How Random is a Coin Toss? Bayesian Inference and the Symbolic Dynamics of Deterministic Chaos

Christopher C. Strelioff
Center for Complex Systems Research and Department of Physics
University of Illinois at Urbana-Champaign
Urbana, IL 61801
and
Center for Computational Science and Engineering
University of California at Davis
Davis, CA 95616
`strelioff@uiuc.edu`

James P. Crutchfield
Center for Computational Science and Engineering and Physics Department
University of California at Davis
Davis, CA 95616
`chaos@cse.ucdavis.edu`

Abstract

Symbolic dynamics has proven to be an invaluable tool in analyzing the mechanisms that lead to unpredictability and random behavior in nonlinear dynamical systems. Surprisingly, a discrete partition of continuous state space can produce a coarse-grained description of the behavior that accurately describes the invariant properties of an underlying chaotic attractor. In particular, measures of the rate of information production—the topological and metric entropy rates—can be estimated from the outputs of Markov or generating partitions. Here we develop Bayesian inference for k-th order Markov chains as a method for finding generating partitions and estimating entropy rates. To the best of our knowledge, this is the first time inference methods have been applied to the search for generating partitions from finite samples of data. The combination of partition selection and model inference enables us to analyze the resulting complexity of the coarse-grained model in ways not considered before.

1 Introduction

Research on chaotic dynamical systems during the last forty years produced a new vision of the origins of randomness. It is now widely understood that observed randomness can be generated by low-dimensional deterministic systems that exhibit a chaotic attractor. 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. This awareness, though, requires a new way of looking at apparently random data since chaotic dynamics are very sensitive to the measurement process [1]. As it turns out, this is both a blessing and a curse.

Symbolic dynamics, as one of a suite of tools in dynamical systems theory, in its most basic form addresses this issue by considering a coarse-grained view of a continuous dynamics. [1] In this sense, any finite-precision instrument that measures a chaotic system induces a symbolic representation of the underlying continuous-valued behavior.

To effectively model time series of discrete data from a continuous-state system two concerns must be addressed. First, we must consider the measurement 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, describing 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 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 data.

Here we address the interplay between optimal instruments and optimal models by analyzing a relatively simple nonlinear system. We consider the design of binary-output instruments for chaotic maps with additive noise. We then use Bayesian inference of a k-th order Markov chain to model the resulting data stream. Our model system is a one-dimensional chaotic map with additive noise [4, 5]

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

where $t = 0, 1, 2, \ldots$, $x_t \in [0, 1]$, and $\xi_t \sim \mathrm{N}(0, \sigma^2)$ is Gaussian random variable with mean zero and variance σ^2. 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 frame of reference.

The construction of a symbolic dynamics representation of a continuous-state system goes as follows [2]. We assume time is discrete and consider a map f from the *state space* M to itself $f : M \rightarrow M$. This space can 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 intervals: $f(I_i) = \cup_j I_j, \forall\, i$. Second, the map $f(I_i)$, restricted to an interval, 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. One might be concerned with how these methods scale to problems in higher dimensions. These ideas have been successfully applied to two dimensional maps and systems of ordinary differential equations [2]. In practice, these examples have employed comparison of the system of interest with one-dimensional maps and used approximate generating partitions with great success.

Given any partition $\mathcal{P} = \{I_i\}$, then, a series of continuous-valued states $\mathbf{X} = x_0 x_1 \ldots x_{N-1}$ 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 $\mathcal{A}$ 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 dynamics as a sequence of symbols. The choice of partition then is equivalent to our instrument-design problem.

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* $\mathrm{s}^L =$

[1] For a recent overview consult [2] and for a review of current applications see [3] and references therein.

$s_i s_{i+1} \ldots s_{i+L-1}$. 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 [6, 7]. In particular, the true entropy rate maximizes the estimated entropy rates:

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

Thus, translated into a statement about experiment design, the results tell us to design an instrument so that it maximizes the observed entropy rate. This reflects the fact that we want each measurement to produce the most information possible.

As a useful benchmark on this, useful only in the case when we know $f(x)$, *Piesin's Identity* [8] tells us that the value of h_μ is equal to the sum of the positive Lyapunov characteristic exponents: $h_\mu = \sum_i \lambda_i^+$. For one-dimensional maps there is a single Lyapunov exponent which is numerically estimated from the map f and observed trajectory $\{x_t\}$ using

$$\lambda = \lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} \log_2 |f'(x_t)| \; . \tag{6}$$

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, [4, 5] demonstrated the methods developed above are robust in the presence of noise.

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 one-dimensional map on a chaotic attractor. This means, in particular, that the 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 unimodal map with a single critical point. A decision point at this critical value produces the assumed binary alphabet and results in the most compact coarse-graining of the data. In principle, larger alphabets could be considered, but this would affect the inference process by creating fewer samples of a larger alphabet without gaining any new information. As a result we choose the simple binary alphabet.

2 Bayesian inference of k-th order Markov chains

Given a method for instrument design the next step is to estimate a model from the observed measurements. Here we choose to use the model class of k-order Markov chains and Bayesian inference as the model estimation and selection paradigm.

The k-th 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 the previous k symbols in the data sample. We introduce the more compact notation $\overleftarrow{s}_t^k = s_{t-k+1} \ldots s_t$ 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 $p(s_t | \overleftarrow{s}_t^k)$. The second assumption is stationarity. This means the probability of observed sequences does not change with the time position in the data sample: $p(s_t | \overleftarrow{s}_t^k) = p(s | \overleftarrow{s}^k)$

for any index t. As noted above, this assumption is satisfied by the data streams produced. 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.

The k-th order Markov chain model class $\mathbf{M}_k$ has a set of parameters $\theta_k = \{p(s|\overleftarrow{s}^k) : s \in \mathcal{A}, \overleftarrow{s}^k \in \mathcal{A}^k\}$. In the Bayesian inference of the model parameters θ_k we must 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 Bayes' theorem

$$P\left(\theta_k|D, \mathbf{M}_k\right) = \frac{P\left(D|\theta_k, \mathbf{M}_k\right)\ P\left(\theta_k|\mathbf{M}_k\right)}{P\left(D|\mathbf{M}_k\right)} \ . \tag{7}$$

The posterior describes the distribution of model parameters θ_k given the model class $\mathbf{M}_k$ 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 [9, 10].

2.1 Likelihood

Within the $\mathbf{M}_k$ model class, the likelihood of an observed data sample is given by

$$P(D|\theta_k, \mathbf{M}_k) = \prod_{s \in \mathcal{A}} \prod_{\overleftarrow{s}^k \in \mathcal{A}^k} p(s|\overleftarrow{s}^k)^{n(\overleftarrow{s}^k s)} \ , \tag{8}$$

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. (8) is conditioned on the start sequence $\overleftarrow{s}^k_{k-1} = s_0 s_1 \ldots s_{k-1}$.

2.2 Prior

The prior is used to describe knowledge about the model class. In the case of the $\mathbf{M}_k$ model class, we choose a product of Dirichlet distributions—the so-called *conjugate prior* [9, 10]. Its form is

$$P(\theta_k|\mathbf{M}_k) = \prod_{\overleftarrow{s}^k \in \mathcal{A}^k} \frac{\Gamma(\alpha(\overleftarrow{s}^k))}{\prod_{s \in \mathcal{A}} \Gamma(\alpha(\overleftarrow{s}^k s))} \, \delta(1 - \sum_{s \in \mathcal{A}} p(s|\overleftarrow{s}^k)) \prod_{s \in \mathcal{A}} p(s|\overleftarrow{s}^k)^{\alpha(\overleftarrow{s}^k s) - 1} \ , \tag{9}$$

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 parameters $\{\alpha(\overleftarrow{s}^k s) : s \in \mathcal{A}, \overleftarrow{s}^k \in \mathcal{A}^k\}$ are assigned to reflect knowledge of the system at hand and must be real and positive. An intuition for the meaning of the parameters can be obtained by considering the mean of the Dirichlet prior, which is

$$\mathbf{E}_{\text{prior}}[p(s|\overleftarrow{s}^k)] = \frac{\alpha(\overleftarrow{s}^k s)}{\alpha(\overleftarrow{s}^k)} \ . \tag{10}$$

In practice, a common assignment is $\alpha(\overleftarrow{s}^k s) = 1$ for all parameters. This produces a uniform prior over the model parameters, reflected by the expectation $\mathbf{E}_{\text{prior}}[p(s|\overleftarrow{s}^k)] = 1/|\mathcal{A}|$. Unless otherwise stated, all inference in the following uses the uniform prior.

2.3 Evidence

The evidence can be seen as a simple normalization term in Bayes' theorem. However, when model comparison of different orders and estimation of entropy rates are considered, this term becomes a fundamental part of the analysis. The evidence is defined

$$P(D|\mathbf{M}_k) = \int d\theta_k \ P(D|\theta_k, \mathbf{M}_k) P(\theta_k|\mathbf{M}_k) \ . \tag{11}$$

It gives the probability of the data D given the model order $\mathbf{M}_k$. For the likelihood and prior derived above, the evidence is found analytically

$$P(D|\mathbf{M}_k) = \prod_{\overleftarrow{s}^k \in \mathcal{A}^k} \frac{\Gamma(\alpha(\overleftarrow{s}^k))}{\prod_{s \in \mathcal{A}} \Gamma(\alpha(\overleftarrow{s}^k s))} \frac{\prod_{s \in \mathcal{A}} \Gamma(n(\overleftarrow{s}^k s) + \alpha(\overleftarrow{s}^k s))}{\Gamma(n(\overleftarrow{s}^k) + \alpha(\overleftarrow{s}^k))} \ . \tag{12}$$

2.4 Posterior

The posterior distribution is constructed from the elements derived above according to Bayes' theorem Eq. (7), resulting in a product of Dirichlet distributions. This form is a result of choosing the conjugate prior and generates the familiar form

$$\begin{aligned} P(\theta_k|D,\mathbf{M}_k) &= \prod_{\overleftarrow{s}^k \in \mathcal{A}^k} \frac{\Gamma(n(\overleftarrow{s}^k) + \alpha(\overleftarrow{s}^k))}{\prod_{s\in\mathcal{A}} \Gamma(n(\overleftarrow{s}^k s) + \alpha(\overleftarrow{s}^k s))} \\ &\times \; \delta(1 - \sum_{s\in\mathcal{A}} p(s|\overleftarrow{s}^k)) \prod_{s\in\mathcal{A}} p(s|\overleftarrow{s}^k)^{n(\overleftarrow{s}^k s)+\alpha(\overleftarrow{s}^k s)-1} \; . \end{aligned} \tag{13}$$

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

$$\mathbf{E}_{\text{post}}[p(s|\overleftarrow{s}^k)] = \frac{n(\overleftarrow{s}^k s) + \alpha(\overleftarrow{s}^k s)}{n(\overleftarrow{s}^k) + \alpha(\overleftarrow{s}^k)} \; . \tag{14}$$

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.

3 Model comparison of orders k

Bayesian model comparison is very similar to the parameter estimation process discussed above. 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

$$P(M_k|D,\mathcal{M}) = \frac{P(D|M_k,\mathcal{M})P(M_k|\mathcal{M})}{P(D|\mathcal{M})} \; . \tag{15}$$

where the denominator is given by the sum $P(D|\mathcal{M}) = \sum_{M_{k'}\in\mathcal{M}} P(D|M_{k'},\mathcal{M})P(M_{k'}|\mathcal{M})$. This expression is driven by two components: the evidence $P(D|M_k,\mathcal{M})$ derived above and the prior over model orders $P(M_k|\mathcal{M})$. Two common priors are a uniform prior over orders and an exponential penalty for the size of the model $P(M_k|\mathcal{M}) = \exp(-|\mathbf{M}_k|)$. For a k-th order Markov chain the size of the model, or number of free parameters, is given by $|M_k| = |\mathcal{A}|^k(|\mathcal{A}| - 1)$. To illustrate the method we will consider only the prior over orders k with a penalty for model size.

4 Estimating entropy rates

The entropy rate of an inferred Markov chain can be estimated by extending the method for independent identically distributed (IID) models of discrete data [11] using *type theory* [12]. In simple terms, type theory shows that the probability of an observed sequence can be suggestively rewritten in terms of the *Kullback-Leibler* (KL) distance and the entropy rate Eq. (3). This form suggests a connection to statistical mechanics and this, in turn, allows us to find average information-theoretic quantities over the posterior by taking derivatives. In the large data limit, the KL distance vanishes and we are left with the desired estimation of the Markov chain's entropy rate.

We introduce this new method for computing the entropy rate for two reasons. First, the result is a true average over the posterior distribution, reflecting an adherence to Bayesian methods. Second, this result provides a computationally efficient method for entropy rate estimation without need for 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, but will appear elsewhere. 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. (9) and likelihood Eq. (8) $P(\theta_k|\mathbf{M}_k)P(D|\theta_k,\mathbf{M}_k) = P(D,\theta_k|\mathbf{M}_k)$. This product forms a joint distribution over the observed data D and model parameters θ_k given the model class $\mathbf{M}_k$. 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 $h_\mu[\cdot]$

$$P(D,\theta_k|\mathbf{M}_k) = Z\, 2^{-\beta_k(\mathcal{D}[Q\|P]+h_\mu[Q])} 2^{+|\mathcal{A}|^{k+1}(\mathcal{D}[U\|P]+h_\mu[U])} \; , \tag{16}$$

where $\beta_k = \sum_{\overleftarrow{s}^k, s} n(\overleftarrow{s}^k s) + \alpha(\overleftarrow{s}^k s)$. The set of probabilities used above are

$$\begin{aligned} Q &= \left\{ q(\overleftarrow{s}^k) = \frac{n(\overleftarrow{s}^k) + \alpha(\overleftarrow{s}^k)}{\beta_k}, q(s|\overleftarrow{s}^k) = \frac{n(\overleftarrow{s}^k s) + \alpha(\overleftarrow{s}^k s)}{n(\overleftarrow{s}^k) + \alpha(\overleftarrow{s}^k)} \right\} && (17) \\ U &= \left\{ q(\overleftarrow{s}^k) = \frac{1}{|\mathcal{A}|^k}, q(s|\overleftarrow{s}^k) = \frac{1}{|\mathcal{A}|} \right\} , && (18) \end{aligned}$$

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

$$\begin{aligned} \mathcal{D}[Q\|P] &= \sum_{s, \overleftarrow{s}^k} q(\overleftarrow{s}^k) q(s|\overleftarrow{s}^k) \log_2 \frac{q(s|\overleftarrow{s}^k)}{p(s|\overleftarrow{s}^k)} && (19) \\ h_\mu[Q] &= -\sum_{s, \overleftarrow{s}^k} q(\overleftarrow{s}^k) q(s|\overleftarrow{s}^k) \log_2 q(s|\overleftarrow{s}^k) \, . && (20) \end{aligned}$$

The form of Eq. (16) and its relation to the evidence motivates the connection to statistical mechanics. We think of 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 $\mathcal{D}[Q\|P] + h_\mu[Q]$ plays the role of an internal energy and β_k is comparable to an inverse temperature. We take advantage of the known form for the evidence provided in Eq. (12) to calculate the desired expectation resulting in

$$\begin{aligned} \mathbf{E}_{\text{post}}[\, D[Q\|P] + h[Q] \,] &= \frac{1}{\log 2} \sum_{\overleftarrow{s}^k} q(\overleftarrow{s}^k) \psi^{(0)} \left[\beta_k q(\overleftarrow{s}^k) \right] && (21) \\ &- \frac{1}{\log 2} \sum_{\overleftarrow{s}^k, s} q(\overleftarrow{s}^k) q(s|\overleftarrow{s}^k) \psi^{(0)} \left[\beta_k q(\overleftarrow{s}^k) q(s|\overleftarrow{s}^k) \right] , \end{aligned}$$

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. (21) 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

$$\mathbf{E}_{\text{post}}[\, \mathcal{D}[Q\|P] + h_\mu[Q] \,] = H_{k+1}[Q] - H_k[Q] + \frac{1}{2\beta_k} |\mathcal{A}|^k (|\mathcal{A}| - 1) \, . \qquad (22)$$

From this expansion we can see that the first two terms make up the entropy rate $h_{\mu k}[Q] = H_{k+1}[Q] - H_k[Q]$. And 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. (22) 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.

5 Experimental setup

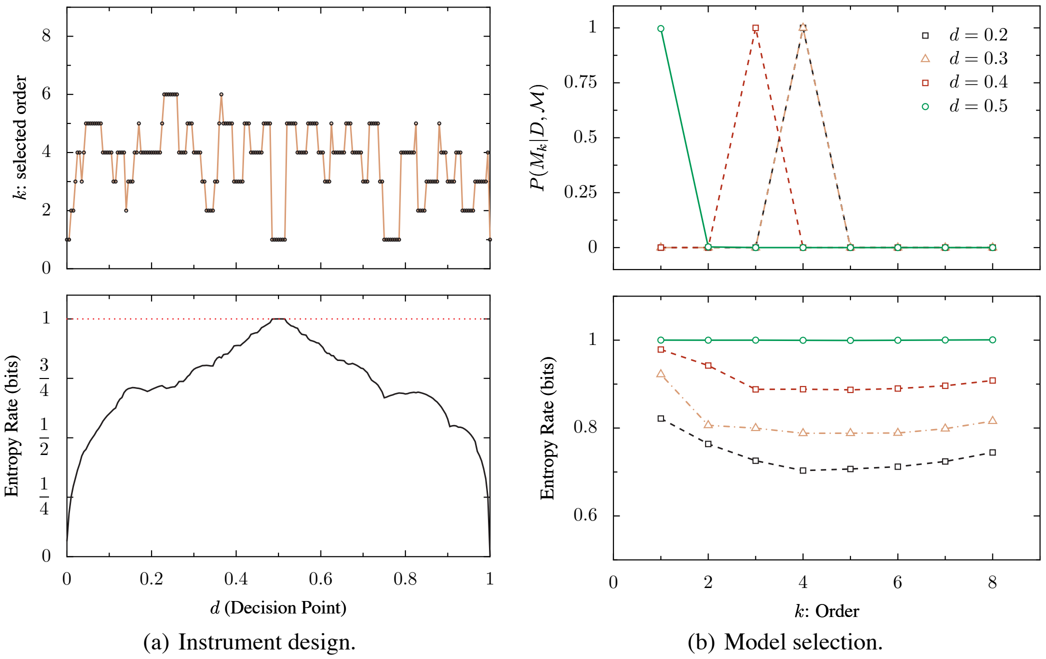
Now that we have our instrument design and model inference methods fully specified we can describe the experimental setup used to test them. Data from simulations of the one-dimensional logistic map, given by $f(x_t) = r x_t (1 - x_t)$, at the chaotic value of $r = 4.0$ was the basis for the analysis. A fluctuation level of $\sigma = 10^{-3}$ was used for the added noise. 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 $N = 10^4$ was produced.

A family of binary partitions $\mathcal{P}(d) = \{\text{"0"} \sim x \in [0, d), \text{ "1"} \sim x \in [d, 1]\}$ of the continuous-valued states was produced for two hundred decision points d between 0 and 1. That is, values in the state time series which satisfied $x_t < d$ were assigned symbol 0 and all others were assigned 1. Given the symbolic representation of the data for a particular partition $\mathcal{P}(d)$, Markov chains from order $k = 1$ to $k = 8$ were inferred and model comparison was used to select the order that most effectively described the data. Then, using the selected model, values of entropy rate $h_\mu(d)$ versus decision point d were produced.

6 Results

The results of our experiments are presented in Fig. 1. The bottom panel of Fig. 1(a) shows the entropy rate $h_\mu(d)$ versus decision point estimated using Eq. (21). Note the nontrivial d dependence of $h_\mu(d)$. The dashed line shows an accurate numerical estimate of the Lyapunov exponent using Eq. (6). It is also known to be $\lambda = 1$ bit per symbol from analytic results. We note that $h_\mu(d)$ is zero at the extremes of $d = 0$ and $d = 1$; the data stream there is all 1s or all 0s, respectively. The entropy rate estimate reaches a maximum at $d = 1/2$. For this decision point the estimated entropy rate is approximately equal to the Lyapunov exponent, indicating this instrument results in a generating partition and satisfies Piesin's identity. In fact, this value of d is also known to produce a Markov partition.

The top panel of Fig. 1(a) shows the Markov chain order k used to produce the entropy rate estimate for each value of d. This dependence on d is also complicated in ways one might not expect. The order k has two minima (ignoring $d = 0$ and $d = 1$) at $d = 1/2$ and $d = f^{-1}(1/2)$. These indicate that the model size is minimized for those instruments. This is another indication of the Markov partition for $r = 4.0$ and $d = 1/2$. These results confirm that the maximum entropy-rate instrument produces the most effective instrument for analysis of deterministic chaos in the presence of dynamical noise. The model order is minimized at the generating partition.



(a) Instrument design.

(b) Model selection.

Figure 1: Analysis of a single data stream of length $N = 10^4$ from the logistic map at $r = 4.0$ with a noise level $\sigma = 10^{-3}$. Two hundred evenly spaced decision points $d \in [0, 1]$ were used to define measurement partitions.

Now let's consider the model-order estimation process directly. The bottom panel of Fig. 1(b) shows the estimated entropy rate $h_\mu(k)$ versus model order for four different decision points. A relative minimum in the entropy rate for a given d selects the model order. This reflects an optimization for the most structure and smallest Markov chain representation of the data produced by a given instrument. The top panel in this figure shows the model probability versus k for the same set of decision points, illustrating exactly this point. The prior over model orders, which penalizes for model size, selects the Markov chain with lowest k and smallest entropy rate.

7 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. 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 precede model inference. In fact, performing the steps in the reverse order leads to nonsensical results, such as using the one or the other extreme decision point $d = 0$ or $d = 1$.

The lessons learned are very simply summarized: Use all of the data and nothing but the data. For deterministic chaos careful decision point analysis coupled with Bayesian inference and model comparison accomplishes both of theses goals.

Acknowledgments

This research was supported by the Santa Fe Institute and the Computational Science and Engineering Center at the University of California at Davis.

References

[1] E. M. Bollt, T. Stanford, Y.-C. Lai, and K. Zyczkowski. Validity of threshold-crossing analysis of symbolic dynamics from chaotic time series. *Phys. Rev. Lett.*, 85(16):3524 – 3527, 2000.

[2] B.-L. Hao and W.-M. Zheng. *Applied Symbolic Dynamics and Chaos*. World Scientific, 1998.

[3] C. S. Daw, C. E. A. Finney, and E. R. Tracy. A review of symbolic analysis of experimental data. *Rev. Sci. Instr.*, 74(2):915 – 930, 2003.

[4] J. P. Crutchfield and N. H. Packard. Symbolic dynamics of one-dimensional maps: Entropies, finite precision, and noise. *Int. J. Theo. Phys.*, 21:433–466, 1982.

[5] J P. Crutchfield and N. H. Packard. Symbolic dynamics of noisy chaos. *Physica D*, 7D:201–223, 1983.

[6] A. N. Kolmogorov. A new metric invariant of transitive dynamical systems and of endomorphisms of lebesgue spaces. *Dokl. Akad. Nauk SSSR*, 119(5):861–864, 1958.

[7] A. N. Kolmogorov. On the entropy as a metric invariant of automorphisms. *Dokl. Akad. Nauk SSSR*, 124(4):754–755, 1959.

[8] Ya. B. Piesin. *Uspek. Math. Nauk.*, 32:55, 1977.

[9] P. Baldi and S. Brunak. *Bioinformatics: The Machine Learning Approach*. MIT Press, 2001.

[10] D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[11] I. Samengo. Estimating probabilities from experimental frequencies. *Phys. Rev. E*, 65:046124, 2002.

[12] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.