Santa Fe Institute Working Paper 06-11-042
arXiv:cs/0611054 [cs.LG]

Optimal Instruments and Models for Noisy Chaos

Christopher C. Strelioff[1, 2, *] and James P. Crutchfield[1, †]

[1] *Center for Computational Science & Engineering and Physics Department, University of California at Davis, One Shields Avenue, Davis, CA 95616*
[2] *Center for Complex Systems Research and Physics Department, University of Illinois at Urbana-Champaign, 1110 West Green Street, Urbana, Illinois 61801*

(Dated: September 9, 2007)

Analysis of finite, noisy time series data leads necessarily to modern statistical inference methods. Here we adapt Bayesian inference for applied symbolic dynamics. We show that reconciling Kolmogorov's maximum-entropy partition with the methods of Bayesian model selection requires the use of two separate optimizations. First, instrument design produces a maximum-entropy symbolic representation of time series data. Second, Bayesian model comparison with a uniform prior selects a minimum-entropy model, with respect to the considered Markov chain orders, of the symbolic data. We illustrate these steps using a binary partition of time series data from the logistic and Hénon maps as well as the Rössler and Lorenz attractors with dynamical noise. In each case we demonstrate the inference of effectively generating partitions and kth-order Markov chain models.

PACS numbers: 05.45.Ac,02.50.Tt,05.45.Tp

Continuous-valued time series data from a low-dimensional deterministic chaotic attractor are commonly transformed into a symbolic sequence. If done correctly, the resulting discrete sequence captures the behavior of the underlying dynamical system and one can obtain, for example, accurate estimates of its fundamental invariants. However, if the projection is not done with care, the resulting sequence can substantially misrepresent the true dynamics, making them appear either simpler or more complicated than they are. We consider the effects of dynamical noise and finite data samples. We show how to find appropriate symbolic representations and estimate the observed randomness due to both deterministic and stochastic elements. We break this problem into two parts: (i) instrument design and (ii) model inference. Instrument design, the first step, attempts to minimize the distortions that arise from projecting continuous dynamics onto a finite alphabet. It has been known for some time that an instrument should be designed so that the symbol sequences it produces have maximum entropy rate, thereby extracting the most information with each measurement. Here we develop a Bayesian method for optimal instrument design. We use model inference, the second step, to estimate an optimal kth-order Markov chain from the resulting symbolic data. We develop Bayesian model comparison to select a minimum entropy-rate model in order to find regularity in the symbol sequence. Our central conclusion is that the two separate optimizations—instrument design and model inference—must be done together to avoid misleading models and unnecessary errors in estimating system properties.

I. INTRODUCTION

Research on chaotic dynamical systems during the last forty years has produced a new vision of the origins of randomness [1]. It is now widely understood that observed randomness can be generated by low-dimensional deterministic systems that exhibit a chaotic attractor [2]. In addition, additive noise and finite data samples can add to the real or apparent randomness [3, 4]. Today, when confronted with what appears to be a high-dimensional stochastic process, one now asks whether or not the process is instead a hidden low-dimensional, but nonlinear dynamical system.

Previous research on analyzing time series from systems that display deterministic or noisy chaos focused on estimating maps, systems of ordinary differential equations (ODEs), or local (non)linear models of the dynamics [5–8]. Often, these methods are quite effective and can be adapted to address dynamical noise in their modeling of the dynamics. However, many do not take direct advantage of modern mathematical statistics methods, such as maximum likelihood or Bayesian estimation. This is a distinct disadvantage since, without such methods, estimates of uncertainty in inference and in model comparison are generally unavailable. These concerns motivate our investigation of an alternative method for time series from nonlinear dynamical systems.

Symbolic dynamics, as one of a suite of tools in dynamical systems theory, considers a coarse-grained view of a continuous dynamics [9]. Transforming continuous-state time series into a symbol sequence opens the door to

*Electronic address: streliof@uiuc.edu
†Electronic address: chaos@cse.ucdavis.edu

applying well developed methods for discrete-valued random processes from statistical physics, information theory, and mathematical statistics. However, converting from a continuous state space to a finite set of symbols must be done with great care since chaotic dynamics are sensitive to the measurement process [10]. As it turns out, this sensitivity is both a blessing and a curse.

Previous developments of symbolic dynamics concentrated on deterministic chaos without additive noise and assumed knowledge of the dynamical system at hand; see, for example, Ref. [5, 9] and references therein. Recent work has attempted to extend the successful analysis of one-dimensional maps to higher-dimensional systems. Examples include analysis of the Hénon map [11] and a coupled map lattice [12]. In addition, application of symbolic dynamics methods to ordinary differential equations have been considered in [9, 13, 14].

Current research in applied symbolic dynamics techniques focuses on extracting information from available time series data. It is assumed the analyst has little or no knowledge of the dynamical system under investigation and is confronted with effects from finite data samples and additive or observational noise [15]. Under these conditions, the main results of symbolic dynamics are not strictly valid. As a result, a variety of methods for finding an approximate *generating partition* have been developed [16–19]. In addition, methods for estimating topological and metric entropies from symbolic data have been pursued as a means to identify effective coarse-grainings [3, 20–22]. We note that recent work on estimating entropies from finite data samples of a more general nature should be considered [23–26] in addition to the references above.

Here we address the interplay between optimal instruments and optimal models by analyzing a set of relatively simple nonlinear systems, though the principles and procedures generalize to more complex systems. To effectively model time series of discrete data from a continuous-state system two concerns must be addressed. First, we must consider the measuring instrument and the representation of the true dynamics which it provides. In the process of instrument design we consider the effect of projecting a continuous state space onto a finite set of disjoint regions—a model of measurement with finite resolution. Second, we must consider the inference of models based on this data. The relation between these steps is more subtle than one might expect. As we will demonstrate, on the one hand, in the measurement of chaotic data, the instrument should be designed to maximize the entropy rate of the resulting symbolic data stream. This allows one to extract as much information from each measurement as possible. On the other hand, model inference strives to minimize the apparent randomness (entropy rate) over a class of alternative models. This reflects a search for determinism and structure in the symbolic data. These criteria are clearly at odds, and so they must be implemented with care.

To address these issues Sec. II considers the design and evaluation of instruments for chaotic maps with additive noise. This is predominantly a review of previously developed methods for symbolic dynamics of noise-free chaotic maps. However, it provides guiding principles for the noisy and finite data-sample setting we will consider. In Sec. III we describe Bayesian inference of a kth-order Markov chain to model the resulting symbolic data stream. We provide sufficient detail for the reader to follow the principles under investigation. However, for a more detailed development of the ideas presented in the section the reader should consult [27]. In Sec. IV we describe experiments using the logistic and Hénon maps as well as the Rössler and Lorenz attractors with additive noise. The simplicity of these examples serves to highlight several technical issues and trade-offs in implementing the multi-level Bayesian modeling scheme that we advocate. Finally, in Sec. V we discuss the consequences of our results, drawing general lessons for modeling more complex systems.

II. INSTRUMENT DESIGN

A. Symbolic dynamics

Our model system is a one-dimensional chaotic map with additive noise [3]

$$x_{t+1} = f(x_t) + \xi_t \ , \tag{1}$$

where $t = 0, 1, 2, \ldots$, $x_t \in [0, 1]$, and ξ_t is an independent identically distributed (IID) random variable with uniform density on the interval $[-\epsilon/2, \epsilon/2]$. We will refer to this as a *noise level* of ϵ. To start, we consider the design of instruments in the zero-noise limit. This is the regime of most previous work in symbolic dynamics and provides a convenient initial frame of reference.

The construction of a symbolic dynamics representation of a continuous-state system goes as follows [9]. We assume time is discrete and consider a map f from the *state space* M to itself $f : M \to M$. This space can be partitioned into a finite set $\mathcal{P} = \{I_i : \cup_i I_i = M; I_i \cap I_j = \emptyset, i \neq j\}$ of nonoverlapping regions in many ways. The most powerful is called a *Markov partition* and must satisfy two conditions. First, the image of each region I_i must be a union of partition elements: $f(I_i) = \cup_j I_j, \forall\, i$. Second, the map $f(I_i)$, restricted to a partition element, must be one-to-one and onto. If a Markov partition cannot be found for the system under consideration, the next best coarse-graining is called a *generating partition*. For one-dimensional maps, these are often easily found using the extrema of $f(x)$—its *critical points*. The critical points in the map are used to divide the state space into intervals I_i over which f is monotone. Note that Markov partitions are generating, but the converse is not generally true.

Given any partition $\mathcal{P} = \{I_i : i = 0, 1, \ldots, K-1\}$, then, a series $\mathbf{X} = x_0 x_1 \ldots x_{N-1}$ of continuous-valued

states can be projected onto its symbolic representation $\mathbf{S} = s_0 s_1 \ldots s_{N-1}$. The latter is simply the associated sequence of partition-element indices. This is done by defining an operator $\pi(x_t) = s_t$ that returns a unique symbol $s_t = i$ for each I_i from an alphabet $\mathcal{A} = \{0, 1, \ldots, K-1\}$ when $x_t \in I_i$.

The central result in symbolic dynamics establishes that, using a generating partition, increasingly long sequences of observed symbols identify smaller and smaller regions of the state space. Starting the system in such a region produces the associated measurement symbol sequence. In the limit of infinite symbol sequences, the result is a discrete-symbol representation of a continuous-state system—a representation that, as we will show, is often much easier to analyze. In this way a chosen partition creates a symbol sequence $\pi(\mathbf{X}) = \mathbf{S}$ which describes the continuous-valued dynamics. The choice of partition is equivalent to our instrument-design problem.

B. Evaluating an instrument

The effectiveness of a partition (in the zero-noise limit) can be quantified by estimating the entropy rate of the resulting symbolic sequence. To do this we consider length-L *words* $\mathbf{s}^L = s_t s_{t+1} \ldots s_{t+L-1}$. The *block entropy* of length-L sequences obtained from partition $\mathcal{P}$ is then

$$H_L(\mathcal{P}) = -\sum_{\mathbf{s}^L \in \mathcal{A}^L} p(\mathbf{s}^L) \log_2 p(\mathbf{s}^L) \ , \tag{2}$$

where $p(\mathbf{s}^L)$ is the probability of observing the word $\mathbf{s}^L \in \mathcal{A}^L$. From the block entropy the *entropy rate* can be estimated as the following limit

$$h_\mu(\mathcal{P}) = \lim_{L \to \infty} \frac{H_L(\mathcal{P})}{L} \ . \tag{3}$$

In practice, it is often more accurate to calculate the length-L estimate of the entropy rate using

$$h_{\mu L}(\mathcal{P}) = H_L(\mathcal{P}) - H_{L-1}(\mathcal{P}) \ . \tag{4}$$

Another key result in symbolic dynamics says that the entropy of the original continuous system is found using generating partitions [28, 29]. In particular, the true entropy rate $h_\mu(f)$ maximizes the estimated entropy rates:

$$h_\mu(f) = \max_{\{\mathcal{P}\}} h_\mu(\mathcal{P}) \ . \tag{5}$$

Thus, translated into a statement about experiment design, this result tells us to design an instrument so that it maximizes the observed entropy rate $h_{\mu L}(\mathcal{P})$. This reflects the fact that we want each measurement to produce the most information possible about the behavior of the measured system.

As a useful benchmark on this, useful only in the case when we know $f(x)$, *Pesin's Identity* [30] tells us that the value of $h_\mu(f)$ is equal to the sum of the positive Lyapunov characteristic exponents:

$$h_\mu(f) = \sum_i \lambda_i^+ \ , \tag{6}$$

where λ_i^+ indicate positive exponents. For 1-d maps there is a single Lyapunov exponent λ which is numerically estimated from the map f and a trajectory $\{x_t : t = 1, \ldots, N\}$ using

$$\lambda = \lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} \log_2 |f'(x_t)| \ . \tag{7}$$

A related set of equations for multi-dimensional maps and flows can be obtained using the Jacobian matrix and time series data; see, for example, Ref. citeKantz06a.

Taken altogether, these results tell us how to design our instrument for effective observation of deterministic chaos. Notably, in the presence of noise no such theorems exist. However, Ref. [3] demonstrated that the methods developed above are robust in the presence of noise. However, we expect that the equality in Eq. (6) no longer applies when the addition of dynamical noise is considered. For noisy chaos, the entropy rate should in general be greater than the sum of the Lyapunov exponents estimated from the noise-free system. This is not incorrect, dynamical noise should increase the entropy rate because it affects the dynamics of the system of interest.

In any case, we view the output of the instrument as a stochastic process. A sample realization D of length N with measurements taken from a finite alphabet is the basis for our inference problem: $D = s_0 s_1 \ldots s_{N-1}$, $s_t \in \mathcal{A}$. For our purposes here, the sample is generated by a partition of continuous-state sequences from iterations of a low-dimensional map or flow on a chaotic attractor. This means, in particular, that the resulting discrete-valued stochastic process is stationary. We assume, in addition, that the alphabet is binary $\mathcal{A} = \{0, 1\}$. This assumption is motivated by our application to a variety of systems which display return maps that can be approximated by single peak, one-dimensional maps. A partition divider located at the critical point of these maps produces the assumed binary alphabet and results in a compact coarse-graining of the data. In principle, larger alphabets could be considered, but, at fixed data length N, this affects the inference process by creating fewer samples of larger-alphabet sequences without gaining any new information. As a result, we choose the simple binary alphabet.

III. MODELING SYMBOLIC DATA

A. Bayesian inference of kth-order Markov chains

Given a method for instrument design the next step is to estimate a model from the observed measurements.

Here we employ the model class of kth-order Markov chains and Bayesian inference as the model estimation and selection paradigm following our development in Ref. [27].

The kth-order Markov chain model class makes two strong assumptions about the data sample. The first is an assumption of finite memory. In other words, the probability of s_t depends only on some previous k symbols in the data sample. We introduce the more compact notation $\overleftarrow{s}^k_t = s_{t-k+1} \ldots s_t$ to indicate a length-k sequence of measurements ending at time t. The finite memory assumption is then equivalent to saying the probability of the observed data can be factored into the product of terms with the form $p(s_{t+1}|\overleftarrow{s}^k_t)$. The second assumption is stationarity. This means the probability of observed sequences does not change with position in the data sample: $p(s_{t+1}|\overleftarrow{s}^k_t) = p(s|\overleftarrow{s}^k)$ for any index t. As noted above, this assumption is satisfied by the data streams considered here. The first assumption, however, is often not true of chaotic systems. They can generate time series with infinitely long temporal correlations. Thus, in some cases, we may be confronted with *out-of-class* modeling. Note, however, that in the case that the instrument defines a Markov partition, the Markov chain model class is exactly appropriate.

A kth-order Markov chain model $\mathbf{M}_k$ has a set of parameters

$$\theta_k = \{p(s|\overleftarrow{s}^k) : s \in \mathcal{A}, \overleftarrow{s}^k \in \mathcal{A}^k\} \ . \tag{8}$$

The notation θ is conventional in mathematical statistics for all model parameters that are to be estimated. For example, a $k = 1$ Markov chain has parameters $\theta_1 = \{p(0|0), p(1|0), p(0|1), p(1|1)\}$. Consistent with this, for integrals we write $d\theta_1 = dp(0|0)\, dp(1|0)\, dp(0|1)\, dp(1|1)$. A similar pattern holds for larger Markov chain order k.

In Bayesian inference we connect the model parameters θ_k for model $\mathbf{M}_k$ with the observed data D. To do this, we write down the *likelihood* $P(D|\theta_k, \mathbf{M}_k)$ and the *prior* $P(\theta_k|\mathbf{M}_k)$ and then calculate the *evidence* $P(D|\mathbf{M}_k)$. The *posterior distribution* $P(\theta_k|D, \mathbf{M}_k)$ is obtained from these using Bayes' theorem:

$$P\left(\theta_k|D, \mathbf{M}_k\right) = \frac{P\left(D|\theta_k, \mathbf{M}_k\right)\ P\left(\theta_k|\mathbf{M}_k\right)}{P\left(D|\mathbf{M}_k\right)} \ . \tag{9}$$

The posterior describes the distribution of model parameters θ_k given the model $\mathbf{M}_k$ and observed data D. From this the expectation of the model parameters can be found along with estimates of the uncertainty in the expectations. In the following sections we outline the specification of these quantities following Ref. [27, 31, 32].

1. Likelihood

Within the Markov chain model class, the likelihood of an observed data sample is given by:

$$P(D|\theta_k, \mathbf{M}_k) = \prod_{s \in \mathcal{A}} \prod_{\overleftarrow{s}^k \in \mathcal{A}^k} p(s|\overleftarrow{s}^k)^{n(\overleftarrow{s}^k s)} \ , \tag{10}$$

where $n(\overleftarrow{s}^k s)$ is the number of times the word $\overleftarrow{s}^k s$ occurs in sample D. We note that Eq. (10) is conditioned on the start sequence $\overleftarrow{s}^k_{k-1} = s_0 s_1 \ldots s_{k-1}$.

2. Prior

The prior is used to describe knowledge about the model class. In the case of $\mathbf{M}_k$ models, we choose a product of Dirichlet distributions—the so-called *conjugate prior* [31, 32]. Its form is:

$$\begin{aligned} P(\theta_k|\mathbf{M}_k) \ &= \ \prod_{\overleftarrow{s}^k \in \mathcal{A}^k} \frac{\Gamma(\alpha(\overleftarrow{s}^k))}{\prod_{s \in \mathcal{A}} \Gamma(\alpha(\overleftarrow{s}^k s))} \\ &\times \ \delta(1 - \sum_{s \in \mathcal{A}} p(s|\overleftarrow{s}^k)) \ \prod_{s \in \mathcal{A}} p(s|\overleftarrow{s}^k)^{\alpha(\overleftarrow{s}^k s) - 1} \ , \end{aligned} \tag{11}$$

where $\alpha(\overleftarrow{s}^k) = \sum_{s \in \mathcal{A}} \alpha(\overleftarrow{s}^k s)$ and $\Gamma(x)$ is the gamma function. The prior's *hyperparameters* $\{\alpha(\overleftarrow{s}^k s) : s \in \mathcal{A}, \overleftarrow{s}^k \in \mathcal{A}^k\}$ are assigned to reflect knowledge of the system at hand and must be real and positive. An intuition for the meaning of the hyperparameters can be obtained by considering the mean of the Dirichlet prior:

$$\mathbf{E}_{\text{prior}}[p(s|\overleftarrow{s}^k)] = \frac{\alpha(\overleftarrow{s}^k s)}{\alpha(\overleftarrow{s}^k)} \ . \tag{12}$$

In practice, a common assignment is $\alpha(\overleftarrow{s}^k s) = 1$ for all hyperparameters. This produces a uniform prior over all of the Markov chain parameters, reflected by the expectation $\mathbf{E}_{\text{prior}}[p(s|\overleftarrow{s}^k)] = 1/|\mathcal{A}|$.

The uniform prior for the Markov chain parameters results in a distribution of entropy rates peaked at the maximum of $\log_2 |\mathcal{A}|$ bits per symbol. As a result, estimates of the entropy rate for any order k will approach the true value from above.

3. Evidence

The evidence is often seen as merely a normalization term in Bayes' theorem—a term that can be ignored. However, when model comparison of different orders and estimation of entropy rates are considered, it becomes a fundamental part of the analysis. The evidence is defined as:

$$P(D|\mathbf{M}_k) = \int \ d\theta_k \ P(D|\theta_k, \mathbf{M}_k) P(\theta_k|\mathbf{M}_k) \ , \tag{13}$$

and provides the average of the likelihood with respect to the prior. Equivalently, it gives the probability of the data D given the model $\mathbf{M}_k$. For the likelihood and prior derived above, the evidence is found analytically [27]:

$$\begin{aligned} P(D|\mathbf{M}_k) \ &= \ \prod_{\overleftarrow{s}^k \in \mathcal{A}^k} \frac{\Gamma(\alpha(\overleftarrow{s}^k))}{\prod_{s \in \mathcal{A}} \Gamma(\alpha(\overleftarrow{s}^k s))} \\ &\times \ \frac{\prod_{s \in \mathcal{A}} \Gamma(n(\overleftarrow{s}^k s) + \alpha(\overleftarrow{s}^k s))}{\Gamma(n(\overleftarrow{s}^k) + \alpha(\overleftarrow{s}^k))} \ . \end{aligned} \tag{14}$$

4. Posterior

The posterior distribution is constructed from the elements derived above according to Bayes' theorem Eq. (9), resulting in a product of Dirichlet distributions. This is a result of choosing the conjugate prior and generates the familiar form:

$$\begin{aligned} P(\theta_k|D,\mathbf{M}_k) &= \prod_{\overleftarrow{s}^k \in \mathcal{A}^k} \frac{\Gamma(n(\overleftarrow{s}^k)+\alpha(\overleftarrow{s}^k))}{\prod_{s\in\mathcal{A}}\Gamma(n(\overleftarrow{s}^k s)+\alpha(\overleftarrow{s}^k s))} \\ &\times \delta(1-\sum_{s\in\mathcal{A}} p(s|\overleftarrow{s}^k)) \\ &\times \prod_{s\in\mathcal{A}} p(s|\overleftarrow{s}^k)^{n(\overleftarrow{s}^k s)+\alpha(\overleftarrow{s}^k s)-1} \ . \end{aligned} \tag{15}$$

The mean for the model parameters θ_k according to the posterior distribution is then:

$$\mathbf{E}_{\text{post}}[p(s|\overleftarrow{s}^k)] = \frac{n(\overleftarrow{s}^k s)+\alpha(\overleftarrow{s}^k s)}{n(\overleftarrow{s}^k)+\alpha(\overleftarrow{s}^k)} \ . \tag{16}$$

Given these estimates of the model parameters θ_k, the next step is to decide which order k is best for a given data sample.

B. Model comparison of orders k

Bayesian model comparison is very similar to the parameter estimation process discussed above. Following Ref. [27] we start by enumerating the set of model orders to consider $\mathcal{M} = \{\mathbf{M}_k : k \in [k_{min}, k_{max}]\}$. The probability of a particular order can be found by considering two factorings of the joint distribution $P(\mathbf{M}_k, D|\mathcal{M})$. Solving for the probability of a particular order we obtain:

$$P(\mathbf{M}_k|D,\mathcal{M}) = \frac{P(D|\mathbf{M}_k,\mathcal{M})P(\mathbf{M}_k|\mathcal{M})}{P(D|\mathcal{M})} \ , \tag{17}$$

where the denominator is given by the sum

$$P(D|\mathcal{M}) = \sum_{\mathbf{M}'_k\in\mathcal{M}} P(D|\mathbf{M}'_k,\mathcal{M})P(\mathbf{M}'_k|\mathcal{M}) \ . \tag{18}$$

This expression is driven by two components: the evidence $P(D|\mathbf{M}_k,\mathcal{M}) = P(D|\mathbf{M}_k)$ in Eq. (14) and the *prior over model orders* $P(\mathbf{M}_k|\mathcal{M})$. Two common priors are a uniform prior over orders and an exponential penalty for the size of the model $P(\mathbf{M}_k|\mathcal{M}) \propto \exp(-|\mathbf{M}_k|)$. For a kth-order Markov chain the size of the model, or number of free parameters, is given by $|\mathbf{M}_k| = |\mathcal{A}|^k(|\mathcal{A}|-1)$. To illustrate the method we will consider only the prior with a penalty for model size. Under these assumptions, the probability of order $\mathbf{M}_k$ is given by

$$P(\mathbf{M}_k|D,\mathcal{M}) = \frac{P(D|\mathbf{M}_k,\mathcal{M})\exp(-|\mathbf{M}_k|)}{\sum_{\mathbf{M}'_k} P(D|\mathbf{M}'_k,\mathcal{M})\exp(-|\mathbf{M}'_k|)} \ . \tag{19}$$

In the examples to follow, the model order $\mathbf{M}_k$ which has the highest probability according to the above expression will be used for estimation of the entropy rate.

C. Entropy rate estimation

The entropy rate of an inferred Markov chain can be estimated by extending the method for independent identically distributed (IID) models of discrete data [24] using *type theory* [33]. In simple terms, type theory shows that the probability of an observed sequence can be suggestively rewritten in terms of the *relative entropy* (*Kullback-Leibler* (KL) distance) and the entropy rate Eq. (3). This form suggests a connection to statistical mechanics. This, in turn, allows us to define a partition function and so to find average information-theoretic quantities over the posterior by taking derivatives. In the large data limit, the relative entropy vanishes and we are left with the desired estimation of the Markov chain's entropy rate.

We introduce this method for computing the entropy rate for two reasons. First, the result is a true average over the posterior distribution, reflecting a strict adherence to Bayesian methods. Second, this result provides a computationally efficient method for entropy rate estimation without, for example, needing numerical linear algebra packages. This provides a distinct benefit when large alphabets or Markov chain orders k are considered. The complete development is beyond our scope here; see Ref. [27]. However, we will provide a brief sketch of the derivation and quote the resulting estimator.

The connection we draw between inference and information theory starts by considering the product of the prior Eq. (11) and likelihood Eq. (10):

$$P(\theta_k|\mathbf{M}_k)P(D|\theta_k,\mathbf{M}_k) = P(D,\theta_k|\mathbf{M}_k) \ . \tag{20}$$

This product forms a joint distribution over the observed data D and model parameters θ_k given the model $\mathbf{M}_k$. Writing the normalization constant from the prior as Z to save space, this joint distribution can be written, without approximation, in terms of conditional relative entropies $\mathcal{D}[\cdot\|\cdot]$ and entropy rates $h_\mu[\cdot]$:

$$\begin{aligned} P(D,\theta_k|\mathbf{M}_k) &= Z\, 2^{-\beta_k(\mathcal{D}[Q\|P]+h_\mu[Q])} \\ &\times 2^{+|\mathcal{A}|^{k+1}(\mathcal{D}[U\|P]+h_\mu[U])} \ , \end{aligned} \tag{21}$$

where $\beta_k = \sum_{\overleftarrow{s}^k,s}\left[n(\overleftarrow{s}^k s)+\alpha(\overleftarrow{s}^k s)\right]$. The probabilities used above are:

$$\begin{aligned} Q &= \left\{ q(\overleftarrow{s}^k) = \frac{n(\overleftarrow{s}^k)+\alpha(\overleftarrow{s}^k)}{\beta_k} \ , \right. \\ &\qquad \left. q(s|\overleftarrow{s}^k) = \frac{n(\overleftarrow{s}^k s)+\alpha(\overleftarrow{s}^k s)}{n(\overleftarrow{s}^k)+\alpha(\overleftarrow{s}^k)} \right\} \end{aligned} \tag{22}$$

$$U = \left\{ q(\overleftarrow{s}^k) = \frac{1}{|\mathcal{A}|^k}, q(s|\overleftarrow{s}^k) = \frac{1}{|\mathcal{A}|} \right\} \ , \tag{23}$$

where Q is the distribution defined by the posterior mean, U is a uniform distribution, and $P = \{p(\overleftarrow{s}^k), p(s|\overleftarrow{s}^k)\}$ are the "true" parameters given the model class. The information theory quantities are given by

$$\mathcal{D}[Q\|P] = \sum_{s,\overleftarrow{s}^k} q(\overleftarrow{s}^k) q(s|\overleftarrow{s}^k) \log_2 \frac{q(s|\overleftarrow{s}^k)}{p(s|\overleftarrow{s}^k)} \quad (24)$$

$$h_\mu[Q] = -\sum_{s,\overleftarrow{s}^k} q(\overleftarrow{s}^k) q(s|\overleftarrow{s}^k) \log_2 q(s|\overleftarrow{s}^k) \ . \quad (25)$$

The form of Eq. (21) and its relation to the evidence motivates the connection to statistical mechanics. We interpret the evidence $P(D|\mathbf{M}_k) = \int d\theta_k P(D, \theta_k|\mathbf{M}_k)$ as a *partition function* $\mathcal{Z} = P(D|\mathbf{M}_k)$. Using conventional techniques from statistical mechanics, the expectation and variance of $\mathcal{D}[Q\|P] + h_\mu[Q]$ are obtained by taking derivatives of $-\log \mathcal{Z}$ with respect to β_k. In this sense $E(Q,P) = \mathcal{D}[Q\|P] + h_\mu[Q]$ plays the role of an energy and β_k is comparable to an inverse temperature. We take advantage of the known form for the evidence provided in Eq. (14) to exactly calculate the desired expectation, resulting in:

$$\mathbf{E}_{\text{post}}[\, E(Q,P)\,] = \frac{1}{\log 2} \sum_{\overleftarrow{s}^k} q(\overleftarrow{s}^k) \psi^{(0)}\left[\beta_k q(\overleftarrow{s}^k)\right] - \frac{1}{\log 2} \sum_{\overleftarrow{s}^k, s} q(\overleftarrow{s}^k) q(s|\overleftarrow{s}^k) \psi^{(0)}\left[\beta_k q(\overleftarrow{s}^k) q(s|\overleftarrow{s}^k)\right] \ , \quad (26)$$

where the polygamma function is defined as $\psi^{(n)}(x) = d^{n+1}/dx^{n+1} \log \Gamma(x)$.

The meaning of the terms on the RHS of Eq. (26) is not immediately clear. However, we can use an expansion of the $n = 0$ polygamma function $\psi^{(0)}(x) = \log x - 1/2x + \mathcal{O}(x^{-2})$, which is valid for $x \gg 1$, to find the asymptotic form:

$$\mathbf{E}_{\text{post}}[\, E(Q,P)\,] = H[q(\overleftarrow{s}^k) q(s|\overleftarrow{s}^k)] - H[q(\overleftarrow{s}^k)] + \frac{1}{2\beta_k} |\mathcal{A}|^k (|\mathcal{A}| - 1) + \mathcal{O}(1/\beta_k^2) \ . \quad (27)$$

The values of $H[q(\overleftarrow{s}^k) q(s|\overleftarrow{s}^k)]$ and $H[q(\overleftarrow{s}^k)]$ are block entropies over words of length $k+1$ and k, respectively. These entropies can be calculated using the form in Eq. (2) and the distribution Q in Eq. (22). From this expansion we can see that the first two terms make up the entropy rate $h_{\mu k}[Q] = H[q(\overleftarrow{s}^k) q(s|\overleftarrow{s}^k)] - H[q(\overleftarrow{s}^k)]$. The last term must be associated with the conditional relative entropy between the *posterior mean estimate* (PME) distribution Q and the true distribution P. Assuming the conditions for the approximation in Eq. (27) hold, the factor $1/\beta_k$ tells us that the desired expectation will approach the entropy rate as $1/N$, where N is the length of the data sample.

IV. EXAMPLES

Now that we have our instrument design and model inference methods fully specified we can describe the experiments used to test them. We consider the logistic and Hénon maps as well as the Rössler and Lorenz ODEs. The specifics of the time series generation for each system are detailed below. For each time series, a family of binary partitions, appropriate to the system under study is applied to the suitably processed data.

These examples serve to demonstrate the principles that are the basis of above methods: (1) instrument design should maximize the entropy rate of the generated symbol sequence and (2) model selection should minimize the entropy rate of the chosen model order. The first principle relies on the applicability of abstract symbolic dynamics, as laid out in Sec. II, to the dynamical noise setting. The second principle can be understood as connecting model comparison with entropy rate estimation: the Bayesian evidence appears in both model comparison and entropy-rate estimation. Careful inspection reveals that minimizing the entropy rate for a given order k maximizes the corresponding model probability.

A. Logistic Map

Data from simulations of the one-dimensional logistic map, given by

$$x_{t+1} = r x_t (1 - x_t) + \xi_t \ , \quad (28)$$

at the chaotic value of $r = 3.7$ were the basis for the analysis in our first example. A set of noise levels from $\epsilon = 10^{-4}$ to $\epsilon = 10^{-1}$ was used for the added fluctuations. Recall that noise level ϵ means a uniform density on the interval $[-\epsilon/2, \epsilon/2]$ was sampled.

For each value of ϵ considered, a random initial condition in the unit interval was generated and one thousand transient steps, not analyzed, were generated to find a typical state on the chaotic attractor. Next, a single time series $x_0, x_1, \ldots, x_{N-1}$ of length N was produced.

The family of partitions used to analyze the logistic map data were given by

$$\mathcal{P}(d) = \{\text{"0"} \sim x \in [0, d), \ \text{"1"} \sim x \in [d, 1]\} \ , \quad (29)$$

and parametrized by the *decision point* $d \in [0, 1]$.

In Figures 1 to 3 we analyze time series from the logistic map at $r = 3.7$—a typical chaotic parameter setting at which deterministic dynamics plus noise stays within the unit interval. Figure 1 illustrates the instrument design procedure for a single time series of length $N = 10^5$ generated with $\epsilon = 10^{-3}$. Panel (b) shows the entropy rate $h_\mu(d)$ for four hundred decision points d. The dependence on d is nontrivial with many local maxima. In fact, there are two prominent peaks, one at $d = 1/2$ and the other at $d = f^{-1}(1/2)$. One also sees that values of

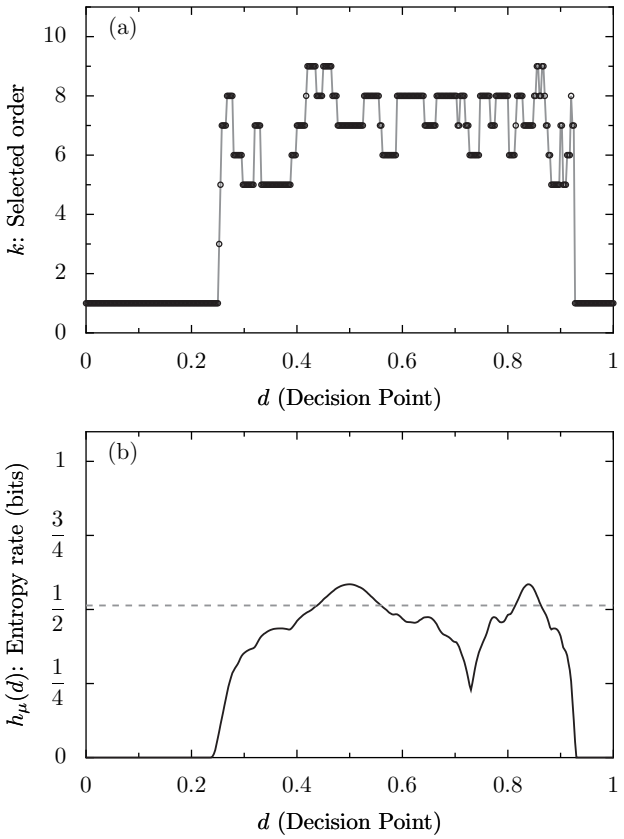

FIG. 1: Optimizing instrument design for the logistic map at $r = 3.7$ and with noise level $\epsilon = 10^{-3}$. Panel (b) plots entropy rate $h_\mu(d)$ versus decision point d. The dashed horizontal line provides the noise-free Lyapunov exponent estimate of $\lambda \approx 0.51$. Panel (a) plots the selected optimal Markov chain order for each decision point.

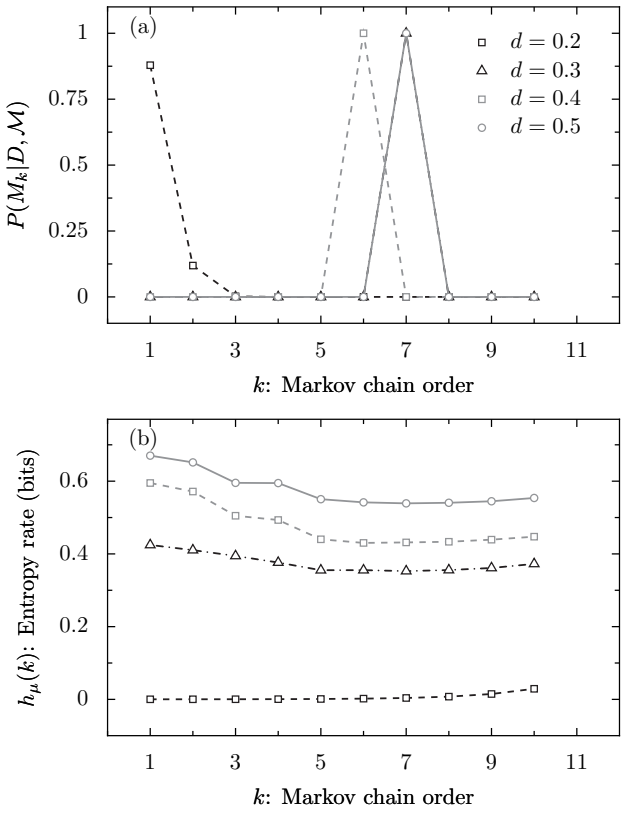

FIG. 2: The model-order selection process for a sample of decision points at $r = 3.7$. Panel (b) shows the estimated entropy rate $h_\mu(k)$ versus model order k. Panel (a) plots the posterior probability of the model order, illustrating the selection of the optimal order with minimum entropy rate.

d off the attractor produce sequences of all 0s or 1s and so have zero entropy rate.

The dashed gray line provides a numerical estimate of the Lyapunov exponent for this r value: $\lambda \approx 0.512$ bits per iteration. The entropy rate estimate for $d = 1/2$ is $h_\mu \approx 0.560$ bits per symbol. This larger estimate relative to the noise-free value of the Lyapunov exponent is not unexpected due to the additive noise and modest sample length. Unlike the noise-free form for Pesin's Identity given in Eq. (6), we expect the observed entropy rate to contain contributions from local spreading, added fluctuations, and finite data [3, 34].

In panel (a) of Fig. 1 we consider selecting an optimal Markov chain order versus d. The selected order k was used to estimate the entropy rate for each value of d. Order $k = 1$ is only selected for the values of d which produce all 0s or 1s. This reflects the fact that $d = 1/2$ is not a Markov partition for $r = 3.7$. However, this decision point is still generating. This is indicated by the entropy rate $h_\mu(d)$ reaching its maximum there. It is important to note that *simpler* models, associated with lower selected order k, can be found away from the generating partition. These are examples of apparently simpler models being inferred due to misplaced decision points—decision points with nonmaximal entropy rate.

Next, consider the model-order estimation process for $r = 3.7$ presented in Fig. 2. In panel (b) we plot the estimated entropy rate $h_\mu(k)$ versus model order k for a variety of decision points. In this case only the trivial decision point of $d = 0.2$ selects the $k = 1$ Markov chain. However, the approximate generating partition at $d = 1/2$ selects $k = 7$. Here the appropriate decision point results in a more complicated model. In panel (a) of Fig. 2 we plot the posterior model probability versus k. Again, this demonstrates that the model order with minimum entropy rate $h_\mu(k)$ is selected for each decision point.

Finally, we consider a variety of fluctuation levels from $\epsilon = 10^{-4}$ to $\epsilon = 10^{-1}$ for the logistic map at $r = 3.7$ and using a sample size $N = 10^5$ in Fig. 3. In panel (b) we plot the estimated entropy rate $h_\mu(d)$ versus decision point for each noise level. In addition, the Lyapunov exponent for the noise-free map is plotted as a dashed horizontal line. An obvious pattern, as discussed above, emerges: the entropy rate increases with ϵ for constant r and sample size N. However, the appropriate decision

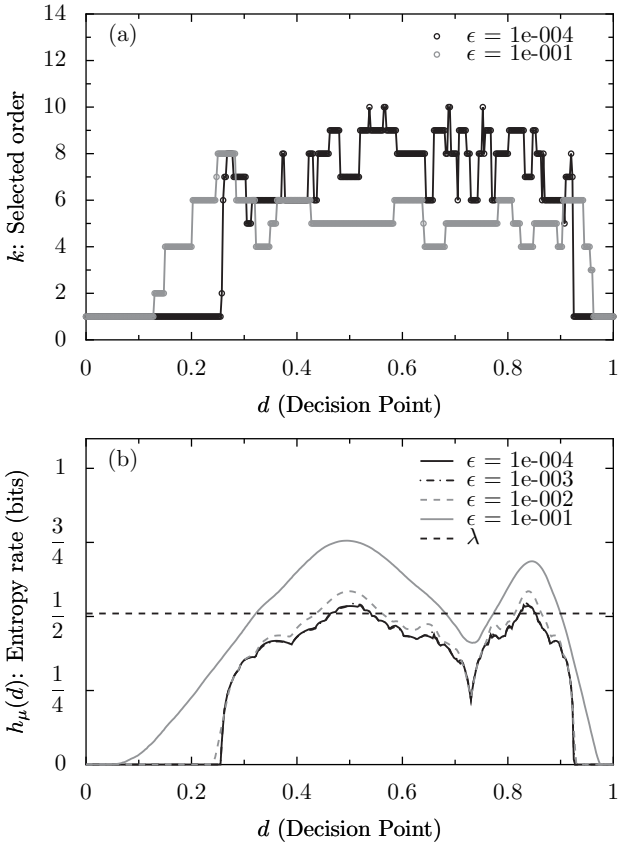

FIG. 3: Analysis of instrument design at a variety of noise levels for the logistic map at $r = 3.7$ and sample size $N = 10^5$. Panel (b) plots entropy rate $h_\mu(d)$ versus decision point d for $\epsilon = 10^{-4}$ to $\epsilon = 10^{-1}$. The dashed gray line provides the Lyapunov exponent $\lambda \approx 0.51$ for the logistic map at $r = 3.7$. Panel (a) plots the selected optimal Markov chain order for each decision point. The largest and smallest noise levels are considered.

point of $d = 1/2$ is still selected despite the increasing fluctuations. In panel (a) we consider the order of the model selected for the largest and smallest noise level, $\epsilon = 10^{-1}$ and $\epsilon = 10^{-4}$ respectively. As the noise level increases there are two interesting effects. First, the dynamics at large ϵ occur on a larger portion of the unit interval. This results in higher-order Markov chains at decision points off of the noise-free attractor. Second, the Markov chain order for regions on the noise-free attractor are *decreased* for increasing ϵ. In this, we see that fluctuations mask the underlying deterministic structure, reducing the range of temporal correlations.

B. Hénon Map

Simulations from the two-dimensional Hénon map, given by

$$\begin{aligned} x_{t+1} &= 1 - a x_t^2 + y_t + \xi_t^x \\ y_{t+1} &= b x_t + \xi_t^y \ , \end{aligned} \tag{30}$$

at $a = 1.4$ and $b = 0.3$, provided the time series for our second example. A set of noise levels from $\epsilon = 10^{-4}$ to $\epsilon = 10^{-2}$ were considered for ξ_t^x and ξ_t^y. Larger ϵ resulted in orbits that diverged. For each value of ϵ, a random initial condition was generated and one thousand transient steps, not analyzed, were generated to find a typical state on the noisy attractor. Next, a single time series of length $N = 10^5$ was produced.

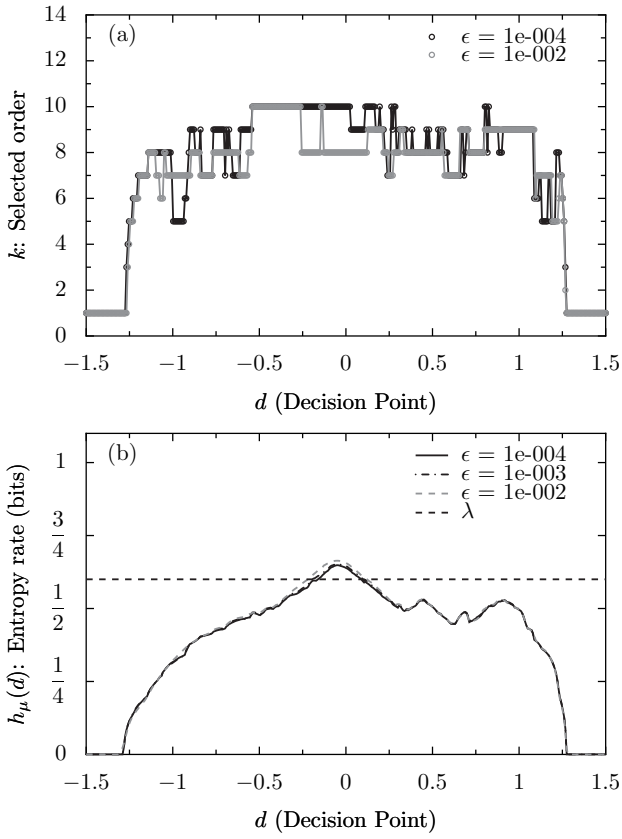

FIG. 4: The instrument design process for the Hénon map at $a = 1.4$ and $b = 0.3$. Panel (b) shows the estimated entropy rate $h_\mu(d)$ versus decision point d for various noise levels. Panel (a) plots the selected model order versus decision point for the largest and smallest noise levels considered.

For the Hénon map we consider the family of partitions

$$\mathcal{P}(d) = \{\text{“0”} \sim x \in [-1.5, d),\ \text{“1”} \sim x \in [d, 1.5]\} \ , \tag{31}$$

which dissects the $x-y$ plane into two by a line at $x = d$. (y-values of the generated time series are effectively ignored.) Although this might seem to be an overly simple instrument, the results demonstrate that it is quite effective.

In Fig. 4 we consider instrument design for the Hénon map for the parameters above. In panel (b) we plot the entropy rate $h_\mu(d)$ versus decision point d for a variety of fluctuation levels. In addition, a numerical estimation of the positive Lyapunov exponent for the noise-free Hénon map is plotted as a dashed horizontal line. As with the

logistic map, we see a nontrivial dependence of the entropy rate on the decision point with a variety of local maxima. However, the range of noise levels has a relatively small effect on the estimated entropy rate, increasing the value only slightly. As discussed above, we found values of ϵ greater than $\epsilon = 10^{-2}$ caused trajectories to diverge. For each of the fluctuation levels considered, an optimal decision point near $d = -0.05$ was found. The estimated noise-free Lyapunov exponent (calculated with a time series of length 10^5) was found to be $\lambda \approx 0.61$ bits per iteration. The estimated values of the entropy rate at the chosen decision point were found to range from $h_\mu \approx 0.65$ to 0.66 bits per symbol from smallest ϵ to largest.

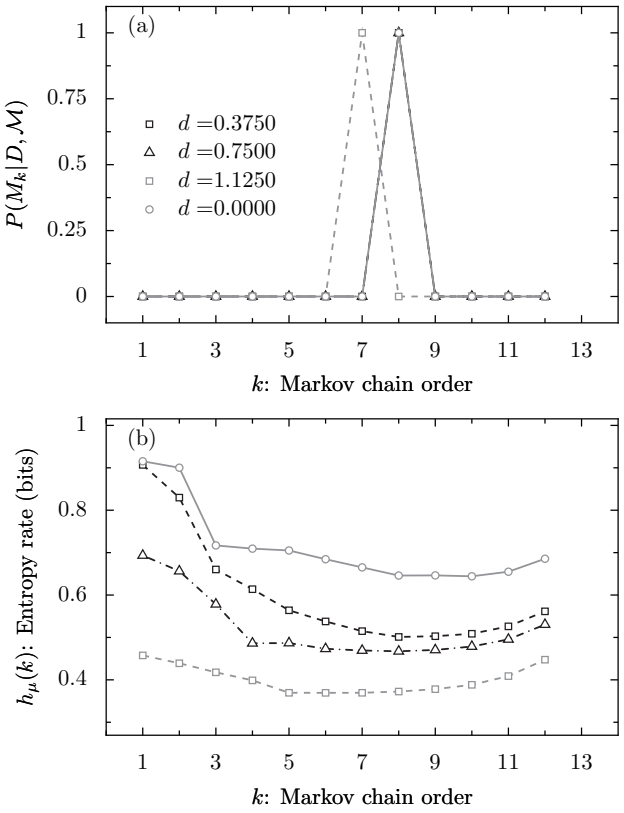

FIG. 5: The model-order selection process for the Hénon map at $a = 1.4, b = 0.3$ and fluctuation level $\epsilon = 10^{-3}$. Panel (b) shows the estimated entropy rate $h_\mu(k)$ versus model order k for a set of decision points. Panel (a) plots the posterior model probability versus model order k for a variety of decision points. Note that the results for all decision points except $d = 1.125$ overlap in this panel and select $k = 8$.

In panel (a) of Fig. 4 we show the selected model-order versus decision point for the largest and smallest fluctuations considered. The results are comparable to panel (a) of Fig. 3 for the logistic map. As we saw there, increased fluctuations generally resulted in a decrease in the selected order. We see a similar effect in the results from the Hénon map, although the effects are not as dramatic due to the limited range of ϵ which could be considered without divergence.

Finally, consider the model-selection process for the Hénon map shown in Fig. 5 for a variety of decision points. Panel (b) shows the entropy rate $h_\mu(k)$ versus order k. As with the logistic map, the decision point closest to the selected value, in this case $d = 0$, displays the largest value of $h_\mu(k)$ for all k. In addition, the minimum of $h_\mu(k)$ with respect to k identifies the selected order k. Panel (a) plots the posterior model probability versus order k. Here we see that the order k with a minimum in $h_\mu(k)$ in panel (b) is selected by model comparison, reflected by a high probability in panel (a).

C. Rössler Attractor

With our third example we analyze a symbolic time series from a system of ODEs, rather than a discrete time map. The continuous-time and continuous-state data must first be transformed so that we can apply a simple binary partition and so estimate Markov chain model. We start by describing the system of interest and the techniques used to generate and sample the data.

The Rössler flow is given by the set of ODEs [35]

$$\begin{aligned} \frac{dx}{dt} &= -y - z + \xi^x \\ \frac{dy}{dt} &= x + ay + \xi^y \\ \frac{dz}{dt} &= b + z(x - c) + \xi^z \ , \end{aligned} \tag{32}$$

where $a = 0.2$, $b = 0.2$, and $c = 5.7$ were used for our simulations. Data was generated by using a 4th-order Runge-Kutta integrator with step size $dt = 0.01$. Each noise term ξ^x, ξ^y, and ξ^z used a noise level of $\epsilon = 3$. The fluctuations were added at each integration time step. A random initial condition was generated and 50 000 time steps, not considered, were integrated to remove transients. Next, a single time series of 10^7 time steps was created and a sampling rate of one data point per 25 integration time steps was employed. In this way, a time series of $\{x_t, y_t, z_t\}$ was created.

Next, the time series data was processed in a manner that takes advantage of the attractor's low dimension. The $\{x_t\}$ time series was used to create a *Lorenz map* [36]. This is done by scanning the data for maxima in the time series. Each time a local maxima is encountered, the value is recorded as demonstrated in Fig. 6. In this way, a return map can be constructed by considering the sequence of maxima. The result of this process is shown in Fig. 7, where the first 1 000 data points from the Lorenz map are plotted. A noisy, single hump map is clearly present.

Following the recipe described above results in a set of maxima of length $N = 17\,104$. This was the time series that we analyzed.

For the Rössler data we consider a family of partitions

$$\mathcal{P}(d) = \{\text{“0”} \sim x \in [0, d),\ \text{“1”} \sim x \in [d, 15]\} \ , \tag{33}$$

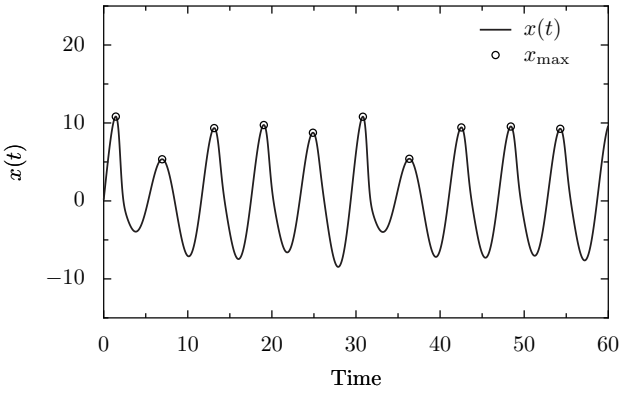

FIG. 6: The x-dynamics from the Rössler attractor at $a = 0.2$, $b = 0.2$, and $c = 5.7$. Each time a maximum is reached, the value $x_{\max}$ is recorded.

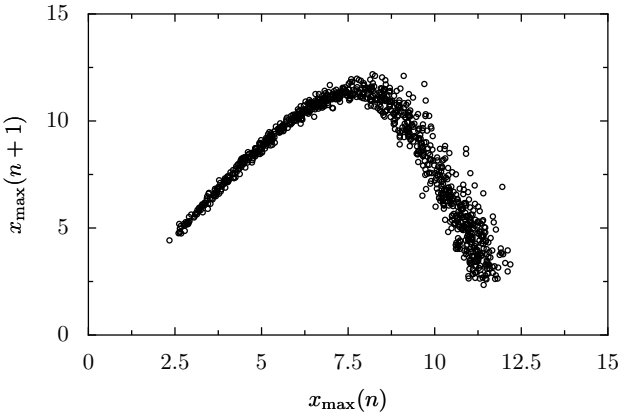

FIG. 7: Lorenz map created from the x-dynamics of the Rössler attractor. For this plot a fluctuation level of $\epsilon = 3$ was used for each degree of freedom. A noisy unimodal return map, similar to the logistic map, is clearly present.

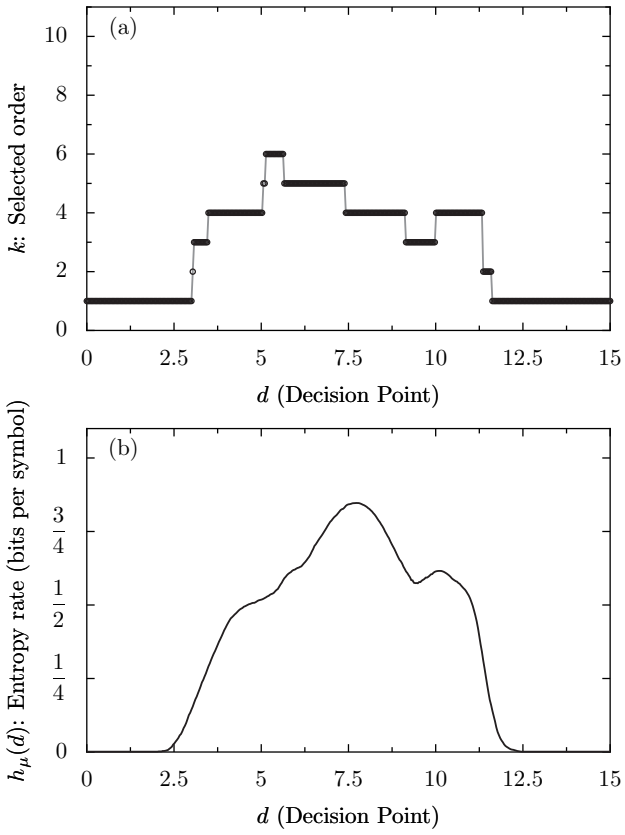

FIG. 8: The instrument design process for the Rössler attractor at $a = 0.2$, $b = 0.2$, and $c = 5.7$. Panel (b) shows the estimated entropy rate $h_\mu(d)$ versus model order d for noise level $\epsilon = 3$. Note the entropy rate is given in bits per symbol and so should not be directly compared to the positive Lyapunov exponent estimated from the flow. (See discussion in the text.) Panel (a) plots the selected model order versus decision point. Note low orders are selected due to the relatively small data sample $N = 17\,104$.

parametrized by the decision point d. The minimum of $d = 0$ and maximum of $d = 15$ were chosen to cover the range of $x_{\max}$ values exhibited plus a margin.

The results of the instrument design process for the Rössler attractor data are presented in Fig. 8. Panel(b) plots the entropy rate $h_\mu(d)$ versus decision point for four hundred values between $d = 0$ and $d = 15$. As with the previous examples, we see a nontrivial relation between the entropy rate and the decision point. There are multiple peaks with a clear global maximum of $h_\mu(d) \approx 0.85$ bits per symbol at $d = 7.725$. If we compare this with the return map in Fig. 7 we see that this decision point identifies the apparent critical point.

The entropy rate is quoted in bits per symbol and so is not numerically comparable to Lyapunov exponents given in the ODEs' natural time units. For example, in Ref. [37] the single positive Lyapunov exponent is quoted as $\lambda^+ = 0.0714$ (base-e) for noise-free dynamics. To directly compare the value we must divide our value of $h_\mu(d)$ by the average time between maxima (found to be $\Delta t = 5.85$) and multiply by $\log 2$ to convert to base-e. This results in estimated entropy rate of $h_{\text{est}} \approx 0.1$. As we saw in previous examples, a small data set with added noise is expected to display a larger value for the entropy rate. With increasing data we expect h_{est} to approach $\lambda^+ = 0.0714$. However, we do not expect equality of these estimates due to contributions from the dynamical noise.

Finally, in panel (a) of Fig. 9 we consider the model-order selected versus decision point for the Rössler attractor data. As we saw in previous examples, $k = 1$ is selected for instruments that label data as all 0s or all 1s. However, a range of orders is selected when meaningful decision points are used. The generally lower order of the Markov chains selected in this example are due to the smaller data sample of $N = 17\,104$ rather than 10^5.

D. Lorenz Attractor

In our final example we consider another well known chaotic attractor found in the Lorenz ODEs. This system models the three dominant Fourier modes of a two-dimensional fluid heated from below. As with the Rössler system, we will demonstrate that an effective symbolic representation can be found using the Lorenz map technique.

The Lorenz chaotic attractor is exhibited by the ODEs [36]

$$\begin{aligned} \frac{dx}{dt} &= \alpha(y-x)+\xi^x \\ \frac{dy}{dt} &= x(r-z)-y+\xi^y \\ \frac{dz}{dt} &= xy-bz+\xi^z \ , \end{aligned} \tag{34}$$

with $\alpha = 10$, $r = 28$, and $b = 8/3$. Data was generated by using a 4th-order Runge-Kutta integrator with step size $dt = 0.001$. Each noise term ξ^x, ξ^y, and ξ^z used a fluctuation level of $\epsilon = 6$ and noise was added at each integration time step. A larger fluctuation level was employed here due to the large size of the attractor when compared with the previous example. As before, a random initial condition was generated and 50 000 time steps, not considered, were integrated to remove transients. Next, a single time series of 10^7 time steps was created and a sampling rate of one data point per 25 integration time steps was employed, creating a time series of $\{x_t, y_t, z_t\}$.

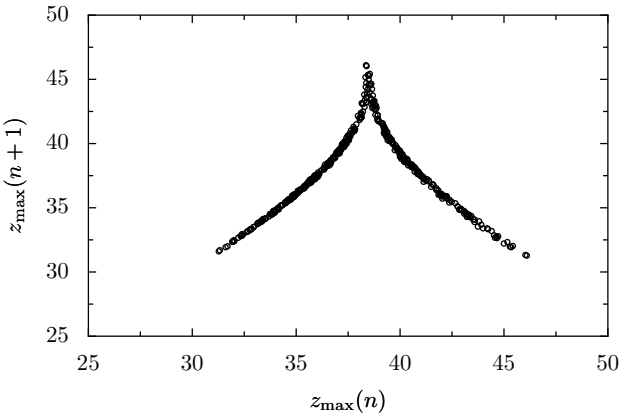

FIG. 9: The Lorenz map extracted from the z-dynamics of the Lorenz attractor. For this plot a noise level of $\epsilon = 6$ was applied to each coordinate. A noisy cusp map is associated with this attractor.

Unlike the Rössler, where the time series of x-maxima reveals the attractor's low dimension, the Lorenz system is most effectively reduced when the sequence of $\{z_{max}\}$ is considered. Using the same technique as described for Fig. 6, a sequence of z-maxima are collected resulting in a data set of size $N = 13\,311$. However, when the return map is considered in Fig. 9 a cusp map is revealed rather than the relatively smooth map of Fig. 7.

For the Lorenz data we consider a family of binary partitions

$$\mathcal{P}(d) = \{\text{“0”} \sim x \in [25, d),\ \text{“1”} \sim x \in [d, 50]\} \ . \tag{35}$$

parametrized by the decision point d. The minimum and maximum values of d were chosen so that they cover the range of observed $z_{\max}$ values plus a buffer.

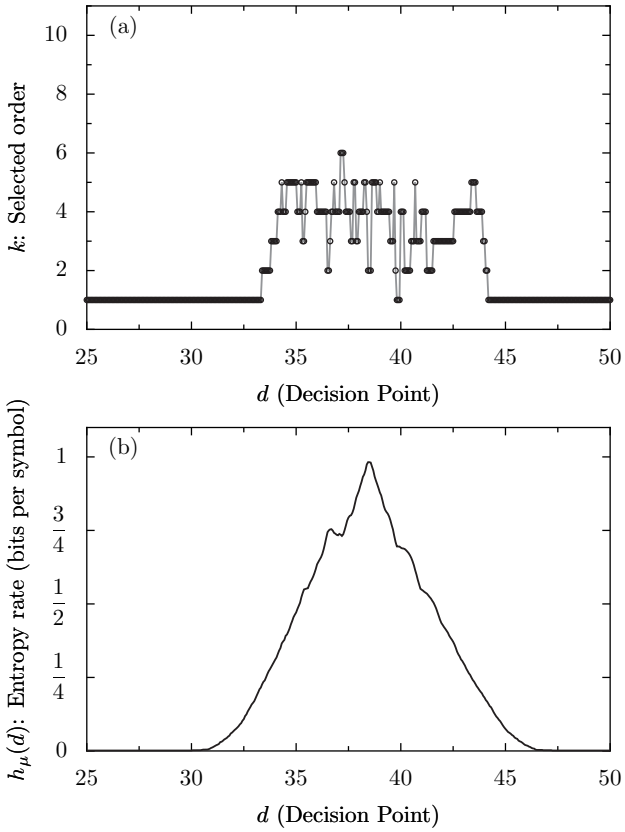

FIG. 10: The instrument design process for the Lorenz attractor at $\alpha = 10$, $r = 28$, and $b = 8/3$. Panel (b) shows the estimated entropy rate $h_\mu(d)$ versus model order d for noise level $\epsilon = 6$. Panel (a) plots the selected model order versus decision point.

In Fig. 10 we consider the instrument design process for the Lorenz data. In panel (b) we plot the entropy rate $h_\mu(d)$ versus decision point for four hundred values from $d = 25$ to $d = 50$, as given in Eq. (35). In this data set a maximum entropy rate of $h_\mu(d) = 0.982$ bits per symbol is found at $d = 38.44$. As discussed with the Rössler example, the entropy rate value is not directly comparable with the quoted positive Lyapunov exponent of $\lambda^+ = 0.9056$ (base-e) in Ref. [37]. To do this we divide by the average time per symbol, found to be $\Delta t = 0.7512$ in this case, and multiply by $\log 2$ to produce a base-e estimate. We find $h_{\text{est}} \approx 0.902$.

V. CONCLUSION

We analyzed the degree of randomness generated by deterministic chaotic systems with a small amount of additive noise. Appealing to the well developed theory of symbolic dynamics, we demonstrated that this required a two-step procedure: first, the careful design of a measuring instrument and, second, effective model-order inference from the resulting data stream. Two modeling principles emerged. The instrument should be designed to be maximally informative and the model inference should produce the most compact description in the model class. In carrying these steps out an apparent conflict appeared: in the first step of instrument design, the entropy rate was maximized; in the second, it was minimized. Moreover, it was seen that instrument design must not be simultaneously combined with model inference. In fact, treating instrument design as part of the model inference process leads to nonsensical results, such as using a trivial decision point that labels all data as a 0.

We suggest that these modeling principles can be applied to other instrument-model combinations, including those appropriate for higher-dimensional time series. In principle, an instrument is any classification method. For example, neural networks, clustering algorithms, or genetic algorithms can be employed to discretely partition a time series. The model of the resulting symbolic data stream can then be used to monitor the effectiveness of the instrument and to estimate the underlying dynamical system's invariant properties. The model class used can be Markov chains, hidden Markov models, context-trees, or ϵ-machines, to name a few options.

The lessons learned are very simply summarized, though: *Use all of the data and nothing but the data.* For noisy chaos careful measurement partition analysis coupled with Bayesian inference and Bayesian model comparison accomplish both of these goals.

Acknowledgments

This research was partially funded by the Center for Computational Science and Engineering at the University of California at Davis through support from Intel Corporation. The authors would also like to acknowledge thoughtful and constructive comments by the anonymous referees.

[1] J. Ford, Physics Today **36**, 40 (1983).
[2] J. P. Eckmann and D. Ruelle, Rev. Mod. Phys. **57(3)**, 617 (1985).
[3] J. P. Crutchfield and N. H. Packard, Physica D **7D**, 201 (1983).
[4] J. P. Crutchfield and D. P. Feldman, Chaos **13(1)**, 25 (2003).
[5] H. Kantz and T. Schrieber, *Nonlinear Time Series Analysis* (Cambridge University Press, Cambridge, UK, 2006), 2nd ed.
[6] J. P. Crutchfield and B. S. McNamara, Complex Systems **1**, 417 (1987).
[7] J. L. Breeden and A. Hübler, Phys. Rev. A **42**, 5817 (1990).
[8] H. D. I. Abarbanel et al., Rev. Mod. Phys. **65**, 1331 (1993).
[9] B.-L. Hao and W.-M. Zheng, *Applied Symbolic Dynamics and Chaos* (World Scientific, 1998).
[10] E. M. Bollt, T. Stanford, Y.-C. Lai, and K. Zyczkowski, Phys. Rev. Lett. **85**, 3524 (2000).
[11] P. Grassberger, H. Kantz, and U. Moenig, Jour. Phys. A **22**, 5217 (1989).
[12] S. D. Pethel, N. J. Corron, and E. Bollt, Phys. Rev. Lett. **96**, 34105 (2006).
[13] A. Fraser, in *Information Dynamics*, edited by H. Atmanspacher and H. Scheingraber (Plenum, New York, 1991), vol. Series B: Physics Vol. 256 of *NATO ASI Series*, p. 125.
[14] H.-P. Fang and B.-L. Hao, Chaos, Solitons & Fractals **7**, 217 (1996).
[15] C. S. Daw, C. E. A. Finney, and E. R. Tracy, Rev. Sci. Instrum. **74**, 915 (2003).
[16] R. L. Davidchack, Y.-C. Lai, E. M. Bollt, and M. Dhamala, Phys. Rev. E **61**, 1353 (2000).
[17] M. B. Kennel and M. Buhl, Phys. Rev. Lett. **91**, 84102 (2003).
[18] Y. Hirata, K. Judd, and D. Kilminster, Phys. Rev. E **70**, 016215 (2004).
[19] M. Buhl and M. B. Kennel, Phys. Rev. E **71**, 046213 (2005).
[20] T. Schurmann and P. Grassberger, Chaos **6(3)**, 414 (1996).
[21] M. B. Kennel and A. I. Mees, Phys. Rev. E **66**, 56209 (2002).
[22] Y. Hirata and A. I. Mees, Phys. Rev. E **67**, 26205 (2003).
[23] D. H. Wolpert and D. R. Wolf, Phys. Rev. E **52**, 6841 (1995).
[24] I. Samengo, Phys. Rev. E **65**, 46124 (2002).
[25] I. Nemenman, W. Bialek, and R. de Ruyter van Steveninck, Phys. Rev. E **69**, 056111 (2004).
[26] M. B. Kennel, J. Shlens, H. D. I. Abarbanel, and E. J. Chichilnisky, Neural Comput. **17**, 1531 (2005).
[27] C. C. Strelioff, J. P. Crutchfield, and A. W. Hübler, Phys. Rev. E **76**, 011106 (2007).
[28] A. N. Kolmogorov, Dokl. Akad. Nauk SSSR **119**, 861 (1958).
[29] A. N. Kolmogorov, Dokl. Akad. Nauk SSSR **124**, 754 (1959).
[30] Y. B. Pesin, Russ. Math Surv. **32**, 55 (1977).
[31] P. Baldi and S. Brunak, *Bioinformatics: The Machine Learning Approach* (MIT Press, Cambridge, 2001).
[32] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms* (Cambridge University Press, Cambridge, 2003).
[33] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley-Interscience, New York, 1991).

[34] B. Shraiman, C. E. Wayne, and P. C. Martin, Phys. Rev. Lett. **46**, 935 (1981).
[35] O. E. Rössler, Phys. Lett. A **57**, 397 (1976).
[36] E. N. Lorenz, J. Atmos. Sci. **20**, 130 (1963).
[37] J. C. Sprott, *Chaos and Time-Series Analysis* (Oxford University Press, Oxford, UK, 2003).