are intended to be applied toward stationary and ergodic stochastic processes which produce categorical time-series data. For our intents we can think of a stochastic process as a collection of probability distributions $\Pr_\mu(x_1 \dots x_L)$ 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 $(\dots, x_{-1}, x_0, x_1, \dots) \in \mathcal{X}^{\mathbb{Z}}$. These sorts of processes can be generated by a number of systems with widely varying complexity. Most popularly studied are those methods often characterized as having some degree of "finite memory": Markov models, hidden Markov models, and observable operator models (also termed generalized hidden Markov models) [10, 34]. Beyond these one can also generate processes using probabilistic grammars, such as probabilistic context-free and indexed grammars [8]. Additionally, coarse-grained data from dynamical systems (such as the logistic map) can display behavior varying widely in complexity [4]. In this paper we will refer back frequently to the following example processes:

(1) 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 for the constraint that 1's can only appear in contiguous blocks of even size. The even process has infinite Markov order but can be generated by a two-state hidden Markov process [5].
(2) The $a^n b^n$ *process* can be generated by choosing a random integer $n \geq 1$ (we will suppose via a Poisson process) and writing n a's followed by an equal number of b's, and then repeating this process indefinitely. The result will be sequences where any contiguous block of a's is followed by a block of b's of equal size. The $\mathsf{a}^n\mathsf{b}^n$ process cannot be modeled by any hidden Markov model, though it is a simple example of a probabilistic context-free grammar [9].
(3) The $x + f(x)$ *process* is a probabilistic context-free grammar modeling simple mathematical expressions. It has terminal symbols $\{(\ ,\ )\ ,\ ;\ ,\ +\ ,\ \mathsf{f}\ ,\ \mathsf{x}\}$ and non-terminals $\{A, B, C\}$, and starts with a sequence osf A's with the production rules:

$$A \mapsto B + C\ ;\ |\ C\ ;$$
$$B \mapsto B + C\ |\ C$$
$$C \mapsto \mathsf{f}(B)\ |\ \mathsf{x}$$

(4) The $a^n b^n c^n$ *process* is a probabilistic indexed grammar [9] which is analagous to $\mathsf{a}^n\mathsf{b}^n$ except after writing the blocks of a's and b's, we also write a block of c's of length n.
(5) The *Feigenbaum process* is generated by sampling from the time series of the logistic map at critical parameter $r \approx 3.56995$:

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

and then coarse-graining the data by taking $x_t = 0$ if $0 < y_t \leq \frac{1}{2}$ and $x_t = 1$ if $\frac{1}{2} < y_t < 1$ [12]. 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 result of the Feigenbaum process is indexed-context free [4].

3 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 $\overrightarrow{x}_L = (x_L, x_{L+1}, \dots)$. 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. The fractal nature of sequence-space is encoded in the structure of its *product topology*.

We can exploit 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 can associate some unique integer between 0 and $|\mathcal{X}| - 1$ inclusive; call this $J(x)$. Then there is a function $C : \mathcal{X}^{\mathbb{N}} \to [0, 1]$ which maps every sequence to a positive real number:

$$C(x_1, x_2, \dots) = \sum_{k=1}^{\infty} \frac{2J(x_k)}{(2|\mathcal{X}| - 1)^k} \tag{3}$$

For instance, suppose that $|\mathcal{X}| = 2$ has two elements; then C maps the sequence to a point the traditional Cantor set fractal. For a finite sequence of length L, truncate the sum at $k = L$.

Remarkably, the embedding C has the property that for any continuous function f on $[0, 1]$, the function $F(\overrightarrow{x}) = f(C(\overrightarrow{x}))$ is continuous on $\mathcal{X}^{\mathbb{N}}$; further, if F is continuous on $\mathcal{X}^{\mathbb{N}}$, then $f(y) = F(C^{-1}(y))$ is continuous on the image. Thus the embedding C respects the basic structure of the product topology [12].

Stationary processes, due to their time-translation invariance, inherit the fractal temporality of sequence space. This can be neatly visualized: given a length-L sample $x_1 \dots x_L$, and some $n, k > 0$, take a sliding window of pasts and futures, $(x_{t-n+1} \dots x_t, x_{t+1} \dots x_k)$ for $t = n, \dots, L - k$. For each past-future pair, compute the truncated Cantor embeddings on the *reversed* past and (unreversed) future: $(C(x_t \dots x_{t-n+1}), C(x_{t+1} \dots x_k))$. The resulting pairs of real numbers can be plotted as (x, y)-values on a scatter plot. The fractal which emerges contains, in essence, all information necessary to understand the temporal structures of the process. See Fig. 1 for examples and guidance on how to interpret the visualization.

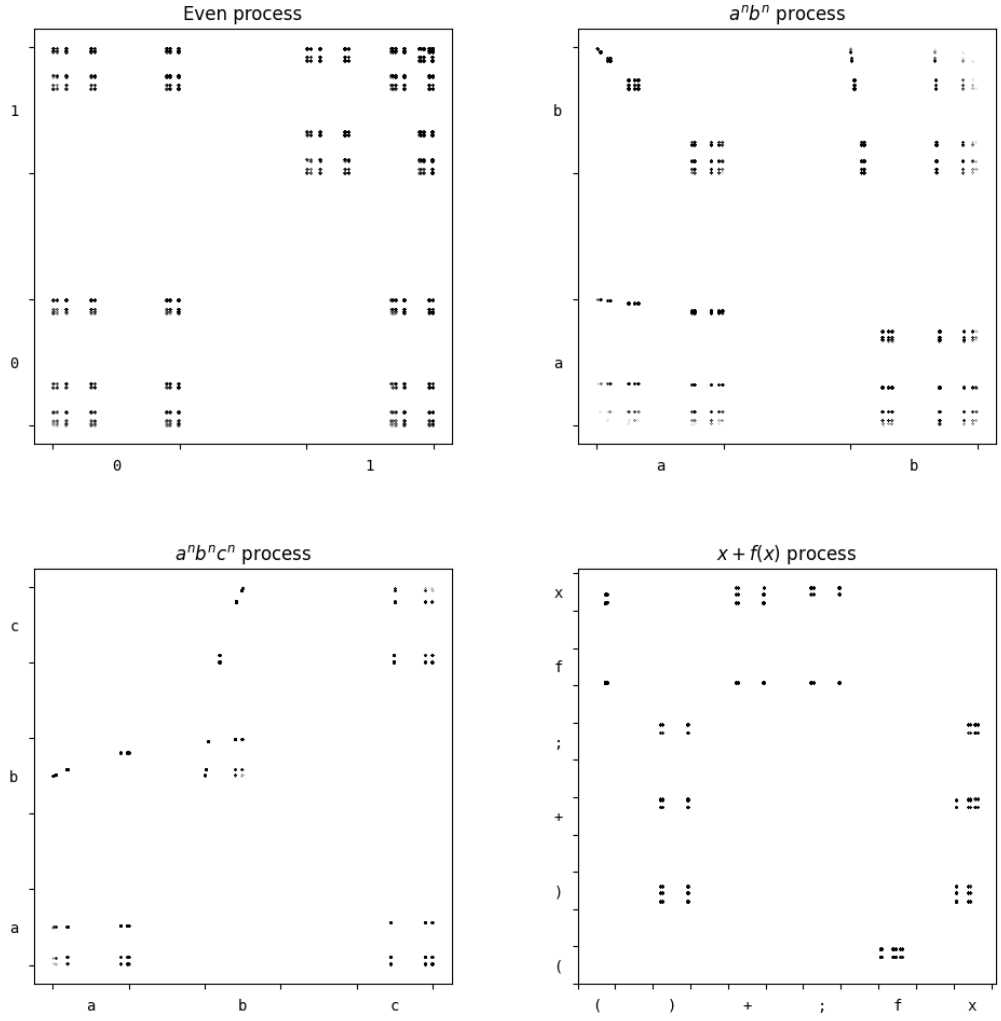

Figure 1: Cantor plots for the even, $\mathsf{a}^n\mathsf{b}^n$, $\mathsf{a}^n\mathsf{b}^n\mathsf{c}^n$, and $\mathsf{x}+\mathsf{f}(\mathsf{x})$ processes. Each point (x, y) corresponds to a pair of sequences corresponding to the past and future respectively. The symbol on the x (y) axis indicate that 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 would indicate the value of the second, third, *etc.* symbols. For instance, one can read from the $\mathsf{x}+\mathsf{f}(\mathsf{x})$ fractal that any past which ends in f must be paired with a future beginning in $(\mathsf{f}$ or $(\mathsf{x}$.

It is worth noting that for $|\mathcal{X}| > 2$, the embedding C also introduces additional structure which may or may not be desirable. The association of each symbol x with an integer j_x endows an ordinal structure on the set $\mathcal{X}$; this ordinality is present in the macroscopic geometry of $C\left(\mathcal{X}^{\mathbb{N}}\right)$. Later we will examine higher-dimension embeddings which do not have this ordinal aspect; however, they will come at the cost of increased computational complexity for the Wasserstein distance computation, so for now we will assume that ordinal artifacts are either desired or sufficiently tolerable as to not outweigh the computational benefits of working in one dimension.

4 WASSERSTEIN DISTANCE ON PREDICTIVE STATES

The Cantor fractals in Fig. 1 can be thought of as probability distributions. We can consequently interpret a vertical slices 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 analyzing the Cantor fractal of the even process, one may notice 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 which generates the even process.

Algorithm 1: Steps for converting a sequence of categorical time-series data into a labeled collection of empirical distributions of Cantor-embedded futures, and a matrix of Wasserstein distances from said distributions.

function CantorWasserstein $(n, k, x_1 \dots x_L)$;
input : Integers n, k 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**
  $\overleftarrow{\mathbf{x}} \leftarrow [x_{t-n+1}, \dots, x_t]$;
  $\overrightarrow{\mathbf{x}} \leftarrow [x_{t+1}, \dots, x_{t+k}]$;
  $p \leftarrow \sum_{\ell=1}^{L} \frac{2J(\overrightarrow{x}_\ell)}{(2|\mathcal{X}|-1)^\ell}$;
  if $\overleftarrow{\mathbf{x}} \in$ UnqPasts **then**
    append $\overleftarrow{\mathbf{x}}$ **to** UnqPasts;
    append $[p]$ **to** Cantors;
  else
    $j \leftarrow$ index $\left(\overleftarrow{\mathbf{x}}, \text{UnqPasts}\right)$;
    append p **to** Cantors$_j$;
  end
end
$K \leftarrow$ length(UnqPasts);
Wass $\leftarrow$ Matrix(K, K);
for $k \leftarrow 1$ **to** K **do**
  for $\ell \leftarrow 1$ **to** K **do**
    Wass$_{k\ell} \leftarrow$ Wasserstein(Cantors$_k$, Cantors$_\ell$);
    Wass$_{\ell k} \leftarrow$ Wass$_{k\ell}$;
  end
end
Result: UnqPasts,Cantors,Wass

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

The intuitive distance metric between probability measures for capturing this underlying geometry is the Wasserstein metric [18]. Given two measures μ, 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. This can be interpreted as the minimal cost for "shifting" the probability mass from one distribution to match shape of the other.

$W(\mu, \nu)$ the solution to a constrained linear optimization. As a function of distributions, $W(\mu, \nu)$ is continuous with respect to convergence in distribution; in fact, convergence under the Wasserstein distance is equivalent to convergence in distribution on compact spaces. This makes $W(\mu, \nu)$ ideal for measuring geometry between predictive states, since empirical estimates of these are known to converge in distribution.

When $\mathcal{M} \subseteq \mathbb{R}$, there is in fact a closed-form solution to the Wasserstein optimization problem [32]. Let F and G respectively be the cumulative distributions functions of μ and ν. Then

$$W(\mu, \nu) = \int_{-\infty}^{\infty} |F(t) - G(t)| \, dt \tag{4}$$

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

The combination of the Cantor embedding and the Wasserstein distance lays out a fairly straightforward programme for analyizing a categorical time series:

(1) Apply Algorithm 1 to the data stream $x_1 \dots x_L$ for a specified past length k and future length n, retrieving the set of unique observed pasts, the empirical Cantor distributions corresponding to each past, and the matrix of Wasserstein distances computed from these distributions.
(2) Use the Wasserstein distance matrix to perform additional methods of geometric data analysis, such as hierarchical clustering [17] and multidimensional scaling [2], in order to elucidate the relative geometry of the predictive states.

In the next two sections we examine the results of this approach.

5 INTERPRETABLE PREDICTIVE STATES WITH HIERARCHICAL CLUSTERING

In Fig. 2 we display the result of collecting the Cantor-embedded empirical predictions for all pasts of a given length, for four processes (even, $\mathsf{a}^n\mathsf{b}^n$, $\mathsf{a}^n\mathsf{b}^n\mathsf{c}^n$ and $\mathsf{x} + \mathsf{f}(\mathsf{x})$). In each case the Wasserstein distance between every pair of predictions was computed and used to hierarchically cluster the pasts (using the Ward method [17]) with others that produced similar predictions.

The resulting clustered Cantor plots offer a highly interpretable visualization of the relationship between pasts and futures, and of the geometry of the predictive states. They are, in a certain sense, a sorting of the columns in the Cantor fractals of Fig. 1, with the whitespace between columns removed. For instance, the clustered Cantor plot of the Even process clearly contains the two major states, with a third "transient" state visible (corresponding 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 remarking on. Close observation will show the reader that the hierarchical clustering allows for the

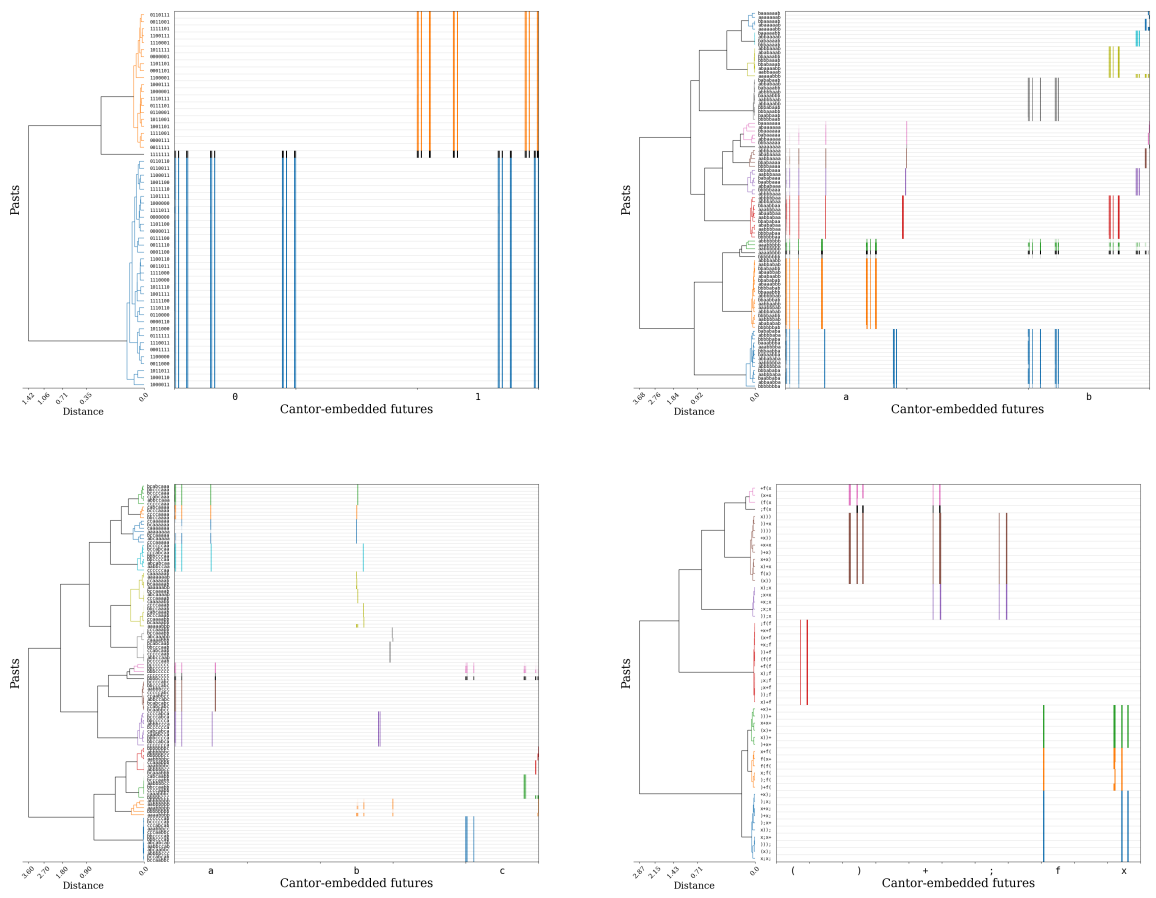

Figure 2: From upper left to lower right, the clustered Cantor diagrams of the even, $a^n b^n$, $a^n b^n c^n$, and $x + f(x)$ processes (zoom for detail). On each diagram, the vertical axis shows all pasts of a given length (length $k = 8$ for the even, $a^n b^n$ and $a^n b^n c^n$ processes and $k = 4$ for the $x + f(x)$ process), along with their hierarchically clustered dendrogram. For the purposes of this graphic the coloring threshold was chosen to aid in 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.

(mostly) scale-free distinctions between pasts with subtle differences. For the $a^n b^n$ process, pasts of the form $\ldots ba^n$ are distinguished for different n, as each involves a distinct number of b's appearing in the near future; meanwhile, the clustering scheme carefully distinguishes pasts of the form $\ldots ba^n b^{n-k}$ for different k but *not* for different n, as k is the essential variable for predicting the remaining number of b's. (The scale-free discernment of the algorithm begins to break down past $n = 5$, which is the scale at which sampling error becomes relevant for our chosen sample size.)

Similar discernment can be noticed for the $a^n b^n c^n$ and $x + f(x)$ processes as well. We draw attention to the manner in which the presence of a semicolon in pasts from $x + f(x)$ affects the comparison of predictions.

By analyzing clustered Cantor plots, one can gain 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 natural geometry of the Cantor set for visualizing the future forecasts associated with each cluster of predictions.

6 PREDICTIVE STATE GEOMETRY WITH MULTIDIMENSIONAL SCALING

If we are willing to sacrifice direct visualization of future predictions, we may gain a more intuitive picture of the geometry of predictive state space. We can apply any desired dimensional reduction algorithm to the matrix of Wasserstein distances between predictions and obtain a coordinate representation of the similarities between predictive states.

In Fig. 3 we have plotted the first two dimensions of a multidimensional scaling (MDS) decomposition [2] for the even process

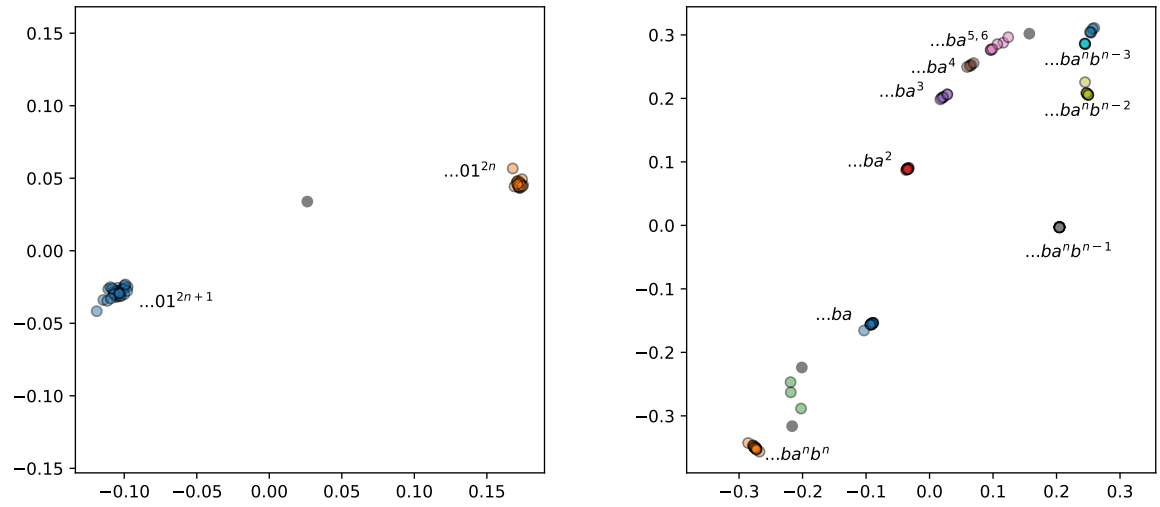

Figure 3: Two scatterplots of the first two MDS coordinates of the reconstructed predictive states, for the even and $a^n b^n$ processes respectively. The clusters are colored according to the scheme determined by the dendrogram in Fig. 2 and the label on each cluster describes the pattern which uniquely characterizes the pasts in that cluster.

and the $a^n b^n$ process, with clusters colored in the same manner as in Fig. 2 and labeled by the specific pattern which distinguishes the pasts in some of the clusters. It should be noted that the clusters and the labels are directly drawn from 2 for reference, and are not a 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 we have seen thus far, has two dominant clusters of predictions, corresponding to the predictive states which result from seeing an even-sized block of 1's (or, equivalently, no 1's), and that resulting from seeing an odd-sized block of 1's. The $a^n b^n$ plot is much more sophisticated. Intriguingly, its geometry not only clearly distinguishes predictively distinct states, but it organizes them in a manner highly suggestive of an *stack*, particularly appropriate given that stack automata are the natural analogue of hidden Markov models for context-free processes. As more a's are observed, we push further up the stack towards the upper-right corner of the figure, and as more b's are observed we pop out of the stack, back towards the lower left (representing equality between a's and b's).

The geometric approach is particularly insightful when we compute the Wasserstein matrix between predictions estimated from Feigenbaum process data. Recall that the Feigenbaum process is just a coarse-graining of the iterated logistic map $y_{t+1} = r y_t (1 - y_t)$ at the critical parameter $r \approx 3.56995$. The resulting stream of 0's and 1's is an infamous instance of high complexity at the "boundary of order and chaos", as values of r on either side tend to result in coarse-grainings which can be generated by hidden Markov models, but the Feigenbaum process is context-sensitive and therefores requires several orders higher of model complexity to capture its behavior.

Regardless, we show that the predictive state geometry reconstructed from a sufficiently large sample of the Feigenbaum process is capable of recovering the neighborhoods of $[0, 1]$ which are relevant to the dynamics of the original logistic map. What we mean by this is that there is a correspondence between each past $x_{-n+1} \ldots x_0$ and a subset $V_{x_{-n+1} \ldots x_0}$, such that $V_{x_1 \ldots x_n}$ 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 $x_{-n+1} \ldots x_0$ whose predictive states are close under the Wasserstein distance are also pasts for which the sets $f(V_{x_{-n+1} \ldots x_0})$ are close (that is, they correspond to predictively similar ranges of the logistic map variable).

This relationship between the reconstructed predictive states of the Feigenbaum process, neighborhoods of the logistic variable y, and the logistic map dynamics is visualized in Fig. 4. In short, despite the fact that the Feigenbaum 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.

We should note that, because of the deterministic nature of the Feigenbaum 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 no asymptotically no overlap between predictions. The choice of the Cantor map allows us to place close together forecasts which match up to a certain time in the future, and the Wasserstein distance allows us to directly compare predictions whose supports are geometrically close. It is therefore the combination of these

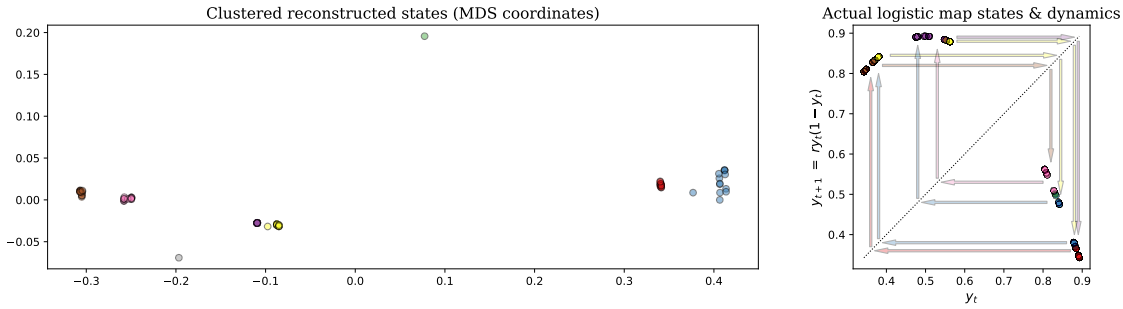

Figure 4: On the left is a scatterplot of the first two MDS coordinates of the reconstructed predictive states for the Feigenbaum process, color-coded by cluster. On the right, we display a scatterplot of the corresponding points in the domain of the logistic map, plotting for each point both the present 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.

two approaches which enables the straightforward recovery of the underlying logistic geometry.

7 CONCLUDING REMARKS

We have presented a general approach for predictive state analysis—Cantor fractal embedding sequences and Wasserstein distance comparison of predictions—and 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.

In comparison to the use of reproducing kernel Hilbert spaces, which is dominant approach to predictive states at present [1, 3, 13, 27, 28], our choice to combine the Cantor set with the Wasserstein distance may appear rather idiosyncratic, but there are strong benefits to both methods, 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, but the Wasserstein distance depends on far fewer parameters (such as the choice of the eponymous kernel in RKHS approaches), and 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 by 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 the clustered Cantor diagrams of Fig. 2.

8 REPRODUCIBILITY STATEMENT

For the purposes of reproducibility, we have provided a GitHub repository which contains the code necessary to generate this paper and its figures, including a notebook for generating the data we used for our examples and building both static and interactive figures for further exploration.

REFERENCES

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

[2] I. Borg and P. J. F. Groenen. 2005. *Modern Multidimensional Scaling: Theory and Applications* (second ed.). Springer, New York, NY.

[3] N. Brodu and J. P. Crutchfield. 2020. Discovering Causal Structure with Reproducing-Kernel Hilbert Space ϵ-Machines. *arXiv:2011.14821* (2020).

[4] J. P. Crutchfield. 1994. The Calculi of Emergence: Computation, Dynamics, and Induction. *Physica D* 75 (1994), 11–54.

[5] J. P. Crutchfield and D. P. Feldman. 2003. Regularities Unseen, Randomness Observed: Levels of Entropy Convergence. *CHAOS* 13, 1 (2003), 25–54.

[6] J. P. Crutchfield and K. Young. 1989. Inferring Statistical Complexity. *Phys. Rev. Let.* 63 (1989), 105–108. https://doi.org/10.1103/PhysRevLett.63.105

[7] J. Emenheiser, A. Chapman, M. Posfai, J. P. Crutchfield, M. Mesbahi, and R. M. D'Souza. 2016. Patterns of patterns of synchronization: Noise induced attractor switching in rings of coupled nonlinear oscillators. *Chaos* 26, 9 (2016), 094816. https://doi.org/10.1063/1.4960191

[8] S. Geman and M. Johnson. 2000. Probabilistic Grammars and their Applications. In *In International Encyclopedia of the Social & Behavioral Sciences. N.J. Smelser and P.B.* 12075–12082.

[9] J. E. Hopcroft, R. Motwani, and J. D. Ullman. 2006. *Introduction to Automata Theory, Languages, and Computation* (third ed.). Prentice-Hall, New York.

[10] H. Jaeger. 2000. Observable Operator Models for Discrete Stochastic Time Series. *Neural Computation* 12, 6 (2000), 1371–1398. https://doi.org/10.1162/089976600300015411

[11] O. Kallenberg. 2001. *Foundations of Modern Probability* (2 ed.). Springer, New York.

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

[13] S. P. Loomis and J. P. Crutchfield. 2021. Topology, Convergence, and Reconstruction of Predictive States. (2021). arXiv:2109.09203.

[14] S. Marzen and J. P. Crutchfield. 2016. Statistical Signatures of Structural Organization: The case of long memory in renewal processes. *Phys. Lett. A* 380, 17 (2016), 1517–1525. https://doi.org/10.1016/j.physleta.2016.02.052

[15] S. Marzen, M. R. DeWeese, and J. P. Crutchfield. 2015. Time Resolution Dependence of Information Measures for Spiking Neurons: Scaling and Universality. *Front. Comput. Neurosci.* 9 (2015), 109. https://doi.org/10.3389/fncom.2015.00105

[16] S. E. Marzen and J. P. Crutchfield. 2019. Probabilistic Deterministic Finite Automata and Recurrent Networks, Revisited. (2019). arxiv.org:1910.07663 [cs.LG].

[17] D. Müllner. 2011. Modern hierarchical, agglomerative clustering algorithms. (2011). 1109.2378v1.

[18] V. M. Panaretos and Y. Zemel. 2019. Statistical Aspects of Wasserstein Distances. *Annu. Rev. Stat. Appl.* 6 (2019), 405–431.

[19] A. Rupe and J. P. Crutchfield. 2020. Spacetime Autoencoders Using Local Causal States. *AAAI Fall Series 2020 Symposium on Physics-guided AI for Accelerating Scientific Discovery* (2020). arXiv:2010.05451.

[20] A. Rupe, K. Kashinath, N. Kumar, V. Lee, Prabhat, and J. P. Crutchfield. [n. d.]. Towards Unsupervised Segmentation of Extreme Weather Events. *arxiv:1909.07520* ([n. d.]).

[21] A. Rupe, N. Kumar, V. Epifanov, K. Kashinath, O. Pavlyk, F. Schimbach, M. Patwary, S. Maidanov, V. Lee, Prabhat, and J. P. Crutchfield. 2019. 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)*. 75–87. https://doi.org/10.1109/MLHPC49564.2019.00013

[22] A. Salova, J. Emenheiser, J. P. Crutchfield, and R. M. D'Souza. 2019. Koopman Operator and its Approximations for Systems with Symmetries. *Chaos* 29 (2019), 093128. https://doi.org/10.1063/1.5099091

[23] C. R. Shalizi. 2001. *Causal Architecture, Complexity and Self-Organization in Time Series and Cellular Automata.* Ph. D. Dissertation. University of Wisconsin, Madison, Wisconsin.

[24] C. R. Shalizi and A. Kontorovich. 2013. Almost None of the Theory of Stochastic Processes. Lecture notes.

[25] C. R. Shalizi, K. L. Shalizi, and J. P. Crutchfield. [n. d.]. Pattern Discovery in Time Series, Part I: Theory, Algorithm, Analysis, and Convergence. *arXiv.org/abs/cs.LG/0210025* ([n. d.]).

[26] L. Song, B. Boots, S. Siddiqi, G. Gordon, and A. Smola. 2004. Learning and Discovery of Predictive State Representations in Dynamical Systems with Reset. In *Proceedings of the Twenty-first International Conference on Machine Learning*. ACM, 53.

[27] L. Song, B. Boots, S. Siddiqi, G. Gordon, and A. Smola. 2010. Hilbert space embeddings of hidden Markov models. In *Proceedings of the 27th International Conference on Machine Learning*. Omnipress, 991–998.

[28] L. Song, J. Huang, A. Smola, and K. Fukumizu. 2009. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th International Conference on Machine Learning*. ACM, 961–968.

[29] B. Sriperembudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. G. Lanckriet. 2010. Hilbert Space Embeddings and Metrics on Probability Measures. *J. Machine Learn. Res.* 11 (2010), 1517–1561.

[30] S. Still, J. P. Crutchfield, and C. J. Ellison. 2010. Optimal Causal Inference: Estimating Stored Information and Approximating Causal Architecture. *CHAOS* 20, 3 (2010), 037111.

[31] C. C. Strelioff and J. P. Crutchfield. 2014. Bayesian Structural Inference for Hidden Processes. *Phys. Rev. E* 89 (2014), 042119. https://doi.org/10.1103/PhysRevE.89.042119

[32] O. Thas. 2010. *Comparing Distributions.* Springer, New York, NY.

[33] M. Thon and H. Jaeger. 2015. Links Between Multiplicity Automata, Observable Operator Models and Predictive State Representations – a Unified Learning Framework. *J. Mach. Learn. Res.* 16 (2015), 103–147.

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

[35] D. P. Varn, G. S. Canright, and J. P. Crutchfield. 2013. ϵ-Machine spectral reconstruction theory: A direct method for inferring planar disorder and structure from X-ray diffraction studies. *Acta. Cryst. Sec. A* 69, 2 (2013), 197–206.

[36] D. P. Varn and J. P. Crutchfield. 2015. Chaotic Crystallography: How the physics of information reveals structural order in materials. *Curr. Opin. Chem. Eng.* 7 (2015), 47–56. https://doi.org/10.1016/j.coche.2014.11.002