

# Inferring Markov Chains: Bayesian Estimation, Model Comparison, Entropy Rate, and Out-of-class Modeling

Christopher C. Strelhoff,<sup>1,2,\*</sup> James P. Crutchfield,<sup>1,†</sup> and Alfred W. Hübler<sup>2,‡</sup>

<sup>1</sup>*Center for Computational Science & Engineering and Physics Department,  
University of California at Davis, One Shields Avenue, Davis, CA 95616*

<sup>2</sup>*Center for Complex Systems Research and Physics Department,  
University of Illinois at Urbana-Champaign, 1110 West Green Street, Urbana, Illinois 61801*

Markov chains are a natural and well understood tool for describing one-dimensional patterns in time or space. We show how to infer  $k$ -th order Markov chains, for arbitrary  $k$ , from finite data by applying Bayesian methods to both parameter estimation and model-order selection. Extending existing results for multinomial models of discrete data, we connect inference to statistical mechanics through information-theoretic (type theory) techniques. We establish a direct relationship between Bayesian evidence and the partition function which allows for straightforward calculation of the expectation and variance of the conditional relative entropy and the source entropy rate. Finally, we introduce a novel method that uses finite data-size scaling with model-order comparison to infer the structure of out-of-class processes.

PACS numbers: 02.50.Tt,02.50.Ga,05.10.Gg

## I. INTRODUCTION

Statistical inference of models from small data samples is a vital tool in the understanding of natural systems. In many problems of interest data consists of a sequence of *letters* from a finite *alphabet*. Examples include analysis of sequence information in biopolymers [1, 2], investigation of one-dimensional spin systems [3], models of natural languages [4], and coarse-grained models of chaotic dynamics [5, 6]. This diversity of potential application has resulted in the development of a variety of representations for describing discrete-valued data series.

We consider the  $k$ -th order Markov chain model class which uses the previous  $k$  letters in a sequence to predict the next letter. Inference of Markov chains from data has a long history in mathematical statistics. Early work focused on maximum likelihood methods for estimating the parameters of the Markov chain [7–9]. This work often assumed a given fixed model order. That is, no *model comparison* across orders is done. This work also typically relied on the assumed asymptotic normality of the likelihood when estimating regions of confidence and when implementing model comparison. As a result, the realm of application has been limited to data sources where these conditions are met. One consequence of these assumptions has been that data sources which exhibit *forbidden words*, symbol sequences which are not allowed, cannot be analyzed with these methods. This type of data violates the assumed normality of the likelihood function.

More recently, model comparison in the maximum likelihood approach has been extended using various *infor-*

*mation criteria*. These methods for model-order selection are based on extensions of the likelihood ratio and allow the comparison of more than two candidate models at a time. The most widely used are *Akaike's information criteria* (AIC) [10] and the *Bayesian information criteria* (BIC) [11]. (Although the latter is called Bayesian, it does not employ Bayesian model comparison in the ways we will present here.) In addition to model selection using information criteria, methods from information theory and machine learning have also been developed. Two of the most widely employed are *minimum description length* (MDL) [12] and *structural risk minimization* [13]. Note that MDL and Bayesian methods obtain similar results in some situations [14]. However, to the best of our knowledge, structural risk minimization has not been adapted to Markov chain inference.

We consider Bayesian inference of the Markov chain model class, extending previous results [2, 4, 15, 16]. We provide the details necessary to infer a Markov chain of arbitrary order, choose the appropriate order (or weight orders according to their probability), and estimate the data source's entropy rate. The latter is important for estimating the intrinsic randomness and achievable compression rates for an information source [17]. The ability to weight Markov chain orders according their probability is unique to Bayesian methods and unavailable in the model selection techniques discussed above.

In much of the literature just cited, steps of the inference process are divided into (i) point estimation of model parameters, (ii) model comparison (hypothesis testing), and (iii) estimation of functions of the model parameters. Here we will show that Bayesian inference connects all of these steps, using a unified set of ideas. Parameter estimation is the first step of inference, model comparison a second level, and estimation of the entropy rate a final step, intimately related to the mathematical structure underlying the inference process. This view of connecting model to data provides a powerful and unique

---

\*Electronic address: [streliof@uiuc.edu](mailto:streliof@uiuc.edu)

†Electronic address: [chaos@cse.ucdavis.edu](mailto:chaos@cse.ucdavis.edu)

‡Electronic address: [a-hubler@uiuc.edu](mailto:a-hubler@uiuc.edu)

understanding of inference not available in the classical statistics approach to these problems. As we demonstrate, each of these steps is vital and implementation of one step without the others does not provide a complete analysis of the data-model connection.

Moreover, the combination of inference of model parameters, comparison of performance across model orders, and estimation of entropy rates provides a powerful tool for understanding Markov chain models themselves. Remarkably, this is true even when the generating data source is outside of the Markov chain model class. Model comparison provides a sense of the structure of the data source, whereas estimates of the entropy rate provide a description of the inherent randomness. Bayesian inference, information theory, and tools from statistical mechanics presented here touch on all of these issues within a unified framework.

We develop this as follows, assuming a passing familiarity with Bayesian methods and statistical mechanics. First, we discuss estimation of Markov chain parameters using Bayesian methods, emphasizing the use of the complete marginal posterior density for each parameter, rather than point estimates with error bars. Second, we consider selection of the appropriate memory  $k$  given a particular data set, demonstrating that a mixture of orders may often be more appropriate than selecting a single order. This is certainly a more genuinely Bayesian approach. In these first two parts we exploit different forms of Bayes' theorem to connect data and model class.

Third, we consider the mathematical structure of the evidence (or marginal likelihood) and draw connections to statistical mechanics. In this discussion we present a method for estimating entropy rates by taking derivatives of a partition function formed from elements of each step of the inference procedure. Last, we apply these tools to three example information sources of increasing complexity. The first example belongs to the Markov chain model class, but the other two are examples of hidden Markov models (HMMs) that fall outside of that class. We show that the methods developed here provide a powerful tool for understanding data from these sources, even when they do not belong to the model class being assumed.

## II. INFERRING MODEL PARAMETERS

In the first level of Bayesian inference we develop a systematic relation between the data  $D$ , the chosen *model class*  $M$ , and the vector of *model parameters*  $\theta$ . The object of interest in the inference of model parameters is the *posterior probability density*  $P(\theta|D, M)$ . This is the probability of the model parameters given the observed data and chosen model. To find the posterior we first consider the joint distribution  $P(\theta, D|M)$  over the data and model parameters given that one has chosen to model the source with a representation in a certain class  $M$ . This can be factored in two ways:  $P(\theta|D, M)P(D|M)$  or  $P(D|\theta, M)P(\theta|M)$ . Setting these equal and solving

for the posterior we obtain Bayes' theorem:

$$P(\theta|D, M) = \frac{P(D|\theta, M)P(\theta|M)}{P(D|M)}. \quad (1)$$

The *prior*  $P(\theta|M)$  specifies our assumptions regarding the model parameters. We take a pragmatic view of the prior, considering its specification to be a statement of assumptions about the chosen model class. The *likelihood*  $P(D|\theta, M)$  describes the probability of the data given the model. Finally, the *evidence* (or marginal likelihood)  $P(D|M)$  is the probability of the data given the model. In the following sections we describe each of the quantities in detail on our path to giving an explicit expression for the posterior.

### A. Markov chains

The first step in inference is to clearly state the assumptions that make up the model. This is the foundation for writing down the likelihood of a data sample and informs the choice of prior. We assume that a single data set of length  $N$  is the starting point of the inference and that it consists of *symbols*  $s_t$  from a finite alphabet  $\mathcal{A}$ ,

$$D = s_0 s_1 \dots s_{N-1}, \quad s_t \in \mathcal{A}. \quad (2)$$

We introduce the notation  $\overleftarrow{s}_t^k$  to indicate a length- $k$  sequence of letters ending at position  $t$ : e.g.,  $\overleftarrow{s}_4^2 = s_3 s_4$ .

The  $k$ -th order Markov chain model class assumes finite memory and stationarity in the data source. The finite memory condition, a generalization of the conventional Markov property, can be written

$$p(D) = p(\overleftarrow{s}_{k-1}^k) \prod_{t=k-1}^{N-2} p(s_{t+1} | \overleftarrow{s}_t^k), \quad (3)$$

thereby factoring into terms which depend only on preceding words of length- $k$ . The stationarity condition can be expressed

$$p(s_t | \overleftarrow{s}_{t-1}^k) = p(s_{t+m} | \overleftarrow{s}_{t+m-1}^k), \quad (4)$$

for any  $(t, m)$ . Equation 4 results in a simplification of the notation because we no longer need to track the position index,  $p(s_t | \overleftarrow{s}_{t-1}^k) = p(s_t | \overleftarrow{s}^k) = p(s | \overleftarrow{s}^k)$  for any  $t$ . Given these two assumptions, the model parameters of the  $k$ -th order Markov chain  $\mathbf{M}_k$  are

$$\theta_k = \{ p(s | \overleftarrow{s}^k) : s \in \mathcal{A}, \overleftarrow{s}^k \in \mathcal{A}^k \}. \quad (5)$$

A normalization constraint is placed on these parameters  $\sum_{s \in \mathcal{A}} p(s | \overleftarrow{s}^k) = 1$  for each word  $\overleftarrow{s}^k$ .

The next step is to write down the elements of Bayes' theorem specific to the  $k$ -th order Markov chain.

## B. Likelihood

Given a sample of data  $D = s_0 s_1 \dots s_{N-1}$ , the likelihood can be written down using the Markov property of Eq. (3) and the stationarity of Eq. (4). This results in the form

$$P(D|\theta_k, \mathbf{M}_k) = \prod_{s \in \mathcal{A}} \prod_{\overleftarrow{s}^k \in \mathcal{A}^k} p(s|\overleftarrow{s}^k)^{n(\overleftarrow{s}^k s)}, \quad (6)$$

where  $n(\overleftarrow{s}^k s)$  is the number of times the *word*  $\overleftarrow{s}^k s$  occurs in the sample  $D$ . For future use we also introduce notation for the number of times a word  $\overleftarrow{s}^k$  has been observed  $n(\overleftarrow{s}^k) = \sum_{s \in \mathcal{A}} n(\overleftarrow{s}^k s)$ . We note that Eq. (6) is conditioned on the *start sequence*  $\overleftarrow{s}^k = s_0 s_1 \dots s_{k-1}$ .

## C. Prior

The prior  $P(\theta|M)$  is used to specify assumptions about the model to be inferred before the data is considered. Here we use *conjugate priors* for which the posterior distribution has the same functional form as the prior. Our choice allows us to derive exact expressions for many quantities of interest in inference. This provides a powerful tool for understanding what information is gained during inference and, especially, model comparison.

The exact form of the prior is determined by our assignment of *hyperparameters*  $\alpha(\overleftarrow{s}^k s)$  for the prior which balance the strength of the modeling assumptions encoded in the prior against the weight of the data. For a  $k$ -th order Markov chain, there is one hyperparameter for each word  $\overleftarrow{s}^k s$ , given the alphabet under consideration. A useful way to think about the assignment of values to the hyperparameters is to relate them to fake counts  $\tilde{n}(\overleftarrow{s}^k s)$ , such that  $\alpha(\overleftarrow{s}^k s) = \tilde{n}(\overleftarrow{s}^k s) + 1$ . In this way, the  $\alpha(\overleftarrow{s}^k s)$  can be set to reflect knowledge of the data source and the strength of these prior assumptions can be properly weighted in relation to the actual data counts  $n(\overleftarrow{s}^k s)$ .

The conjugate prior for Markov chain inference is a product of Dirichlet distributions, one for each word  $\overleftarrow{s}^k$ . It restates the finite-memory assumption from the model definition:

$$\begin{aligned} P(\theta_k|\mathbf{M}_k) &= \prod_{\overleftarrow{s}^k \in \mathcal{A}^k} \left\{ \frac{\Gamma(\alpha(\overleftarrow{s}^k))}{\prod_{s \in \mathcal{A}} \Gamma(\alpha(\overleftarrow{s}^k s))} \right. \\ &\times \delta\left(1 - \sum_{s \in \mathcal{A}} p(s|\overleftarrow{s}^k)\right) \\ &\times \left. \prod_{s \in \mathcal{A}} p(s|\overleftarrow{s}^k)^{\alpha(\overleftarrow{s}^k s) - 1} \right\}. \end{aligned} \quad (7)$$

(See App. A for relevant properties of Dirichlet distributions.) The prior's hyperparameters  $\{\alpha(\overleftarrow{s}^k s)\}$  must be real and positive. We also introduce the more compact notation  $\alpha(\overleftarrow{s}^k) = \sum_{s \in \mathcal{A}} \alpha(\overleftarrow{s}^k s)$ . The function  $\Gamma(x) = (x-1)!$  is the well known Gamma function. The  $\delta$ -function constrains the model parameters to be properly normalized:  $\sum_{s \in \mathcal{A}} p(s|\overleftarrow{s}^k) = 1$  for each  $\overleftarrow{s}^k$ .

Given this functional form, there are at least two ways to interpret what the prior says about the Markov chain parameters  $\theta_k$ . In addition to considering fake counts  $\tilde{n}(\cdot)$ , as discussed above, we can consider the range of fluctuations in the estimated  $p(s|\overleftarrow{s}^k)$ . Classical statistics would dictate describing the fluctuations via a single value with error bars. This can be accomplished by finding the average and variance of  $p(s|\overleftarrow{s}^k)$  with respect to the prior. The result is:

$$\mathbf{E}_{\text{prior}}[p(s|\overleftarrow{s}^k)] = \frac{\alpha(\overleftarrow{s}^k s)}{\alpha(\overleftarrow{s}^k)}, \quad (8)$$

$$\mathbf{Var}_{\text{prior}}[p(s|\overleftarrow{s}^k)] = \frac{\alpha(\overleftarrow{s}^k s)(\alpha(\overleftarrow{s}^k) - \alpha(\overleftarrow{s}^k s))}{\alpha(\overleftarrow{s}^k)^2(1 + \alpha(\overleftarrow{s}^k))}. \quad (9)$$

A second method, more in line with traditional Bayesian estimation, is to consider the marginal distribution for each model parameter. For a Dirichlet distribution, the marginal for any one parameter will be a Beta distribution. With this knowledge, a probability density can be provided for each Markov chain parameter given a particular setting for the hyperparameters  $\alpha(\overleftarrow{s}^k s)$ . In this way, the prior can be assigned and analyzed in substantial detail.

A common stance in model inference is to assume all things are a-priori equal. This can be expressed by assigning  $\alpha(\overleftarrow{s}^k s) = 1$  for all  $\overleftarrow{s}^k \in \mathcal{A}^k$  and  $s \in \mathcal{A}$ , adding *no* fake counts  $\tilde{n}(\overleftarrow{s}^k s)$ . This assignment results in a uniform prior distribution over the model parameters and a prior expectation:

$$\mathbf{E}_{\text{prior}}[p(s|\overleftarrow{s}^k)] = 1/|\mathcal{A}|. \quad (10)$$

## D. Evidence

Given the likelihood and prior derived above, the evidence  $P(D|M)$  is seen to be a simple normalization term in Bayes' theorem. In fact, the evidence provides the probability of the data given the model  $\mathbf{M}_k$  and so plays a fundamental role in model comparison. Formally, the definition is

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

where we can see that this term can be interpreted as an average of the likelihood over the prior distribution. Applying this to the likelihood in Eq. (6) and the prior in Eq. (7) produces

$$\begin{aligned} P(D|\mathbf{M}_k) &= \prod_{\overleftarrow{s}^k \in \mathcal{A}^k} \left\{ \frac{\Gamma(\alpha(\overleftarrow{s}^k))}{\prod_{s \in \mathcal{A}} \Gamma(\alpha(\overleftarrow{s}^k s))} \right. \\ &\times \left. \frac{\prod_{s \in \mathcal{A}} \Gamma(n(\overleftarrow{s}^k s) + \alpha(\overleftarrow{s}^k s))}{\Gamma(n(\overleftarrow{s}^k) + \alpha(\overleftarrow{s}^k))} \right\}. \end{aligned} \quad (12)$$

As we will see, this analytic expression results in the ability to make useful connections to statistical mechanics techniques when estimating entropy rates. This is another benefit of choosing a conjugate prior with known properties.

## E. Posterior

Using Bayes' theorem Eq. (1) the results of the three previous sections can be combined to obtain the posterior distribution over the parameters of the  $k$ -th order Markov chain. One finds:

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

As noted in selecting the prior, the resulting form is a Dirichlet distribution with modified parameters. This is a result of choosing the conjugate prior: cf. the forms of Eq. (7) and Eq. (13).

From Eq. (13) the estimation of the model parameters  $p(s|\overleftarrow{s}^k)$  and the uncertainty of these estimates can be given using the known properties of the Dirichlet distribution. As with the prior, there are two main ways to understand what the posterior tells us about the fluctuations in the estimated Markov chain parameters. The first uses a point estimate with "error bars". We obtain these from the mean and variance of the  $p(s|\overleftarrow{s}^k)$  with respect to the posterior, finding

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

$$\begin{aligned}
\mathbf{Var}_{\text{post}}[p(s|\overleftarrow{s}^k)] &= \frac{n(\overleftarrow{s}^k s) + \alpha(\overleftarrow{s}^k s)}{(n(\overleftarrow{s}^k) + \alpha(\overleftarrow{s}^k))^2} \\
&\times \frac{(n(\overleftarrow{s}^k) + \alpha(\overleftarrow{s}^k)) - (n(\overleftarrow{s}^k s) + \alpha(\overleftarrow{s}^k s))}{(n(\overleftarrow{s}^k) + \alpha(\overleftarrow{s}^k) + 1)}. \quad (15)
\end{aligned}$$

This is the *posterior mean estimate* (PME) of the model parameters.

A deeper understanding of Eq. (14) is obtained through a simple factoring:

$$\begin{aligned}
\mathbf{E}_{\text{post}}[p(s|\overleftarrow{s}^k)] &= \frac{1}{n(\overleftarrow{s}^k) + \alpha(\overleftarrow{s}^k)} \left[ n(\overleftarrow{s}^k) \left( \frac{n(\overleftarrow{s}^k s)}{n(\overleftarrow{s}^k)} \right) \right. \\
&\left. + \alpha(\overleftarrow{s}^k) \left( \frac{\alpha(\overleftarrow{s}^k s)}{\alpha(\overleftarrow{s}^k)} \right) \right], \quad (16)
\end{aligned}$$

where  $n(\overleftarrow{s}^k s)/n(\overleftarrow{s}^k)$  is the *maximum likelihood estimate* (MLE) of the model parameters and  $\alpha(\overleftarrow{s}^k s)/\alpha(\overleftarrow{s}^k)$  is the prior expectation given in Eq. (8). In this form, it is apparent that the posterior mean estimate is a weighted sum of the MLE and prior expectation. As a result, we can say that the posterior mean and maximum likelihood estimates converge to

the same value for  $n(\overleftarrow{s}^k) \gg \alpha(\overleftarrow{s}^k)$ . Only when the data is scarce, or the prior is set with strong conviction, does the Bayesian estimate add corrections to the MLE.

A second method for analyzing the resulting posterior density is to consider the marginal density for each parameter. As discussed with the prior, the marginal for a Dirichlet is a Beta distribution. As a result, we can either provide regions of confidence for each parameter or simply inspect the density function. The latter provides much more information about the inference being made than the point estimation just given. In our examples, to follow shortly, we plot the marginal posterior density for various parameters of interest to demonstrate the wealth of information this method provides.

Before we move on, we make a final point regarding the estimation of inference uncertainty. The form of the posterior is not meant to reflect the potential fluctuations of the data source. Instead, the width of the distribution reflects the possible Markov chain parameters which are consistent with observed data sample. These are distinct notions and should not be conflated.

## F. Predictive distribution

Once we have an inferred model, a common task is to estimate the probability of a new observation  $D^{(new)}$  given the previous data and estimated model. This is implemented by taking an average of the likelihood of the new data:

$$P(D^{(new)}|\theta_k, \mathbf{M}_k) = \prod_{\overleftarrow{s}^k \in \mathcal{A}^k, s \in \mathcal{A}} p(s|\overleftarrow{s}^k)^{m(\overleftarrow{s}^k s)} \quad (17)$$

with respect to the posterior distribution [18]:

$$\begin{aligned}
P(D^{(new)}|D, \mathbf{M}_k) &= \int d\theta_k P(D^{(new)}|\theta_k, \mathbf{M}_k) \\
&\times P(\theta_k|D, \mathbf{M}_k). \quad (18)
\end{aligned}$$

We introduce the notation  $m(\overleftarrow{s}^k s)$  to indicate the number of times the word  $\overleftarrow{s}^k s$  occurs in  $D^{(new)}$ . This method has the desirable property, compared to point estimates, that it takes into account the uncertainty in the model parameters  $\theta_k$  as reflected in the form of the posterior distribution.

The evaluation of Eq. (18) follows the same path as the calculation for the evidence and produces a similar form; we find:

$$\begin{aligned}
P(D^{(new)}|D, \mathbf{M}_k) &= \prod_{\overleftarrow{s}^k \in \mathcal{A}^k} \left\{ \frac{\Gamma(n(\overleftarrow{s}^k) + \alpha(\overleftarrow{s}^k))}{\prod_{s \in \mathcal{A}} \Gamma(n(\overleftarrow{s}^k s) + \alpha(\overleftarrow{s}^k s))} \right. \\
&\times \left. \frac{\prod_{s \in \mathcal{A}} \Gamma(n(\overleftarrow{s}^k s) + m(\overleftarrow{s}^k s) + \alpha(\overleftarrow{s}^k s))}{\Gamma(n(\overleftarrow{s}^k) + m(\overleftarrow{s}^k) + \alpha(\overleftarrow{s}^k))} \right\}. \quad (19)
\end{aligned}$$

### III. MODEL COMPARISON

With the ability to infer a Markov chain of a given order  $k$ , a common sense question is to ask how do we choose the correct order given a particular data set? Bayesian methods have a systematic way to address this through the use of *model comparison*.

In many ways, this process is analogous to inferring model parameters themselves, which we just laid out. We start by enumerating the set of model orders to be compared  $\mathcal{M} = \{\mathbf{M}_k\}_{k_{min}^{k_{max}}}$ , where  $k_{min}$  and  $k_{max}$  correspond to the minimum and maximum order to be inferred, respectively. Although we will not consider an independent, identically distributed (IID) model ( $k = 0$ ) here, we do note that this could be included using the same techniques described below.

We start with the joint probability  $P(M_k, D|\mathcal{M})$  of a particular model  $M_k \in \mathcal{M}$  and data sample  $D$ , factoring it in two ways following Bayes' theorem. Solving for the probability of a particular model class we obtain

$$P(\mathbf{M}_k|D, \mathcal{M}) = \frac{P(D|\mathbf{M}_k, \mathcal{M})P(\mathbf{M}_k|\mathcal{M})}{P(D|\mathcal{M})}, \quad (20)$$

where the denominator is the sum given by

$$P(D|\mathcal{M}) = \sum_{\mathbf{M}'_k \in \mathcal{M}} P(D|\mathbf{M}'_k, \mathcal{M})P(\mathbf{M}'_k|\mathcal{M}). \quad (21)$$

The probability of a particular model class in the set under consideration is driven by two components: the evidence  $P(D|\mathbf{M}_k, \mathcal{M})$ , derived in Eq. (12), and the prior over model classes  $P(\mathbf{M}_k|\mathcal{M})$ .

Two common priors in model comparison are: (i) all models are equally likely and (ii) models should be penalized for the number of free parameters used to fit the data. In the first instance  $P(\mathbf{M}_k|\mathcal{M}) = 1/|\mathcal{M}|$  is the same for all orders  $k$ . However, this factor cancels out because it appears in both the numerator and denominator. As a result, the probability of models using this prior becomes

$$P(\mathbf{M}_k|D, \mathcal{M}) = \frac{P(D|\mathbf{M}_k, \mathcal{M})}{\sum_{\mathbf{M}'_k \in \mathcal{M}} P(D|\mathbf{M}'_k, \mathcal{M})}. \quad (22)$$

In the second case, a common penalty for the number of model parameters is

$$P(\mathbf{M}_k|\mathcal{M}) = \frac{\exp(-|\mathbf{M}_k|)}{\sum_{\mathbf{M}'_k \in \mathcal{M}} \exp(-|\mathbf{M}'_k|)}, \quad (23)$$

where  $|\mathbf{M}_k|$  is the number of free parameters in the model. For a  $k$ -th order Markov chain, the number of free parameters is

$$|\mathbf{M}_k| = |\mathcal{A}|^k(|\mathcal{A}| - 1), \quad (24)$$

where  $|\mathcal{A}|$  is the alphabet size. Thus, model probabilities under this prior take on the form

$$P(\mathbf{M}_k|D, \mathcal{M}) = \frac{P(D|\mathbf{M}_k, \mathcal{M}) \exp(-|\mathbf{M}_k|)}{\sum_{\mathbf{M}'_k \in \mathcal{M}} P(D|\mathbf{M}'_k, \mathcal{M}) \exp(-|\mathbf{M}'_k|)}. \quad (25)$$

We note that the normalization sum in Eq. (23) cancels because it appears in both the numerator and denominator.

Bayesian model comparison has a natural *Occam's razor* in the model comparison process [18]. This means there is a natural preference for smaller models even when a uniform prior over model orders is applied. In this light, a penalty for the number of model parameters can be seen as a very cautious form of model comparison. Both of these priors, Eq. (22) and Eq. (25), will be considered in the examples to follow.

A note is in order on computational implementation. In general, the resulting probabilities can be extremely small, easily resulting in numerical underflow if the equations are not implemented with care. As mentioned in [16], computation with extended logarithms can be used to alleviate these concerns.

### IV. INFORMATION THEORY, STATISTICAL MECHANICS, AND ENTROPY RATES

An important property of an information source is its *entropy rate*  $h_\mu$ , which indicates the degree of intrinsic randomness and controls the achievable compression. A first attempt at estimating a source's entropy rate might consist of plugging a Markov chain's estimated model parameters into the known expression [17]. However, this does not accurately reflect the posterior distribution derived above. This observation leaves two realistic alternatives. The first option is to sample model parameters from the posterior distribution. These samples can then be used to calculate a set of entropy rate estimates that reflect the underlying posterior distribution. A second option, which we take here, is to adapt methods from type theory and statistical mechanics previously developed for IID models [19] to Markov chains. To the best of our knowledge this is the first time these ideas have been extended to inferring Markov chains; although cf. [20].

In simple terms, type theory shows that the probability of an observed sequence can be written in terms of the *Kullback-Leibler* (KL) *distance* and the entropy rate. When applied to the Markov chain inference problem the resulting form suggests a connection to statistical mechanics. For example, we will show that averages of the KL-distance and entropy rate with respect to the posterior are found by taking simple derivatives of a partition function.

The connection between inference and information theory starts by considering the product of the prior Eq. (7) and likelihood Eq. (6):

$$P(\theta_k|\mathbf{M}_k)P(D|\theta_k, \mathbf{M}_k) = P(D, \theta_k|\mathbf{M}_k). \quad (26)$$

This forms a joint distribution over the observed data  $D$  and model parameters  $\theta_k$  given the model order  $\mathbf{M}_k$ . Denoting the normalization constant from the prior as  $Z$

to save space, this joint distribution is

$$P(D, \theta_k | \mathbf{M}_k) = Z \prod_{\overleftarrow{s}^k, s} p(s | \overleftarrow{s}^k)^{n(\overleftarrow{s}^k s) + \alpha(\overleftarrow{s}^k s) - 1}. \quad (27)$$

This form can be written, without approximation, in terms of conditional relative entropies  $\mathcal{D}[\cdot | \cdot]$  and entropy rate  $h_\mu[\cdot]$ :

$$P(D, \theta_k | \mathbf{M}_k) = Z 2^{-\beta_k(\mathcal{D}[Q|P] + h_\mu[Q])} \times 2^{+|\mathcal{A}|^{k+1}(\mathcal{D}[U|P] + h_\mu[U])}, \quad (28)$$

where  $\beta_k = \sum_{\overleftarrow{s}^k, s} [n(\overleftarrow{s}^k s) + \alpha(\overleftarrow{s}^k s)]$  and the distribution of true parameters is  $P = \{p(\overleftarrow{s}^k), p(s | \overleftarrow{s}^k)\}$ . The distributions  $Q$  and  $U$  are given by

$$Q = \left\{ q(\overleftarrow{s}^k) = \frac{n(\overleftarrow{s}^k) + \alpha(\overleftarrow{s}^k)}{\beta_k}, \right. \quad (29)$$

$$\left. q(s | \overleftarrow{s}^k) = \frac{n(\overleftarrow{s}^k s) + \alpha(\overleftarrow{s}^k s)}{n(\overleftarrow{s}^k) + \alpha(\overleftarrow{s}^k)} \right\}$$

$$U = \left\{ u(\overleftarrow{s}^k) = \frac{1}{|\mathcal{A}|^k}, u(s | \overleftarrow{s}^k) = \frac{1}{|\mathcal{A}|} \right\}, \quad (30)$$

where  $Q$  is the distribution defined by the posterior mean and  $U$  is a uniform distribution. The information-theoretic quantities used above are given by

$$\mathcal{D}[Q|P] = \sum_{s, \overleftarrow{s}^k} q(\overleftarrow{s}^k) q(s | \overleftarrow{s}^k) \log_2 \frac{q(s | \overleftarrow{s}^k)}{p(s | \overleftarrow{s}^k)} \quad (31)$$

$$h_\mu[Q] = - \sum_{s, \overleftarrow{s}^k} q(\overleftarrow{s}^k) q(s | \overleftarrow{s}^k) \log_2 q(s | \overleftarrow{s}^k). \quad (32)$$

The form of Eq. (28) and its relation to the evidence suggests a connection to statistical mechanics: The evidence  $P(D | \mathbf{M}_k) = \int d\theta_k P(D, \theta_k | \mathbf{M}_k)$  is a partition function  $\mathcal{Z} = P(D | \mathbf{M}_k)$ . Using conventional techniques, the expectation and variance of the ‘‘energy’’

$$E(Q, P) = \mathcal{D}[Q|P] + h_\mu[Q] \quad (33)$$

are obtained by taking derivatives of the logarithm of the partition function with respect to  $\beta_k$ :

$$\mathbf{E}_{\text{post}}[E(Q, P)] = - \frac{1}{\log 2} \frac{\partial}{\partial \beta_k} \log \mathcal{Z} \quad (34)$$

$$\mathbf{Var}_{\text{post}}[E(Q, P)] = \frac{1}{\log 2} \frac{\partial^2}{\partial \beta_k^2} \log \mathcal{Z}. \quad (35)$$

The factors of  $\log 2$  in the above expressions come from the decision to use base 2 logarithms in the definition of our information-theoretic quantities. This results in values in *bits* rather than *nats* [17].

To evaluate the above expression, we take advantage of the known form for the evidence provided in Eq. (12). With the definitions  $\alpha_k = \sum_{\overleftarrow{s}^k} \alpha(\overleftarrow{s}^k)$  and

$$R = \left\{ r(\overleftarrow{s}^k) = \frac{\alpha(\overleftarrow{s}^k)}{\alpha_k}, r(s | \overleftarrow{s}^k) = \frac{\alpha(\overleftarrow{s}^k s)}{\alpha(\overleftarrow{s}^k)} \right\} \quad (36)$$

the negative logarithm of the partition function can be written

$$- \log \mathcal{Z} = \sum_{\overleftarrow{s}^k, s} \log \Gamma [\alpha_k r(\overleftarrow{s}^k) r(s | \overleftarrow{s}^k)] \quad (37)$$

$$- \sum_{\overleftarrow{s}^k} \log \Gamma [\alpha_k r(\overleftarrow{s}^k)] + \sum_{\overleftarrow{s}^k} \log \Gamma [\beta_k q(\overleftarrow{s}^k)]$$

$$- \sum_{\overleftarrow{s}^k, s} \log \Gamma [\beta_k q(\overleftarrow{s}^k) q(s | \overleftarrow{s}^k)].$$

From this expression, the desired expectation is found by taking derivatives with respect to  $\beta_k$ ; we find that

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

$$- \frac{1}{\log 2} \sum_{\overleftarrow{s}^k, s} q(\overleftarrow{s}^k) q(s | \overleftarrow{s}^k) \psi^{(0)} [\beta_k q(\overleftarrow{s}^k) q(s | \overleftarrow{s}^k)]. \quad (38)$$

The variance is obtained by taking a second derivative with respect to  $\beta_k$ , producing

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

$$+ \frac{1}{\log 2} \sum_{\overleftarrow{s}^k, s} q(\overleftarrow{s}^k)^2 q(s | \overleftarrow{s}^k)^2 \psi^{(1)} [\beta_k q(\overleftarrow{s}^k) q(s | \overleftarrow{s}^k)]. \quad (39)$$

In both of the above the polygamma function is defined  $\psi^{(n)}(x) = d^{n+1}/dx^{n+1} \log \Gamma(x)$ . (For further details, consult a reference such as [21].)

From the form of Eq. (38) and Eq. (39), the meaning is not immediately clear. We can use an expansion of the  $n = 0$  polygamma function

$$\psi^{(0)}(x) = \log x - 1/2x + \mathcal{O}(x^{-2}), \quad (40)$$

valid for  $x \gg 1$ , however, to obtain an asymptotic form for Eq. (38); we find

$$\mathbf{E}_{\text{post}}[E(Q, P)] = H[q(\overleftarrow{s}^k) q(s | \overleftarrow{s}^k)] - H[q(\overleftarrow{s}^k)]$$

$$+ \frac{1}{2\beta_k} |\mathcal{A}|^k (|\mathcal{A}| - 1) + \mathcal{O}(1/\beta_k^2). \quad (41)$$

From this we see that the first two terms make up the entropy rate  $h_\mu[Q] = H[q(\overleftarrow{s}^k) q(s | \overleftarrow{s}^k)] - H[q(\overleftarrow{s}^k)]$  and the last term is associated with the conditional relative entropy between the posterior mean distribution  $Q$  and true distribution  $P$ .

In summary, we have found the average of conditional relative entropy and entropy rate with respect to the posterior density. This was accomplished by making connections to statistical mechanics through type theory. Unlike sampling from the posterior to estimate the entropy rate, this method results in an analytic form which approaches

$h_\mu[P]$  as the inverse of the data size. This method for approximating  $h_\mu$  also provides a computational benefit. No eigenstates have to be found from the Markov transition matrix, allowing for the storage of values in sparse data structures. This provides a distinct computational advantage when large orders or alphabets are considered.

Finally, it might seem awkward to use the expectation of Eq. (33) for estimation of the entropy rate. This method was chosen because it is the form that naturally appears in writing down the likelihood-prior combination in Eq. (28). As a result of using this method, most of the results obtained above are without approximation. We were also able to show this expectation converges to the desired value in a well behaved manner.

## V. EXAMPLES

To explore how the above produces a robust inference procedure, let's now consider the statistical inference of a series of increasingly complex data sources. The first, called the *golden mean* process, is a first-order Markov chain. The second data source is called the *even process* and cannot be represented by a Markov chain with finite order. However, this source is a deterministic HMM, meaning that the current state and next output symbol uniquely determine the next state. Finally, we consider the *simple nondeterministic source*, so named since its smallest representation is as a nondeterministic HMM. (Nondeterminism here refers to the HMM structure: the current state and next output symbol do not uniquely determine the next state. This source is represented by an infinite-state deterministic HMM [22, 23].)

The golden mean, even, and simple nondeterministic processes can all be written down as models with two internal states—call them  $A$  and  $B$ . However, the complexity of the data generated from each source is of markedly different character. Our goal in this section is to consider the three main steps in inference to analyze them. First, we consider inference of a first-order Markov chain to demonstrate the estimation of model parameters with uncertainty. Second, we consider model comparison for a range of orders  $k$ . This allows us to discover structure in the data source even though the true model class cannot be captured in all cases. Finally, we consider estimation of entropy rates from these data sources, investigating how randomness is expressed in them.

While investigating these processes we consider average data counts, rather than sample counts from specific realizations, as we want to focus specifically on the average performance of Bayesian inference. To do this we take advantage of the known form of the sources. Each is described by a transition matrix  $T$ , which gives transitions between states  $A$  and  $B$ :

$$T = \begin{bmatrix} p(A|A) & p(B|A) \\ p(A|B) & p(B|B) \end{bmatrix}. \quad (42)$$

Although two of our data sources are not finite Markov chains, the transition matrix between internal states is

Markov. This means the matrix is *stochastic* (all rows sum to one) and we are guaranteed an eigenstate  $\vec{\pi}$  with eigenvalue one:  $\vec{\pi}T = \vec{\pi}$ . This eigenstate describes the asymptotic distribution over internal states:  $\vec{\pi} = [p(A), p(B)]$ .

The transition matrix can be divided into labeled matrices  $T^{(s)}$  which contain those elements of  $T$  that output symbol  $s$ . For our binary data sources one has

$$T = T^{(0)} + T^{(1)}. \quad (43)$$

Using these matrices, the average probability of words can be estimated for each process of interest. For example, the probability of word 01 can be found using

$$p(01) = \vec{\pi}T^{(0)}T^{(1)}\vec{\eta}, \quad (44)$$

where  $\vec{\eta}$  is a column vector with all 1's. In this way, for any data size  $N$ , we estimate the average count for a word as

$$n(\overleftarrow{s}^k s) = (N - k) p(\overleftarrow{s}^k s). \quad (45)$$

Average counts, obtained this way, will be the basis for all of the examples to follow.

In the estimation of the true entropy rate for the examples we use the formula

$$h_\mu = - \sum_{v \in \{A, B\}} p(v) \sum_{s \in \mathcal{A}} p(s|v) \log_2 p(s|v) \quad (46)$$

for the the golden mean and even processes, where  $p(s|v) = T_v^{(s)}$  is the probability of a letter  $s$  given the state  $v$  and  $p(v)$  is the asymptotic probability of the state  $v$  which can be found as noted above. For the simple nondeterministic source this closed-form expression cannot be applied and the entropy rate must be found using more involved methods; see [22] for further details.

### A. Golden mean process: In-class modeling

The *golden mean process* can be represented by a simple 1st-order Markov chain over a binary alphabet characterized by a single (shortest) forbidden word  $s^2 = 00$ . The defining labeled transition matrices for this data source are given by

$$T^{(0)} = \begin{bmatrix} 0 & 1/2 \\ 0 & 0 \end{bmatrix}, \quad T^{(1)} = \begin{bmatrix} 1/2 & 0 \\ 1 & 0 \end{bmatrix}. \quad (47)$$

Figure 1 provides a graphical representation of the corresponding hidden Markov chain. Inspection reveals a simple relation between the internal states  $A$  and  $B$  and the output symbols 0 and 1. An observation of 0 indicates a transition to internal state  $B$  and a 1 corresponds to state  $A$ , making this process a Markov chain over 0s and 1s.

For the golden mean the eigenstate is  $\vec{\pi} = [p(A), p(B)] = (2/3, 1/3)$ . With this vector and the labeled transition matrices any desired word count can be found as discussed above.

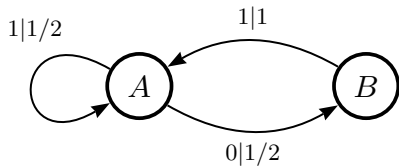


FIG. 1: A deterministic hidden Markov chain for the golden mean process. Edges are labeled with the output symbol and the transition probability: *symbol | probability*.

### 1. Estimation of $M_1$ Parameters

To demonstrate the effective inference of the Markov chain parameters for the golden mean process we consider average counts for a variety of data sizes  $N$ . For each size, the marginal posterior for the parameters  $p(0|1)$  and  $p(1|0)$  is plotted in Fig. 2. The results demonstrate that the shape of the posterior effectively describes the distribution of possible model parameters at each  $N$  and converges to the correct values of  $p(0|1) = 1/2$  and  $p(1|0) = 1$  with increasing data.

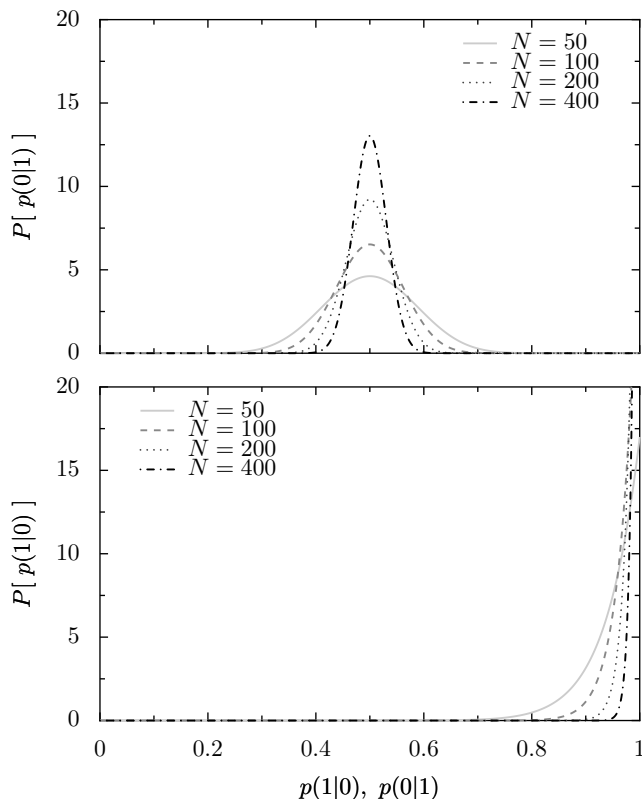


FIG. 2: A plot of the inference of  $M_1$  model parameters for the golden mean process. For each data sample size  $N$ , the marginal posterior is plotted for the parameters of interest:  $p(0|1)$  in the top panel and  $p(1|0)$  in the lower panel. The *true* values of the parameters are  $p(0|1) = 1/2$  and  $p(1|0) = 1$ .

Point estimates with a variance can be provided for each of the parameters, but these numbers by themselves can be misleading. However, the estimate obtained by using the mean and variance of the posterior are a more effective description of the inference process than a maximum likelihood estimate with estimated error given by a Gaussian approximation of the likelihood alone. As Fig. 2 demonstrates, in fact, a Gaussian approximation of uncertainty is an ineffective description of our knowledge when the Markov chain parameters are near their upper or lower limits at 0 and 1. Probably the most effective set of numbers to provide consists of the mean of the posterior and a region of confidence. These would most accurately describe asymmetries in the uncertainty of model parameters. Although we will not do that here, a brief description of finding regions of confidence is provided in App. A 1.

### 2. Selecting the Model Order $k$

Now consider the selection of the appropriate order  $k$  from golden mean realizations. As discussed above, the golden mean process is a first order Markov chain with  $k = 1$ . As a result, we would expect model comparison to select this order from the available possibilities. To demonstrate this, we consider orders  $k = 1 - 4$  and perform model comparison with a uniform prior over orders (Eq. (22)) and with a penalty for the number of model parameters (Eq. (25)).

The results of the model comparisons are given in Fig. 3. The top panel shows the probability for each order  $k$  as a function of the sample size, using a uniform prior. For this prior over orders,  $M_1$  is selected with any reasonable amount of data. However, there does seem to be a possibility to over-fit for small data size  $N \leq 100$ . The bottom panel shows the model probability with a penalty prior over model order  $k$ . This removes the over-fitting at small data sizes and produces an offset which must be overcome by the data before higher  $k$  is selected. This example is not meant to argue for the penalty prior over model orders. In fact, Bayesian model comparison with a uniform prior does an effective job using a relatively small sample size.

### 3. Estimation of Entropy Rate

We can also demonstrate the convergence of the average for  $E(Q, P) = D[Q||P] + h_\mu[Q]$  given in Eq. (38) to the correct entropy rate for the golden mean process. We choose to show this convergence for all orders  $k = 1 - 4$  discussed in the previous section. This exercise demonstrates that all orders greater than or equal to  $k = 1$  effectively capture the entropy rate. However, the convergence to the correct values for higher-order  $k$  takes more data because of a larger initial value of  $D[Q||P]$ . This larger value is simply due to the larger number of parameters for higher-order Markov chains.

In evaluating the value of  $D[Q||P] + h_\mu[Q]$  for different sample lengths, we expect that the PME estimated

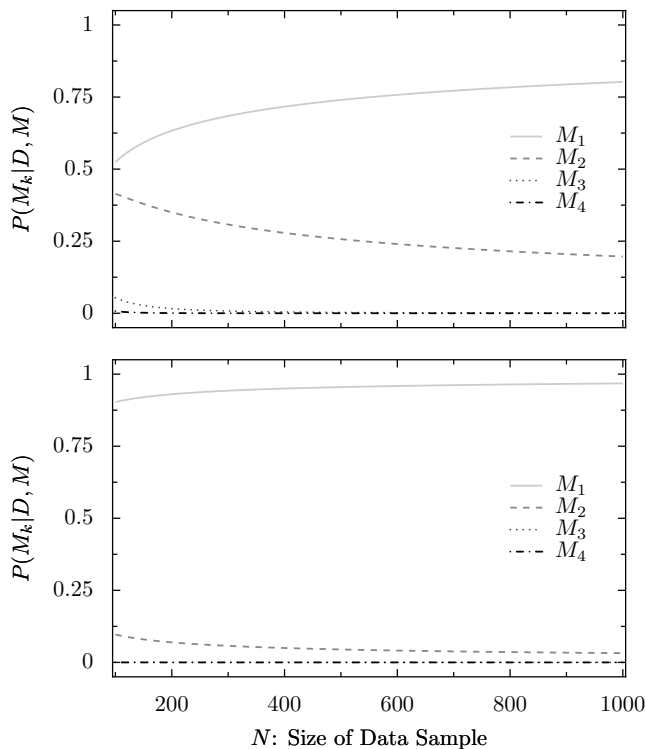


FIG. 3: Model comparison for Markov chains of order  $k = 1 - 4$  using average counts from the golden mean process. Sample sizes from  $N = 100$  to  $N = 1,000$  in steps of  $\Delta N = 5$  are used to generate these plots. The top panel displays the model probabilities using a uniform prior over orders  $k$ . The bottom panel displays the effect of a penalty for model size.

$Q$  will converge to the true distribution  $P$ . As a result, the conditional relative entropy should go to zero with increasing  $N$ . For the golden mean process, the known value of the entropy rate is  $h_\mu = 2/3$  bits per symbol. Inspection of Fig. 4 demonstrates the expected convergence of the average from Eq. (38) to the true entropy rate.

The result of our model comparison from the previous section could also be used in the estimation of the entropy rate. As we saw in Fig. 3, there are ranges of sample length  $N$  where the probability of orders  $k = 1, 2$  are both nonzero. In principle, an estimate of  $h_\mu$  should be made by weighting the values obtained for each  $k$  by the corresponding order probability  $P(\mathbf{M}_k | D, \mathcal{M})$ . As we can see from Fig. 4, the estimates of the entropy rate for  $k = 1, 2$  are also very similar in this range of  $N$ . As a result, this additional step would not have a large effect for entropy rate estimation.

### B. Even process: Out-of-class modeling

We now consider a more difficult data source called the *even process*. The defining labeled transition matrices are

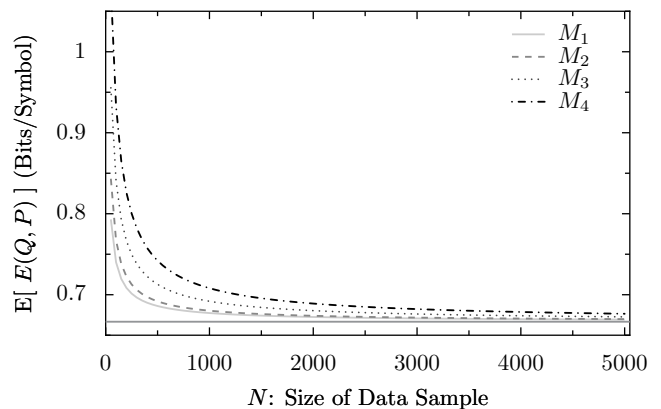


FIG. 4: The convergence of  $\mathbf{E}_{\text{post}}[E(Q, P)]$  to the true entropy rate  $h_\mu = 2/3$  bits per symbol (indicated by the gray horizontal line) for the the golden mean process. As demonstrated in Eq. (41), the conditional relative entropy  $D[Q||P] \rightarrow 0$  as  $1/N$ . This results in the convergence of  $h_\mu[Q]$  to the true entropy rate.

given by

$$T^{(0)} = \begin{bmatrix} 1/2 & 0 \\ 0 & 0 \end{bmatrix}, T^{(1)} = \begin{bmatrix} 0 & 1/2 \\ 1 & 0 \end{bmatrix}. \quad (48)$$

As can be seen in Fig. 5, the node-edge structure is identical to the golden mean process but the output symbols on the edges have been changed slightly. As a result of this shuffle, the states  $A$  and  $B$  can no longer be associated with a simple sequence of 0's and 1's. Whereas the golden mean has the irreducible set of forbidden words  $\mathcal{F} = \{00\}$ , the even process has a countably infinite set  $\mathcal{F} = \{01^{2n+1}0 : n = 0, 1, 2, \dots\}$  [22].

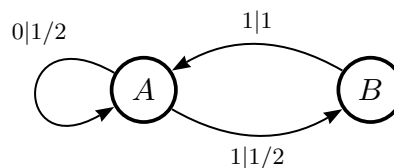


FIG. 5: Deterministic hidden Markov chain representation of the even process. This process cannot be represented as a finite-order (nonhidden) Markov chain over the output symbols 0s and 1s. The set of irreducible forbidden words  $\mathcal{F} = \{01^{2n+1}0 : n = 0, 1, 2, \dots\}$  reflects the fact that the process generates blocks of 1's, bounded by 0s, that are *even* in length, at any length.

In simple terms, the even process produces blocks of 1's which are even in length. This is a much more complicated type of memory than we saw in the golden mean process. For the Markov chain model class, where a word of length  $k$  is used to predict the next letter, this would require an infinite-order  $k$ . It would be necessary to keep

track of all even and odd strings of 1's, irrespective of the length. As a result, the properties of the even process mean that a finite Markov chain *cannot* represent this data source.

This example is then a demonstration of what can be learned in a case of out-of-class modeling. We are interested, therefore, in how well Markov chains approximate the even process. We expect that model comparison will select larger  $k$  as the size of the data sample increases. Does the model selection tell us anything about the underlying data source despite the inability to exactly capture its properties? As we will see, we do obtain intriguing hints of the true nature of the even process from model comparison. Finally, can we estimate the entropy rate of the process with a Markov chain? As we will see, a high  $k$  is needed to do this effectively.

### 1. Estimation of $M_1$ Parameters

In this section we consider an  $M_1$  approximation of the even process. We expect the resulting model to accurately capture length-2 word probabilities as  $N$  increases. In this example, we consider the *true* model to be the best approximation possible by a  $k = 1$  Markov chain. From the labeled transition matrices given above we can calculate the appropriate values for  $p(0|1)$  and  $p(1|0)$  using the methods described above. Starting from the asymptotic distribution  $\vec{\pi} = [p(A), p(B)] = [2/3, 1/3]$  we obtain  $p(0|1) = p(10)/p(1) = 1/4$  and  $p(1|0) = p(01)/p(0) = 1/2$ .

As we can see from Fig. 6, a first-order Markov chain can be inferred without difficulty. The values obtained are exactly as expected. However, these values do not tell us much about the nature of the data source by themselves. This points to the important role of model comparison and entropy rate estimation in understanding this data.

### 2. Selecting the Model Order $k$

Now consider the selection of Markov chain order  $k = 1-4$  for a range of data sizes  $N$ . Recall that the even process cannot be represented by a finite-order Markov chain over the output symbols 0 and 1. As a consequence, we expect higher  $k$  to be selected with increasing data  $N$ , as more data statistically justifies more complex models. This is what happens, in fact, but the way in which orders are selected as we increase  $N$  provides structural information we could not obtain from the inference of a Markov chain of fixed order.

If we consider Fig. 7, an interesting pattern becomes apparent. Orders with even  $k$  are preferred over odd. In this way model selection is hinting at the underlying structure of the source. The Markov chain model class cannot represent the even process in a compact way, but

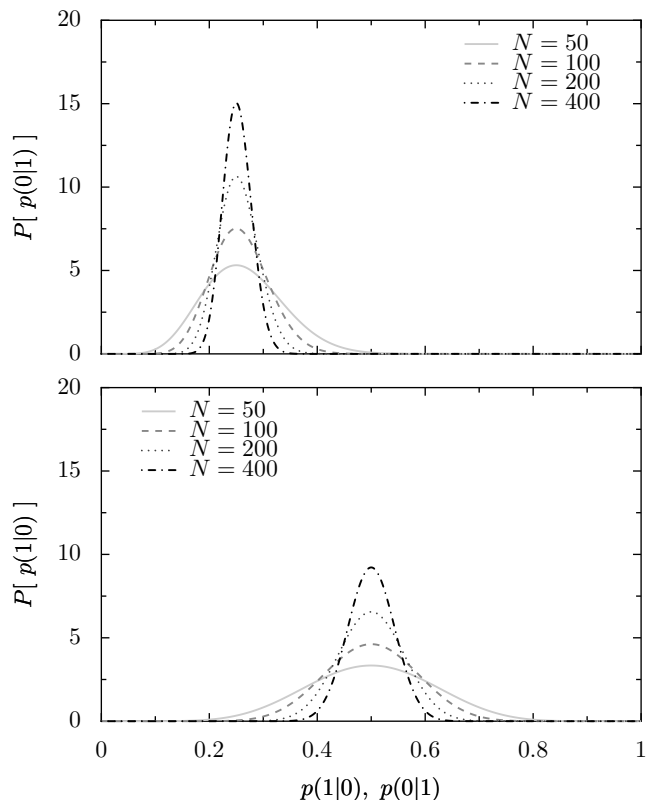


FIG. 6: A plot of the inference of  $M_1$  model parameters for the even process. For a variety of sample sizes  $N$ , the marginal posterior for  $p(0|1)$  (top panel) and  $p(1|0)$  (bottom panel) are shown. The *true* values of the parameters are  $p(0|1) = 1/4$  and  $p(1|0) = 1/2$ .

inference and model comparison combined provide useful information about the hidden structure of the source.

In this example we also have regions where the probability of multiple orders  $k$  are equally probable. The sample size at which this occurs depends on the prior over orders which is employed. When this happens, properties estimated from the Markov chain model class should use a weighted sum of the various orders. As we will see in the estimation of entropy rates, this is not as critical. At sample sizes where the order probabilities are similar, the estimated entropy rates are also similar.

### 3. Estimation of Entropy Rate

Entropy rate estimation for the even process turns out to be a more difficult task than one might expect. In Fig. 8 we see that Markov chains of orders 1–6 are unable to effectively capture the true entropy rate. In fact, experience shows that an order  $k = 10$  Markov chain or higher is needed to get close to the true value of  $h_\mu = 2/3$  bits per symbol. Note also the factor of 20 longer realizations that are required compared, say, to the golden mean example.

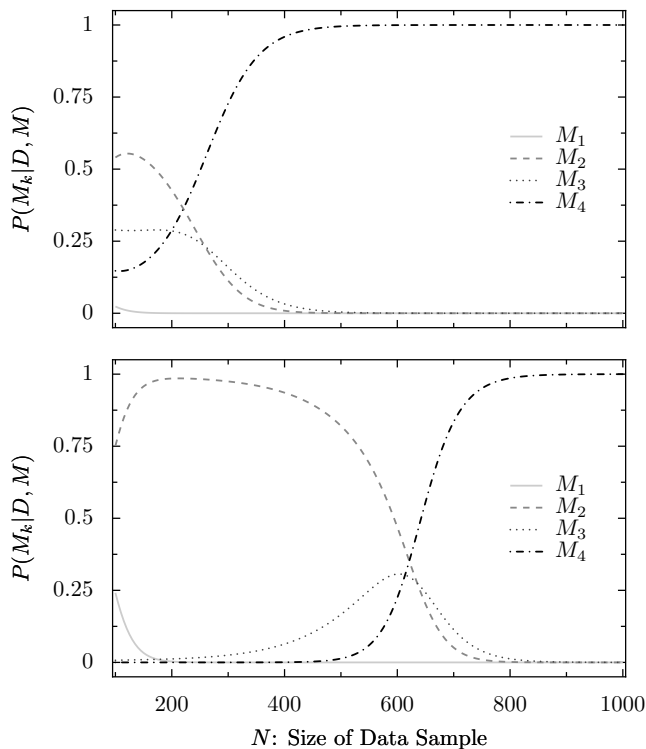


FIG. 7: Model comparison for Markov chains of order  $k = 1 - 4$  for average data from the even process. The top panel shows the model comparison with a uniform prior over the possible orders  $k$ . The bottom panel demonstrates model comparison with a penalty for the number of model parameters. In both cases the  $k = 4$  model is chosen over lower orders as the amount of data available increases.

As discussed above, a weighted sum of  $\mathbf{E}_{\text{post}}[D[Q||P] + h_\mu[Q]]$  could be employed in this example. For the estimate this is not critical because the different orders provide roughly the same value at these points. In fact, these points correspond to where the estimates of  $E(Q, P)$  cross in Fig. 8. They are samples sizes where apparent randomness can be explained by structure and increased order  $k$ .

### C. Simple Nondeterministic Source: Out-of-class modeling

The simple nondeterministic source adds another level of challenge to inference. As its name suggests, it is described by a nondeterministic HMM. Considering Fig. 9 we can see that a 1 is produced on every transition except for the  $B \rightarrow A$  edge. This means there are many paths through the internal states that produce the same observable sequence of 0s and 1s. The defining labeled transition matrices for this process are given by

$$T^{(0)} = \begin{bmatrix} 0 & 0 \\ 1/2 & 0 \end{bmatrix}, T^{(1)} = \begin{bmatrix} 1/2 & 1/2 \\ 0 & 1/2 \end{bmatrix}. \quad (49)$$

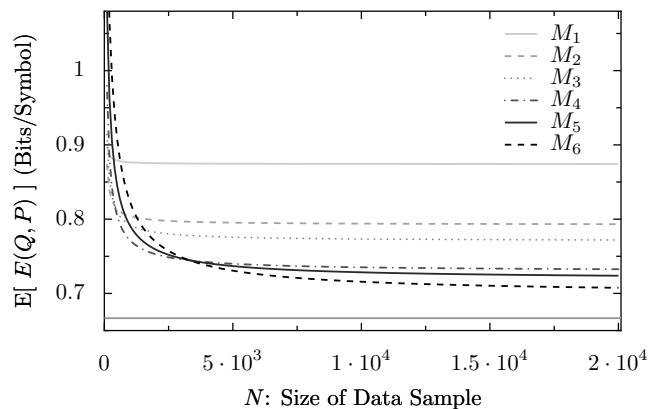


FIG. 8: The convergence of  $\mathbf{E}_{\text{post}}[D[Q||P] + h_\mu[Q]]$  to the true entropy rate  $h_\mu = 2/3$  bits per symbol for the the even process. The true value is indicated by the horizontal gray line. Experience shows that a  $k = 10$  Markov chain is needed to effectively approximate the true value of  $h_\mu$ .

Using the state-to-state transition matrix  $T = T^{(0)} + T^{(1)}$ , we find the asymptotic distribution for the hidden states to be  $\vec{\pi} = [p(A), p(B)] = [1/2, 1/2]$ . Each of the hidden states is equally likely; however, a 1 is always produced from state  $A$ , while there is an equal chance of obtaining a 0 or 1 from state  $B$ .

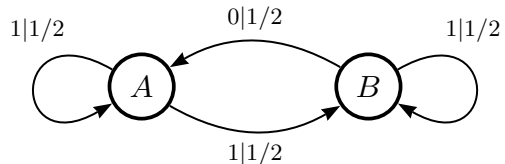


FIG. 9: A hidden Markov chain representation of the simple nondeterministic process. This example also cannot be represented as a finite-order Markov chain over outputs 0 and 1. It, however, is more complicated than the two previous examples: Only the observation of a 0 provides the observer with information regarding the internal state of the underlying process; observing a 1 leaves the internal state ambiguous.

#### 1. Estimation of $M_1$ Parameters

Using the asymptotic distribution derived above, the parameters of an inferred first-order Markov chain should approach  $p(0|1) = p(10)/p(1) = 1/3$  and  $p(1|0) = p(01)/p(0) = 1$ . As we can see from Fig. 10, the inference process captures these values very effectively despite the out-of-class data source.

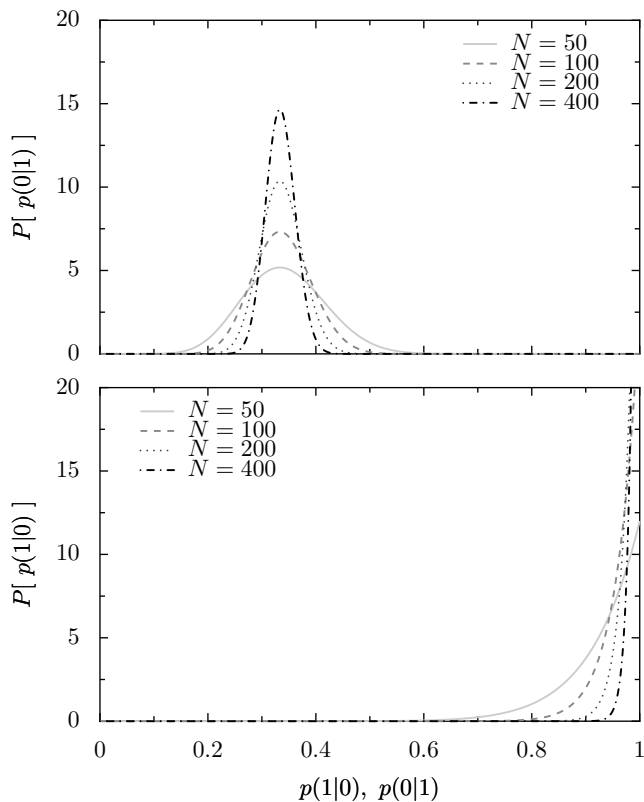


FIG. 10: Marginal density for  $M_1$  model parameters for the simple nondeterministic process: The curves for each data size  $N$  demonstrate a well behaved convergence to the correct values:  $p(0|1) = 1/3$  and  $p(1|0) = 1$ .

### 2. Selecting the Model Order $k$

Here we consider the comparison of Markov chain models of orders  $k = 1 - 4$  when applied to data from the simple nondeterministic source. As with the even process, we expect increasing order to be selected as the amount of available data increases. In Fig. 11 we see that this is exactly what happens.

Unlike the even process, there is no preference for even orders. Instead, we observe a systematic increase in order with larger data sets. We do note that the amount of data need to select a higher order does seem to be larger than for the even process. Here the distribution over words is more important and more subtle than the support of the distribution (those words with positive probability).

### 3. Estimation of Entropy Rate

Estimation of the entropy rate for the simple nondeterministic source provides an interesting contrast to the previous examples. As discussed when introducing the examples, this data source is a nondeterministic HMM and the entropy rate cannot be directly calculated using Eq. (46) [24]. However, a value of  $h_\mu \approx 0.677867$  bits

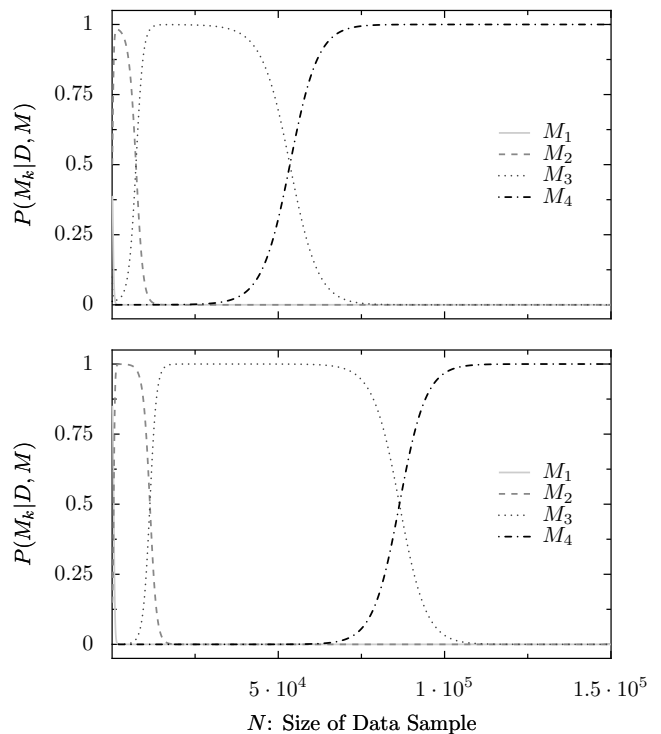


FIG. 11: Model comparison for Markov chains of order  $k = 1 - 4$  for data from the simple nondeterministic process. The top panel shows the model comparison with a uniform prior over the possible orders  $k$ . The bottom panel demonstrates model comparison with a penalty for the number of model parameters. Note the scale on the horizontal axis—it takes much more data for the model comparison to pick out higher orders for this process compared to the previous examples.

per symbol has been obtained in [22].

Figure 12 shows the results of entropy-rate estimation using Markov chains of order  $k = 1 - 6$ . These results demonstrate that the entropy rate can be effectively estimated with low-order  $k$  and relatively small data samples. This is an interesting result, as we might expect estimation of the entropy rate to be most difficult in this example. Instead we find that the even process was a more difficult test case.

## VI. DISCUSSION

The examples presented above provide several interesting lessons in inference, model comparison, and estimating randomness. The combination of these three ideas applied to a data source provides information and intuition about the structure of the underlying system, even when modeling out-of-class processes.

In the examples of  $M_1$  estimates for each of the sources we see that the Bayesian methods provide a powerful and consistent description of Markov chain model parameters. The marginal density accurately describes the uncertainty associated with these estimates, reflecting

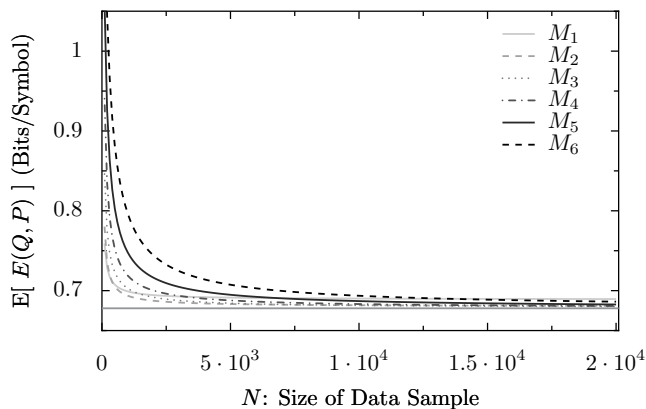


FIG. 12: The convergence of  $\mathbf{E}_{\text{post}}[D[Q||P] + h_\mu[Q]]$  to the true entropy rate  $h_\mu \approx 0.677867$  bits per symbol for the simple nondeterministic source. The true value is indicated by the gray horizontal line.

asymmetries which point estimation with error bars cannot capture. In addition, methods described in App. A 1 can be used to generate regions of confidence of any type.

Although the estimates obtained for the Markov chain model parameters were consistent with the data source for words up to length  $k + 1$ , they did not capture the true nature of the system under consideration. This demonstrates that estimation of model parameters without some kind of model comparison can be very misleading. Only with the comparison of different orders did some indication of the true properties of the data source become clear. Without this step, misguided interpretations are easily obtained.

For the golden mean process, a  $k = 1$  Markov chain, the results of model comparison were predictably uninteresting. This is a good indication that the correct model class is being employed. However, with the even process a much more complicated model comparison was found. In this case, a selection of even  $k$  over odd hinted at the distinguishing properties of the source. In a similar way, the results of model comparison for the simple nondeterministic source selected increasing order with larger  $N$ . In both out-of-class modeling examples, the increase in selected order without end is a good indication that the data source is not in the Markov chain class. (A parallel technique is found in *hierarchical  $\epsilon$ -machine reconstruction* [22].) Alternatively, there is an indication that very high-order dependencies are important in the description of the process. Either way, this information is important since it gives an indication to the modeler that a more complicated dynamic is at work and all results must be treated with caution.

Finally, we considered the estimation of entropy rates for the example data sources. In two of the cases, the golden mean process and the simple nondeterministic source, short data streams were adequate. This is not unexpected for the golden mean, but for the simple nondeterministic source this might be considered surprising.

For the even process, the estimation of the entropy rate was markedly more difficult. For this data source, the countably infinite number of forbidden words makes the support of the word distribution at a given length important. As a result, a larger amount of data and a higher-order Markov chain are needed to find a decent estimate of randomness from that data source. In this way, each of the steps in Bayesian inference allow one to separate structure from randomness.

## VII. CONCLUSION

We considered Bayesian inference of  $k$ -th order Markov chain models. This included estimating model parameters for a given  $k$ , model comparison between orders, and estimation of randomness in the form of entropy rates. In most approaches to inference, these three aspects are treated as separate, but related endeavors. However, we find them to be intimately related. An estimate of model parameters without a sense of whether the correct model is being used is misguided at best. Model comparison provides a window into this problem by comparing various orders  $k$  within the model class. Finally, estimating randomness in the form of an entropy rate provides more information about the trade-off between structure and randomness. To do this we developed a connection to the statistical mechanical partition function, from which averages and variances were directly calculable. For the even process, structure was perceived as randomness and for the simple nondeterministic source randomness was easily estimated and structure was more difficult to find. These insights, despite the out-of-class data, demonstrate the power of combining these three methods into one effective tool for investigating structure and randomness in finite strings of discrete data.

## Acknowledgments

This work was partially supported at the Center for Computational Science and Engineering at the University of California at Davis by Intel Corporation. Work at the Santa Fe Institute was supported under its Computation, Dynamics, and Inference Program core grants from the National Science and MacArthur Foundations. C.S. and A.H. acknowledge support by the National Science Foundation Grant DMS 03-25939 ITR.

## APPENDIX A

### 1. Dirichlet Distribution

We supply a brief overview of the Dirichlet distribution for completeness. For more information, a reference such as [25] should be consulted. In simple terms, the Dirichlet distribution is the multinomial generalization of the

Beta distribution. The probability density function for  $q$  elements is given by

$$\text{Dir}(\{p_i\}) = \frac{\Gamma(\alpha)}{\prod_{i=0}^{q-1} \Gamma(\alpha_i)} \delta\left(1 - \sum_{i=0}^{q-1} p_i\right) \prod_{i=0}^{q-1} p_i^{\alpha_i - 1}. \quad (\text{A1})$$

The variates must satisfy  $p_i \in [0, 1]$  and  $\sum_{i=0}^{q-1} p_i = 1$ . The hyperparameters  $\{\alpha_i\}$  of the distribution, must be real and positive and we use the notation  $\alpha = \sum_{i=0}^{q-1} \alpha_i$ . The average, variance, and covariance of the parameters  $p_i$  are given by, respectively,

$$\mathbf{E}[p_j] = \frac{\alpha_j}{\alpha}, \quad (\text{A2})$$

$$\mathbf{Var}[p_j] = \frac{\alpha_j (\alpha - \alpha_j)}{\alpha^2 (1 + \alpha)}, \quad (\text{A3})$$

$$\mathbf{Cov}[p_j, p_l] = -\frac{\alpha_j \alpha_l}{\alpha^2 (1 + \alpha)}, \quad j \neq l. \quad (\text{A4})$$

## 2. Marginal distributions

An important part of understanding uncertainty in the inference process is the ability to find regions of confidence from a marginal density. The marginal is obtained

from the posterior by integrating out the dependence on all parameters except for the parameter of interest. For a Dirichlet distribution, the marginal density is known to be a Beta distribution [25],

$$\text{Beta}(p_i) = \frac{\Gamma(\alpha)}{\Gamma(\alpha_i)\Gamma(\alpha - \alpha_i)} p_i^{\alpha_i - 1} (1 - p_i)^{\alpha - \alpha_i - 1}. \quad (\text{A5})$$

## 3. Regions of confidence from the marginal density

From the marginal density provided in Eq. (A5) a cumulative distribution function can be obtained using the incomplete Beta integral

$$\Pr(p_i \leq x) = \int_0^x dp_i \text{Beta}(p_i). \quad (\text{A6})$$

Using this form, the probability that a Markov chain parameter will be between  $a$  and  $b$  can be found using  $\Pr(a \leq p_i \leq b) = \Pr(p_i \leq b) - \Pr(p_i \leq a)$ . For a confidence level  $R$ , between zero and one, we then want to find  $(a, b)$  such that  $R = \Pr(a \leq p_i \leq b)$ . The incomplete Beta integral and its inverse can be found using computational methods, see [26–29] for details.

- 
- [1] P. J. Avery and D. A. Henderson, *Appl. Stat.* **48**, 53 (1999).
- [2] J. S. Liu and C. E. Lawrence, *Bioinformatics* **15**, 38 (1999).
- [3] J. P. Crutchfield and D. P. Feldman, *Phys. Rev. E* **55**, R1239 (1997).
- [4] D. J. C. MacKay and L. C. B. Peto, *Nat. Lang. Eng.* **1** (1994).
- [5] J. P. Crutchfield and N. H. Packard, *Physica D* **7D**, 201 (1983).
- [6] B.-L. Hao and W.-M. Zheng, *Applied Symbolic Dynamics and Chaos* (World Scientific, 1998).
- [7] T. W. Anderson and L. A. Goodman, *Ann. Math. Stat.* **28**, 89 (1957).
- [8] P. Billingsley, *Ann. Math. Stat.* **32**, 12 (1961).
- [9] C. Chatfield, *Appl. Stat.* **22**, 7 (1973).
- [10] H. Tong, *Jour. Appl. Prob.* **12**, 488 (1975).
- [11] R. W. Katz, *Technometrics* **23**, 243 (1981).
- [12] J. Rissanen, *IEEE Trans. Inform. Theory* **30**, 629 (1984).
- [13] V. Vapnik, *IEEE Trans. Neur. Net.* **10**, 988 (1999).
- [14] P. M. Vitányi and M. Li, *IEEE Trans. Inform. Theory* **46(2)**, 446 (2000).
- [15] P. Baldi and S. Brunak, *Bioinformatics: The Machine Learning Approach* (MIT Press, Cambridge, 2001).
- [16] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis* (Cambridge University Press, Cambridge, 1998).
- [17] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley-Interscience, New York, 1991).
- [18] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms* (Cambridge University Press, Cambridge, 2003).
- [19] I. Samengo, *Phys. Rev. E* **65**, 46124 (2002).
- [20] K. Young and J. P. Crutchfield, *Chaos, Solitons, and Fractals* **4**, 5 (1994).
- [21] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1965).
- [22] J. P. Crutchfield, *Physica D* **75**, 11 (1994).
- [23] D. R. Upper, Ph.D. thesis, University of California, Berkeley (1997), Published by University Microfilms Intl, Ann Arbor, Michigan.
- [24] D. Blackwell and L. Koopmans, *Ann. Math. Stat.* **28**, 1011 (1957).
- [25] S. S. Wilks, *Mathematical Statistics* (John Wiley & Sons, Inc., New York, 1962).
- [26] K. Majumder and G. Bhattacharjee, *Appl. Stat.* **22**, 411 (1973).
- [27] K. Majumder and G. Bhattacharjee, *Appl. Stat.* **22**, 409 (1973).
- [28] G. Cran, K. Martin, and G. Thomas, *Appl. Stat.* **26**, 111 (1977).
- [29] K. Berry, P.W. Mielke, Jr., and G. Cran, *Appl. Stat.* **39**, 309 (1990).