Bayesian Analysis (2024) **19**, Number 2, pp. 501–529

Posterior Representations for Bayesian Context Trees: Sampling, Estimation and Convergence*

Ioannis Papageorgiou† and Ioannis Kontoyiannis‡

Abstract. We revisit the Bayesian Context Trees (BCT) modelling framework for discrete time series, which was recently found to be very effective in numerous tasks including model selection, estimation and prediction. A novel representation of the induced posterior distribution on model space is derived in terms of a simple branching process, and several consequences of this are explored in theory and in practice. First, it is shown that the branching process representation leads to a simple variable-dimensional Monte Carlo sampler for the joint posterior distribution on models and parameters, which can efficiently produce independent samples. This sampler is found to be more efficient than earlier MCMC samplers for the same tasks. Then, the branching process representation is used to establish the asymptotic consistency of the BCT posterior, including the derivation of an almost-sure convergence rate. Finally, an extensive study is carried out on the performance of the induced Bayesian entropy estimator. Its utility is illustrated through both simulation experiments and real-world applications, where it is found to outperform several state-of-the-art methods.

Keywords: discrete time series, Bayesian context trees, branching processes, exact sampling, consistency, model selection, prediction, entropy estimation, context-tree weighting.

1 Introduction

The statistical modelling and analysis of discrete time series are important scientific and engineering tasks, with a very wide range of applications. Numerous Markovian model classes have been developed in connection with these and related problems, including mixture transition distribution (MTD) models (Raftery, 1985; Berchtold and Raftery, 2002), variable-length Markov chains (VLMC) (Bühlmann and Wyner, 1999; Bühlmann, 2000; Mächler and Bühlmann, 2004) and sparse Markov chains (Jääskinen et al., 2014; Xiong et al., 2016). Alternative approaches also include the use of multinomial logit or probit regression (Zeger and Liang, 1986), categorical regression models (Fokianos and Kedem, 2003), and conditional tensor factorisation (Sarkar and Dunson, 2016).

A popular and useful class of relevant models for discrete time series are the *context-tree sources*, introduced by Rissanen (1983a, 1983b, 1986) as descriptions of variable-

*I. K. was supported in part by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "First Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment grant," project number 1034.

†Department of Engineering, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, UK, ip307@cam.ac.uk

‡Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, Wilberforce Road, Cambridge CB3 0WB, UK, yiannis@maths.cam.ac.uk

 https://doi.org/10.1214/23-BA1362

memory Markov chains, a flexible and rich class of chains that admit parsimonious representations. Their key feature is that the memory length of the chain is allowed to depend on the most recently observed symbols, providing a richer model class than ordinary Markov chains. Context-tree sources have been very successful in information-theoretic applications in connection with data compression (Weinberger et al., 1994; Willems et al., 1995), and the celebrated *context tree weighting* (CTW) algorithm (Willems et al., 1995; Willems, 1998), also based on context-tree sources, has been used widely as an efficient compression method with extensive theoretical guarantees and justifications.

Recently, Kontoyiannis et al. (2022) revisited context-tree models and the CTW algorithm from a Bayesian inference point of view. A general modelling framework, called *Bayesian Context Trees* (BCT), was developed for discrete time series, along with a collection of efficient algorithmic tools both for exact inference and for posterior sampling via Markov chain Monte Carlo (MCMC). The BCT methods were found to be very effective in important statistical tasks, including model selection, estimation, prediction and change-point detection (Papageorgiou et al., 2021; Lungu et al., 2022a,b); see also the R package BCT (Papageorgiou et al., 2020).

In this work we derive an alternative representation of the posterior distribution induced by the BCT framework, and explore several ways in which it facilitates inference, both in theory and in practice. Our first main contribution is described in Sections 3.1 and 3.2, where we show that both the prior and posterior distributions on model space admit explicit representations as simple branching processes. In particular, sampling tree models from the prior or the posterior is shown to be equivalent to generating trees via an appropriate Galton-Watson process (Athreya and Ney, 2004; Harris, 1963), stopped at a given depth. Therefore, in some sense the BCT model prior acts as a 'conjugate' prior for variable-memory Markov chains. An immediate first practical consequence of this representation is that it facilitates direct Monte Carlo (MC) sampling from the posterior, where independent and identically distributed (i.i.d.) samples can be efficiently obtained from the joint posterior distribution on models and parameters. This variable-dimensional sampler and its potential utility in a wide range of applications (including model selection, parameter estimation and Markov order estimation) are described in Section 3.3.

The Bayesian perspective adopted in this work is neither purely subjective nor purely objective. For example, we think of the model posterior distribution as a summary of the most accurate, data-driven representation of the regularities present in a given time series, but we also examine the frequentist properties of the resulting inferential procedures (Gelman et al., 1995; Chipman et al., 2001; Bernardo and Smith, 2009). Indeed, in Section 4 we employ the branching process representation to show that the posterior asymptotically almost surely concentrates on the "true" underlying model (Theorem 4.1), and in Theorem 4.2 we derive an explicit rate for this convergence as a function of the sample size. Analogous results are established in Theorems 4.3 and 4.4 in the case of out-of-class modelling, when the data are generated by a model outside the BCT class. Importantly, the limiting model is explicitly identified in this case. The branching process representation is also used in Proposition 4.1 to provide a simple, explicit representation of the posterior predictive distribution. These theoretical results are the second main contribution of this work.

Our last contribution, in Section 5, is a brief experimental evaluation of the utility of the MC sampler of Section 3, and a careful examination of the performance of the induced Bayesian *entropy estimator*. In Section 5.1, the new i.i.d. sampler is compared with the MCMC samplers introduced by Kontoyiannis et al. (2022) on simulated data. As expected, it is found that the i.i.d. sampler has superior performance, both in terms of estimation accuracy and, as expected, in terms of mixing.

Finally, in Section 5.2 we consider the important problem of estimating the entropy rate of a discrete time series. Starting with the original work of Shannon (1951), many different approaches have been developed for this task, including Lempel-Ziv (LZ) estimators (Ziv and Lempel, 1977; Wyner and Ziv, 1989), prediction by partial matching (PPM) (Cleary and Witten, 1984), the CTW algorithm (Gao et al., 2008), and block sorting methods (Cai et al., 2004); for an extensive review of the relevant literature, see Verdú (2019). Entropy estimation has also received a lot of attention in the neuroscience literature (Strong et al., 1998; London et al., 2002; Nemenman et al., 2004; Paninski, 2003), in an effort to describe and quantify the amount of information transmitted by neurons.

In contrast with most earlier work, here we adopt a fully-Bayesian approach. Since the entropy rate is a functional of the model and associated parameters, the branching process sampler of Section 3 makes it possible to effectively sample from (and hence estimate) the actual posterior distribution of the entropy rate. This of course provides a much richer picture than the simple point estimates employed in most applications. The performance of the BCT entropy estimator is illustrated both on simulated data and real-world applications from neuroscience, finance and animal communication, where it is seen to outperform several of the state-of-the-art methods.

In closing this introduction, we mention that there are, of course, numerous other approaches to the problem of inference for discrete time series. In addition to the extensive review given by Kontoyiannis et al. (2022), those include the class of reversible variable-memory chains examined by Bacallado (2011); Bacallado et al. (2013, 2016), and the Bayesian analyses of discrete models with priors that encourage sparse representations developed in Heiner et al. (2019); Heiner and Kottas (2022).

2 Bayesian context trees

In this section, we briefly review the BCT model class, the associated prior structure, and some relevant properties and results that will be needed in subsequent sections.

The BCT model class consists of variable-memory Markov chains, where the memory length of the process may depend on the values of the most recently observed symbols. Variable-memory Markov chains admit natural representations as context trees. Let $\{X_n\}$ be a dth order Markov chain, for some $d \geq 0$, taking values in the alphabet $A = \{0, 1, \ldots, m-1\}$. The *model* describing $\{X_n\}$ as a variable-memory chain is represented by a proper m-ary tree T as in the example in Figure 1, where a tree T is called *proper* if any node in T that is not a leaf has exactly m children.

Each leaf of the tree T corresponds to a string s determined by the sequence of symbols along the path from the root node λ to that leaf. At each leaf s, there is an associated set of parameters θ_s that form a probability vector, $\theta_s = (\theta_s(0), \theta_s(1), \ldots, \theta_s(m-1))$. At every time n, the conditional distribution of the next symbol X_n, given the past d observations $(x_{n-1}, \ldots, x_{n-d})$, is given by the vector θ_s associated to the unique leaf s of T that is a suffix of $(x_{n-1}, \ldots, x_{n-d})$. Throughout this paper, every variable-memory Markov chain is described by a tree model T and a set of associated parameters $\theta = \{\theta_s \ ; \ s \in T\}$, where T is viewed as the collection of its leaves.

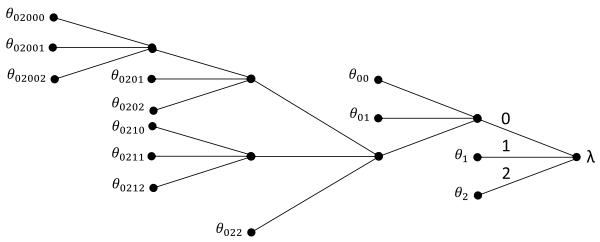


Figure 1: Tree model representation of a 5th order variable-memory chain.

Model prior Given a maximum depth $D \geq 0$, let $\mathcal{T}(D)$ denote the collection of all proper m-ary trees with depth no greater than D. In Kontoyiannis et al. (2022), the following prior distribution is introduced on $\mathcal{T}(D)$,

$$\pi(T) = \pi_D(T; \beta) = \alpha^{|T|-1} \beta^{|T|-L_D(T)}, \tag{2.1}$$

where $\beta \in (0,1)$ is a hyperparameter, α is given by $\alpha = (1-\beta)^{1/(m-1)}$, $|T|$ is the number of leaves of T, and $L_D(T)$ is the number of leaves of T at depth D.

This prior clearly penalises larger trees by an exponential amount, and larger values of β make the penalisation more severe. We adopt the default value $\beta = 1 - 2^{-m+1}$ for β; see Kontoyiannis et al. (2022) for an extensive discussion of the properties of this prior and the choice of β.

Prior on parameters Given a tree model $T \in \mathcal{T}(D)$, an independent Dirichlet prior with parameters $(1/2, 1/2, \ldots, 1/2)$ is placed on each θ_s, so that:

$$\pi(\theta|T) = \prod_{s \in T} \pi(\theta_s) = \prod_{s \in T} \mathrm{Dir}(1/2, 1/2, \ldots, 1/2). \tag{2.2}$$

Let $x = (x_{-D+1}, \ldots, x_0, x_1, \ldots, x_n)$ denote a time series with values in A. For each $i \leq j$, we write x_i^j for the segment $(x_i, x_{i+1}, \ldots, x_j)$, so that x consists of the observations x_1^n along with an initial context x_{-D+1}^0 of length D.

One of the main observations of Kontoyiannis et al. (2022) is that the prior predictive likelihood, averaged over both models and parameters,

$$P(x) = \sum_{T \in \mathcal{T}(D)} \int_\theta P(x|T, \theta) \pi(\theta|T) \pi(T) d\theta, \tag{2.3}$$

can be computed exactly and efficiently by a version of the CTW algorithm (where $P(x|T,\theta)$ denotes the probability of x under model T with parameters θ), which of course facilitates numerous important statistical tasks. For a given time series $x = x_{-D+1}^n$, the CTW algorithm uses the *estimated probabilities* $P_{e,s}$ defined as follows. For any tree model $T \in \mathcal{T}(D)$ and any context s (not necessarily a leaf),

$$P_{e,s} = P_e(a_s) = \frac{\prod_{j=0}^{m-1}[(1/2)(3/2)\cdots(a_s(j)-1/2)]}{(m/2)(m/2+1)\cdots(m/2+M_s-1)}, \tag{2.4}$$

where the elements of each *count vector* $a_s = (a_s(0), a_s(1), \ldots, a_s(m-1))$ are given by,

$$a_s(j) = \#\text{ times symbol } j \in A \text{ follows context } s \text{ in } x_1^n, \tag{2.5}$$

and $M_s = a_s(0) + a_s(1) + \cdots + a_s(m-1)$.

CTW: The context tree weighting algorithm

1. Build the tree T_{MAX}, which is the smallest proper tree that contains all the contexts x_{i-D+1}^i, $i = 1, 2, \ldots, n$, as leaves. Compute $P_{e,s}$ as given in (2.4) for each node s of T_{MAX}.
2. Starting at the leaves and proceeding recursively towards the root, for each node s of T_{MAX} compute the *weighted probabilities* $P_{w,s}$, given by,

$$P_{w,s} = \begin{cases} P_{e,s}, & \text{if } s \text{ is a leaf,} \\ \beta P_{e,s} + (1-\beta)\prod_{j=0}^{m-1} P_{w,sj}, & \text{otherwise,} \end{cases} \tag{2.6}$$

   where sj is the concatenation of context s and symbol j.

As shown in Kontoyiannis et al. (2022), the weighted probability $P_{w,\lambda}$ produced by the CTW algorithm at the root λ, is indeed exactly equal to the prior predictive likelihood in (2.3). A family of Markov chain Monte Carlo (MCMC) samplers for the posterior $\pi(T|x)$ or $\pi(T,\theta|x)$ were also introduced in Kontoyiannis et al. (2022). However, as will be seen below, the representation of Section 3 leads to a simple i.i.d. MC sampler that typically outperforms these MCMC samplers.

3 Branching process representations

In this section we show that the BCT model prior $\pi(T) = \pi_D(T;\beta)$ of (2.1) and the resulting posterior $\pi(T|x)$ both admit natural and easily interpretable representations in terms of simple branching processes. We also discuss how the posterior representation leads to efficient samplers that can be used for model selection and estimation.

3.1 The prior branching process

Given $D \geq 0$ and $\beta \in (0,1)$, let $T = \{\lambda\}$ consist of only the root node λ and consider the following procedure:

- If $D = 0$, stop.
- If $D > 0$, then, with probability β, mark the root as a leaf and stop, or, with probability $(1-\beta)$, add all m children of λ at depth 1 to T. If $D = 1$, stop.
- If $D > 1$, examine each of the m new nodes, and either mark a node as a leaf with probability β, or add all m of its children to T with probability $(1-\beta)$, independently from node to node.
- Continue recursively, at each step examining all non-leaf nodes at depths strictly smaller than D, until no more eligible nodes remain to be examined.
- Output the resulting tree $T \in \mathcal{T}(D)$.

The above construction is a simple Galton-Watson process (Athreya and Ney, 2004; Harris, 1963) with offspring distribution $(\beta, (1-\beta))$ on $\{0, m\}$, stopped at generation D. The following proposition states that the distribution of a tree T generated by this process is exactly the prior $\pi_D(T;\beta)$. Note that this also implies that the expression for $\pi_D(T;\beta)$ given in (2.1) indeed defines a probability distribution on $\mathcal{T}(D)$, giving an alternative proof of Kontoyiannis et al. (2022, Lemma 2.1).

Proposition 3.1. *For any $D \geq 0$ and any $\beta \in (0,1)$, the probability that the above branching process produces any particular tree $T \in \mathcal{T}(D)$ is given by $\pi_D(T;\beta)$ as in (2.1).*

Proof. When $D = 0$, $\mathcal{T}(D)$ consists of a single tree, $T = \{\lambda\}$, which has probability 1 under both π_D and the branching process construction. Assume $D \geq 1$. Note that every tree $T \in \mathcal{T}(D)$ can be viewed as a collection of a number, k, say, of m-branches, since every node in T has either zero or m children. The proposition is proven by induction on k. The result is trivial for $k = 0$, since the only tree with no m-branches is $T = \{\lambda\}$ and its probability under both π_D and the branching process construction is equal to β.

Suppose the claim of the proposition is true for all trees with k m-branches, and let $T' \in \mathcal{T}(D)$ consist of $(k+1)$ m-branches. Then T' can be obtained from some $T \in \mathcal{T}(D)$ that has k m-branches, by adding a single m-branch to one of its leaves, s, say. Two cases are considered.

(i) If s is at depth $D-2$ or smaller, then the probability $\pi_b(T')$ of T' under the branching process construction is,

$$\pi_b(T') = \frac{\pi_b(T)}{\beta}(1-\beta)\beta^m = \frac{\pi_D(T;\beta)}{\beta}(1-\beta)\beta^m = \frac{\alpha^{|T|-1}\beta^{|T|-L_D(T)}}{\beta}\alpha^{m-1}\beta^m,$$

where the second equality follows from the inductive hypothesis and the third from the definition of $\pi_D(T;\beta)$. Therefore, since $|T'| = |T| + m - 1$ and no leaves are added at depth D, so that $L_D(T') = L_D(T)$,

$$\pi_b(T') = \alpha^{[|T|+m-1]-1}\beta^{|T|+m-1-L_D(T)} = \alpha^{|T'|-1}\beta^{|T'|-L_D(T')} = \pi_D(T';\beta),$$

as required.

(*ii*) If s is at depth $D-1$, we similarly find that,

$$\pi_b(T') = \frac{\pi_b(T)}{\beta}(1-\beta) = \frac{\pi_D(T;\beta)}{\beta}(1-\beta) = \frac{\alpha^{|T|-1}\beta^{|T|-L_D(T)}}{\beta}\alpha^{m-1},$$

and since $|T'| = |T| + m - 1$, but now, $L_D(T') = L_D(T) + m$,

$$\pi_b(T') = \alpha^{[|T|+m-1]-1}\beta^{[|T|+m-1]-[L_D(T)+m]} = \alpha^{|T'|-1}\beta^{|T'|-L_D(T')} = \pi_D(T';\beta),$$

completing the proof. □

Apart from being aesthetically appealing, this representation also offers a simple and practical way of sampling from $\pi_D(T;\beta)$. Moreover, using well-known properties of the Galton-Watson process we can perform some direct computations that offer better insight into the nature and specific properties of the BCT prior.

Interpretation and choice of β The branching process description of $\pi_D(T,\beta)$ further clarifies the role of the hyperparameter β: It is exactly the probability that, when a node is added to the tree T, it is marked as a leaf and its children are not included in T.

In terms of choosing the value of the hyperparameter β appropriately, recall that, for a Galton-Watson process, the expected number of children of each node, in this case $\rho = m(1-\beta)$, governs the probability of extinction P_e: If $\rho \le 1$ we have $P_e = 1$, whereas if $\rho > 1$, P_e is strictly less than one. Therefore, in the binary case $m = 2$, the original choice $\beta = 1/2$ used in the CTW algorithm gives an expected number of children equal to the critical value $\rho = 1$. This suggests that a reasonable choice for general alphabets could be $\beta = 1 - 1/m$, which keeps $\rho = 1$, so that the resulting prior would have similar qualitative characteristics with the well-studied binary case. This is also in line with the observation of Kontoyiannis et al. (2022) that β should decrease with m.

Now suppose T is a random model generated by the prior and let $L_d(T)$ denote the number of nodes at depth $d = 0, 1, \ldots, D$. Then, standard Galton-Watson theory (Athreya and Ney, 2004; Harris, 1963) provides the useful expressions,

$$\mathbb{E}\left[L_d(T)\right] = \rho^d, \qquad \mathrm{Var}\left[L_d(T)\right] = \begin{cases} \sigma^2 d, & \text{if } \rho = 1, \\ \sigma^2 \rho^{d-1}\frac{1-\rho^d}{1-\rho}, & \text{if } \rho \neq 1, \end{cases}$$

where $\sigma^2 = m^2\beta(1-\beta)$.

3.2 The posterior branching process

Given a time series $x = x_{-D+1}^{n}$, a maximum depth D, and $\beta \in (0,1)$, for any context s with length strictly smaller than D we define the *branching probabilities* $P_{b,s}$ as,

$$P_{b,s} := \frac{\beta P_{e,s}}{P_{w,s}}, \tag{3.1}$$

where the estimated and weighted probabilities, $P_{e,s}$ and $P_{w,s}$, are defined in (2.4)-(2.6), and with the convention that $P_{b,s} = \beta$ for all contexts s that do not appear in x. Starting with $T = \{\lambda\}$, the following construction produces a sample model $T \in \mathcal{T}(D)$ from the posterior $\pi(T|x)$:

- If $D = 0$, stop.
- If $D > 0$, then, with probability $P_{b,\lambda}$, mark the root as a leaf and stop, or, with probability $(1 - P_{b,\lambda})$, add all m children of λ at depth 1 to T. If $D = 1$, stop.
- If $D > 1$, examine each of the m new nodes and either mark a node s as a leaf with probability $P_{b,s}$, or add all m of its children to T with probability $(1 - P_{b,s})$, independently from node to node.
- Continue recursively, at each step examining all non-leaf nodes at depths strictly smaller than D, until no more eligible nodes remain.
- Output the resulting tree $T \in \mathcal{T}(D)$.

Proposition 3.2. *For any $D \geq 0$ and any $\beta \in (0,1)$, the probability that the above branching process produces any particular tree $T \in \mathcal{T}(D)$ is given by $\pi(T|x)$.*

The proof, which follows along the same lines as that of Proposition 3.1, is given in Section A of the Supplementary Material (Papageorgiou and Kontoyiannis, 2023). It is perhaps somewhat remarkable that the posterior $\pi(T|x)$ on the vast model space $\mathcal{T}(D)$ admits such a simple description. Indeed, the posterior branching process is of exactly the same form as that of the prior, which can then naturally be viewed as a conjugate prior on $\mathcal{T}(D)$.

Model posterior probabilities Proposition 3.2 allows us to write an exact expression for the posterior of any model $T \in \mathcal{T}(D)$ in terms of the branching probabilities $P_{b,s}$,

$$\pi(T|x) = \prod_{s \in T_o} (1 - P_{b,s}) \prod_{s \in T} P_{b,s}, \tag{3.2}$$

where T_o denotes the set of all *internal* nodes of T, and with the convention that $P_{b,s} = 1$ for all leaves of T at depth $d = D$. This expression will be the starting point in the proofs of the asymptotic results of Section 4 for $\pi(T|x)$.

In terms of inference, the main utility of Proposition 3.2 is that it offers a practical way of obtaining exact i.i.d. samples directly from the model posterior, as described in the next section.

3.3 Sampling from the posterior

The branching process representation of $\pi(T|x)$ readily leads to a simple way for obtaining i.i.d. samples $\{T^{(i)}\}$ from the posterior on model space. And since the full conditional density of the parameters $\pi(\theta|T,x)$ is explicitly identified by Kontoyiannis et al. (2022) as a product of Dirichlet densities,

$$\pi(\theta|T,x) = \prod_{s\in T} \mathrm{Dir}\left(1/2 + a_s(0), 1/2 + a_s(1), \ldots, 1/2 + a_s(m-1)\right), \tag{3.3}$$

for each $T^{(i)}$ we can draw a conditionally independent sample $\theta^{(i)} \sim \pi(\theta|T^{(i)},x)$, producing a sequence of exact i.i.d. samples $\{(T^{(i)},\theta^{(i)})\}$ from the joint posterior $\pi(T,\theta|x)$.

This facilitates numerous applications. For example, effective parameter estimation can be performed by simply keeping the samples $\{\theta^{(i)}\}$, which come from the marginal posterior distribution $\pi(\theta|x)$. Similarly, Markov order estimation can be performed by collecting the sequence of maximum depths of the models $\{T^{(i)}\}$. And in model selection tasks, the model posterior can be extensively explored, offering better insight and deeper understanding of the underlying structure and dependencies present in the data.

Although a family of MCMC samplers was introduced and successfully used for the same tasks in Kontoyiannis et al. (2022), MCMC sampling has well-known limitations and drawbacks, including potentially slow mixing, high correlation between samples, and the need for convergence diagnostics (Gelman and Rubin, 1992; Cowles and Carlin, 1996; Robert and Casella, 2004). Partly for these reasons, being able to obtain i.i.d. samples from the posterior is generally much more desirable, as illustrated in Section 5.1.

Estimation of general functionals Consider the general Bayesian estimation problem, where the goal is to estimate an arbitrary functional $F = F(T,\theta)$ of the underlying variable-memory chain, based on data x. Using the above sampler, the entire posterior distribution of the statistic F can be explored, by considering the i.i.d. samples $F^{(i)} = F(T^{(i)},\theta^{(i)})$, distributed according to the desired posterior $\pi(F|x)$.

In connection with classical estimation techniques, and in order to evaluate estimation performance more easily in practice, several reasonable point estimates can also be obtained. The most common choices are the empirical average approximation to the posterior mean,

$$\widehat{F}_{\mathrm{MC}} = \frac{1}{N}\sum_{i=1}^{N} F^{(i)} = \frac{1}{N}\sum_{i=1}^{N} F(T^{(i)},\theta^{(i)}), \tag{3.4}$$

or the posterior mode, i.e., the maximum *a posteriori* probability (MAP) estimate, $\widehat{F}_{\mathrm{MAP}}$. In cases where the conditional mean $\bar{F}(T) = E(F(T,\theta)|x,T)$ can be computed for any model T (as e.g. in the case of parameter estimation), a lower-variance Rao-Blackwellised estimate (Blackwell, 1947; Gelfand and Smith, 1990) for the posterior mean can also be obtained as, $\widehat{F}_{\mathrm{RB}} = \frac{1}{N}\sum_{i=1}^{N} \bar{F}(T^{(i)})$.

Importantly, as this posterior sampler provides access to the entire posterior distribution $\pi(F|x)$ of the statistic of interest F, standard Bayesian methodology can be

applied to quantify the resulting uncertainty of any estimator $\hat{F}$, for example by obtaining credible intervals in terms of the $\pi(F|x)$.

An interesting special case of particular importance in practice is the estimation of the *entropy rate* H of the underlying process, $F(T,\theta) = H(T,\theta)$. The performance of all methods discussed above, with emphasis on the estimation of the entropy rate, is illustrated through simulated experiments and real-world applications in Section 5.2.

4 Theoretical results

Using the branching process representation of Proposition 3.2, we show how to derive precise results on the asymptotic behaviour of the BCT posterior $\pi(T|x)$ on model space, and provide an explicit, useful expression for the posterior predictive distribution.

Let $\{X_n\}$ be a variable-memory chain with model $T \in \mathcal{T}(D)$. The specific model T that describes the chain is typically not unique, for the same reason, e.g., that every i.i.d. process can also trivially be described as a first-order Markov chain: Adding m children to any leaf of T which is not at maximal depth, and giving each of them the same parameters as their parent, leaves the distribution of the chain unchanged. The natural main goal in model selection is to identify the "minimal" model, i.e., the smallest model that can fully describe the distribution of the chain. A model $T \in \mathcal{T}(D)$ is called *minimal* if every m-tuple of leaves $\{sj\ ;\ j = 0, 1, \ldots, m-1\}$ in T contains at least two with non-identical parameters, i.e., there are $j \neq j'$ such that $\theta_{sj} \neq \theta_{sj'}$. It is easy to see that every Dth order Markov chain $\{X_n\}$ has a unique minimal model $T^* \in \mathcal{T}(D)$.

A variable-memory chain $\{X_n\}$ with model $T \in \mathcal{T}(D)$ and associated parameters $\theta = \{\theta_s; s \in T\}$ is *ergodic*, if the corresponding first-order chain $\{Z_n := X_{n-D+1}^n; n \geq 1\}$ taking values in A^D is irreducible and aperiodic. In order to avoid uninteresting technicalities, in most of our results we will assume that the data are generated by a *positive-ergodic* chain $\{X_n\}$, namely that all its parameters $\theta_s(j)$ are nonzero, so that its unique stationary distribution π gives strictly positive probability to all finite contexts s.

4.1 Posterior consistency and concentration

Our first theorem is a strong consistency result, which states that, if the data $x = x_{-D+1}^n$ are generated by an ergodic chain with minimal model $T^* \in \mathcal{T}(D)$, then the model posterior asymptotically almost surely (a.s.) concentrates on T^*. Theorem 4.1 both strengthens and generalises a weaker result on the asymptotic behaviour of the MAP model established in Willems et al. (1993, Theorem 8).

Theorem 4.1. *Let $X_{-D+1}^n = (X_{-D+1}, \ldots, X_0, X_1, \ldots, X_n)$ be a time series generated by a positive-ergodic, variable-memory chain $\{X_n\}$ with minimal model $T^* \in \mathcal{T}(D)$. For any value of the prior hyperparameter $\beta \in (0,1)$, the posterior distribution over models concentrates on T^*, i.e.,*

$$\pi(T^*|X_{-D+1}^n) \to 1, \quad \textit{a.s., as } n \to \infty.$$

Proof. Recalling the posterior representation in (3.2), it suffices to show that, as $n \to \infty$, $P_{b,s} \to 0$ a.s. for all internal nodes of T^*, and $P_{b,s} \to 1$ a.s. for all leaves of T^*. These two claims are established in Lemmas 4.2 and 4.3 below. □

We first recall the following simple bounds on the estimated probabilities $P_{e,s}$; see Krichevsky and Trofimov (1981); Xie and Barron (2000) and Catoni (2004, Ch. 1).

Lemma 4.1. *For every node s with count vector $a_s = (a_s(0), a_s(1), \ldots, a_s(m-1))$ and $M_s = a_s(0) + \cdots + a_s(m-1)$, the estimated probabilities $P_{e,s}$ of (2.4) satisfy:*

$$\log P_{e,s} \geq \sum_{j=0}^{m-1} a_s(j) \log \frac{a_s(j)}{M_s} - \frac{m-1}{2} \log M_s - \log m; \tag{4.1}$$

$$\log P_{e,s} \leq \sum_{j=0}^{m-1} a_s(j) \log \frac{a_s(j)}{M_s} - \frac{m-1}{2} \log \frac{M_s}{2\pi} - \log \frac{\pi^{m/2}}{\Gamma(m/2)}. \tag{4.2}$$

[Throughout the paper, $\log \equiv \log_e$ denotes the natural logarithm.] Lemma 4.2 is a generalisation of Jiao et al. (2013, Lemma 12).

Lemma 4.2. *Under the assumptions of Theorem 4.1, for every internal node s of T^*, the branching probability $P_{b,s} \to 0$, a.s., as $n \to \infty$.*

Proof. As the complete proof is quite involved, only the main and more interesting part of the argument is given here, with the remaining details given in Section B.1 of the Supplementary Material.

We begin by observing that,

$$P_{b,s} = \frac{\beta P_{e,s}}{P_{w,s}} = \frac{\beta P_{e,s}}{\beta P_{e,s} + (1-\beta) \prod_j P_{w,sj}} = \frac{1}{1 + (1-\beta)/\beta \prod_j P_{w,sj}/P_{e,s}} \tag{4.3}$$

$$\leq \frac{1}{1 + c_0 \prod_j P_{e,sj}/P_{e,s}}, \tag{4.4}$$

for some constant c_0, where in the last step we used that either $P_{w,sj} \geq \beta P_{e,sj}$ or $P_{w,sj} = P_{e,sj}$. Therefore, it suffices to show that $P_{e,s} \big/ \prod_j P_{e,sj} \to 0$, a.s., as $n \to \infty$.

Let s be a fixed finite context, and let X and J denote the random variables corresponding to the symbols that follow and precede s, respectively, under the stationary distribution of $\{X_n\}$. Using Lemma 4.1 and the ergodic theorem for Markov chains, it is shown in Section B.1 of the Supplementary Material that,

$$\log P_{e,s} - \sum_j \log P_{e,sj} = -nI(X; J|s)\pi(s) + o(n), \quad \text{a.s.}, \tag{4.5}$$

where $I(X; J|s)$ is the conditional mutual information between X and J given s (Cover and Thomas, 2012, Ch. 2). This mutual information is always nonnegative, and it is zero if and only if X and J are conditionally independent given s.

For any internal node s that is a parent of leaves of T^*, the minimality of T^* implies that X and J are not conditionally independent given s, as there exist $j \neq j'$ such that $\theta_{sj} \neq \theta_{sj'}$, so that θ_{sj} depends on j. Therefore, $I(X;J|s) > 0$ and $\pi(s) > 0$ by assumption, so (4.5) implies that $\log P_{e,s} - \sum_j \log P_{e,sj} \to -\infty$, a.s., as required.

For the general case of internal nodes that may not be parents of leaves, a simple iterative argument is given in Section B.1 of Supplementary Material establishing the same result in that case as well, and completing the proof of the lemma. □

Lemma 4.3. *Under the assumptions of Theorem 4.1, for every leaf s of T^*, the branching probability $P_{b,s} \to 1$, a.s., as $n \to \infty$.*

Proof. As with the previous lemma, we only give an outline of the main interesting steps in the proof here; complete details are provided in Section B.2 of the Supplementary Material.

By definition, for any leaf s of T^* (and also for any 'external' node s, that is, any context s not in T^*), we have $I(X;J|s) = 0$ because of conditional independence. Therefore, we need to consider the higher-order terms in the asymptotic expansion of (4.5). Using Lemma 4.1, the ergodic theorem, and the law of the iterated logarithm (LIL) for Markov chains, it is shown in Section B.2 of the Supplementary Material that here,

$$\sum_j \log P_{e,sj} - \log P_{e,s} \leq -\frac{(m-1)^2}{2} \log n + O(\log \log n), \quad \text{a.s.} \tag{4.6}$$

This implies that $\sum_j \log P_{e,sj} - \log P_{e,s} \to -\infty$, so that $\prod_j P_{e,sj}/P_{e,s} \to 0$, a.s.

The last step of the proof, namely that for any leaf s of T^*, $\prod_j P_{e,sj}/P_{e,s} \to 0$ also implies that $\prod_j P_{w,sj}/P_{e,s} \to 0$, a.s., so that, by (4.3), $P_{b,s} \to 1$, a.s., is given in Section B.2 of the Supplementary Material. □

Our next result is a refinement of Theorem 4.1, which characterises the rate at which the posterior probability of T^* converges to 1.

Theorem 4.2. *Let X_{-D+1}^n be a time series generated by a positive-ergodic variable-memory chain $\{X_n\}$ with minimal model $T^* \in \mathcal{T}(D)$. For any value of the prior hyperparameter $\beta \in (0,1)$ and any $\epsilon > 0$, we have, as $n \to \infty$:*

$$\pi(T^*|X_{-D+1}^n) = 1 - O\left(n^{-\frac{(m-1)^2}{2}+\epsilon}\right), \quad \text{a.s.}$$

The proof of Theorem 4.2 is given in Section B.3 of the Supplementary Material. In fact, as discussed at the end of Section B.3, the proof also reveals that a stronger statement can be made about the rate in the case of full Dth order Markov chains:

Corollary 4.1. *If $\{X_n\}$ is a genuinely Dth order chain in that its minimal model T^* is the complete tree of depth D, then its posterior probability $\pi(T^*|X_{-D+1}^n)$ almost surely converges to 1 at an exponential rate.*

4.2 Out-of-class modelling

In this section we consider the behaviour of the posterior distribution $\pi(T|x)$ on models $T \in \mathcal{T}(D)$ when the time series x is *not* generated by a chain from the model class $\mathcal{T}(D)$, but from a general stationary and ergodic process $\{X_n\}$ with possibly infinite memory. We first give an explicit description of the "limiting" model $T_\infty \in \mathcal{T}(D)$ on which the posterior $\pi(T|x)$ concentrates when the observations are generated by a general process outside $\mathcal{T}(D)$, and then we give conditions under which T_∞ is structurally "as close as possible" to the true underlying model.

Description of T_∞ Recall that, for any context s, we write X and J for the random variables corresponding to the symbols that follow and precede s, respectively. The limiting tree $T_\infty \in \mathcal{T}(D)$ corresponding to a general stationary process $\{X_n\}$ with values in A can be constructed via the following procedure:

- Take T_∞ to be the empty tree.
- Starting with the nodes at depth $d = D-1$, for each such s, if $I(X;J|s) > 0$, then add s to T_∞ along with all its children and all its ancestors; that is, add the complete path from the root λ to the children of s.
- After all nodes s at depth $d = D-1$ have been examined, examine all possible nodes s at depth $d = D-2$ that are not already included in T_∞, and repeat the same process.
- Continue recursively towards the root, until all nodes at all depths $0 \le d \le D-1$ have been examined.
- For any node already in T_∞ at depth $d \le D-1$, such that only some but not all m of its children are included in T_∞, add the missing children to T_∞ so that it becomes proper.
- Output T_∞.

In order to state our results we need to impose two additional conditions on the underlying data-generating process. Suppose $\{X_n\}$ is stationary. Without loss of generality (by Kolmogorov's extension theorem) we may consider the two-sided version of the process, $\{X_n;\ n \in \mathbb{Z}\}$. Its α-mixing coefficients (Ibragimov, 1962; Philipp and Stout, 1975) are defined as,

$$\alpha_n = \sup_{A\in\mathcal{A},\ B\in\mathcal{B}} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|, \tag{4.7}$$

where $\mathcal{A}$ and $\mathcal{B}$ denote the σ-algebras, $\sigma(\ldots, X_{-1}, X_0)$ and $\sigma(X_n, X_{n+1}, \ldots)$, respectively. We will need a mixing condition and a positivity condition for our results:

$$\sum_{n=1}^{\infty} \alpha_n^{\delta/(2+\delta)} < \infty \text{ for some } \delta > 0, \quad \text{and} \quad \mathbb{P}(X_0^D = x_0^D) > 0 \text{ for all } x_0^D \in A^{D+1}. \tag{4.8}$$

Theorem 4.3. *Let X_{-D+1}^n be a time series generated by a stationary ergodic process $\{X_n\}$ satisfying the assumptions (4.8), and let $T_\infty \in \mathcal{T}(D)$ be given by the above construction. Then, for any value of the prior hyperparameter $\beta \in (0,1)$ we have:*

$$\pi(T_\infty | X_{-D+1}^n) \to 1, \quad a.s., \text{ as } n \to \infty.$$

The proof of Theorem 4.3 follows along exactly the same lines as the earlier proof of Theorem 4.1. Instead of the ergodic theorem for Markov chains (Chung, 1967, p. 92) we now use Birkhoff's ergodic theorem (Breiman, 1992, Ch. 6), and instead of the LIL for Markov chains we apply the general LIL for functions of blocks of an ergodic process, which follows, as usual, from the almost-sure invariance principle (Philipp and Stout, 1975; Rio, 1995; Zhao and Woodroofe, 2008). The mixing condition in (4.8) was chosen as one of the simplest ones that guarantee this general version of the LIL.

Regarding the overall structure of the proof, all the earlier asymptotic expansions of the branching probabilities still remain valid, including (4.5) and (4.6). The only possible difference might be at the boundary conditions that are required as a starting point for the iterative argument in the proof of Lemma 4.2, since the actual "leaves" of the true underlying model of $\{X_n\}$ are not necessarily at depth $d \leq D$ here. But, as before, all branching probabilities $P_{b,s}$ tend either to 0 or to 1, depending on whether the mutual information condition that appears in the description of T_∞ holds or not. Specifically, starting from nodes s at depth $d = D-1$, if $I(X;J|s) = 0$ then we are in the same situation as in Lemma 4.3, so that $P_{b,s} \to 1$, and all children of s are pruned. On the other hand, if $I(X;J|s) > 0$, then $P_{b,s} \to 0$, and by the same iterative argument, $P_{b,u} \to 0$ for all ancestors u of node s as well.

Analogous comments apply to the proof of Theorem 4.4, which is again a refinement characterising the rate at which the posterior probability of T_∞ converges to 1.

Theorem 4.4. *Let X_{-D+1}^n be a time series generated by a stationary ergodic process $\{X_n\}$ satisfying the assumptions (4.8), and let T_∞ be its limiting model in $\mathcal{T}(D)$. Then, for any $\beta \in (0,1)$ and any $\epsilon > 0$, as $n \to \infty$ we have:*

$$\pi(T_\infty | X_{-D+1}^n) = 1 - O\left(n^{-\frac{(m-1)^2}{2}+\epsilon}\right), \quad a.s.$$

In general, it is natural to expect that T_∞ should be "as close as possible" in some sense to the true underlying model T^*, and this is indeed what is most often observed in applications: T_∞ being the same as T^* truncated to depth D. But it is not always the case. For example, recall the 3rd order chain $\{X_n\}$ considered in Example 5.2 of Kontoyiannis et al. (2022), also described as having a "bimodal posterior" in Section 5.2. There, T^* is the complete m-ary tree of depth 3 (with $m = 6$), but X_n depends on $(X_{n-1}, X_{n-2}, X_{n-3})$ only via X_{n-3}. For that reason, the limiting model T_∞ in $\mathcal{T}(D)$ with $D = 1$ or $D = 2$ is not T^* truncated at depth D, but rather the empty tree $\{\lambda\}$ consisting of only the root node λ.

Fortunately, we can read a simple necessary and sufficient condition for the "expected" behaviour to occur from the definition of T_∞ itself. Let $\{X_n\}$ be a stationary

and ergodic process on the finite alphabet A. From the description of Csiszár and Talata (2006), it is easy to see that there is a unique minimal context tree model T^* for $\{X_n\}$ of possibly infinite depth. Let $T^*_{|D}$ be T^* truncated at depth D, and write $N_{D-1}(T^*)$ for the set of all internal nodes of T^* at depth $d = D-1$ whose children exist in T^* (at depth D) but are not leaves of T^*. With this notation:

Corollary 4.2. *The limiting model $T_\infty = T^*_{|D}$ if and only if $I(X;J|s) > 0$ for all nodes $s \in N_{D-1}(T^*)$.*

Proof. The result follows directly from the definition of T_∞ combined with the observation that the condition $I(X;J|s) > 0$ is already satisfied for all nodes s of T^* at depth $d = D-1$ whose children are leaves of T^*. □

4.3 The posterior predictive distribution

The branching process representation of the posterior can also be used to facilitate practically useful computations. In Proposition 4.1, an exact expression is given for the posterior predictive distribution $P\left(x_{n+1}|x_{-D+1}^n\right)$ in terms of the branching probabilities $P_{b,s}$.

Proposition 4.1. *The posterior predictive distribution is given by,*

$$P\left(x_{n+1}|x_{-D+1}^n\right) = \sum_{i=0}^{D} \left(\frac{a_{s^{(i)}}\left(x_{n+1}\right) + 1/2}{M_{s^{(i)}} + m/2} \right) \gamma_i, \tag{4.9}$$

where, for $0 \le i \le D$, the string $s^{(i)}$ is the context of length i preceding x_{n+1}, and γ_i is the posterior probability that node $s^{(i)}$ is a leaf, given by,

$$\gamma_i = \begin{cases} P_{b,\lambda}, & i = 0, \\ \prod_{k=0}^{i-1} \left(1 - P_{b,s^{(k)}}\right) P_{b,s^{(i)}}, & 1 \le i \le D-1, \\ \prod_{k=0}^{D-1} \left(1 - P_{b,s^{(k)}}\right), & i = D. \end{cases} \tag{4.10}$$

Proof. Writing $x = x_{-D+1}^n$, the posterior predictive distribution can be expressed as,

$$P\left(x_{n+1}|x\right) = \sum_{T \in \mathcal{T}(D)} P\left(x_{n+1}|T,x\right) \pi(T|x). \tag{4.11}$$

For any tree $T \in \mathcal{T}(D)$, exactly one of the contexts $s^{(i)}$, $0 \le i \le D$, is a leaf of the tree. For every $0 \le i \le D$, define the subset $\mathcal{T}_i(D) \subset \mathcal{T}(D)$ to be the collection of trees $T \in \mathcal{T}(D)$ such that the context of x_{n+1} that is a leaf of T is $s^{(i)}$; these $\mathcal{T}_i(D)$ are disjoint and their union is $\mathcal{T}(D)$. The key observation here is that $P(x_{n+1}|T,x)$ is the same for all trees $T \in \mathcal{T}_i(D)$, since,

$$P(x_{n+1}|T,x) = \int_\theta P(x_{n+1}|T,\theta,x)\pi(\theta|T,x)\, d\theta$$

$$= \int_{\theta_{s^{(i)}}} P(x_{n+1}|T, \theta_{s^{(i)}}, x)\pi(\theta_{s^{(i)}}|T, x)\ d\theta_{s^{(i)}}$$

$$= \int_{\theta_{s^{(i)}}} \theta_{s^{(i)}}(x_{n+1})\,\pi(\theta_{s^{(i)}}|T, x)\ d\theta_{s^{(i)}} = \frac{a_{s^{(i)}}(x_{n+1}) + 1/2}{M_{s^{(i)}} + m/2}, \qquad (4.12)$$

where we used the full conditional density of the parameters in (3.3). So, from (4.11),

$$P(x_{n+1}|x) = \sum_{i=0}^{D} \sum_{T\in\mathcal{T}_i(D)} P(x_{n+1}|T, x)\,\pi(T|x) = \sum_{i=0}^{D} \frac{a_{s^{(i)}}(x_{n+1}) + 1/2}{M_{s^{(i)}} + m/2} \sum_{T\in\mathcal{T}_i(D)} \pi(T|x),$$

which completes the proof upon noticing that the last sum $\sum_{T\in\mathcal{T}_i(D)} \pi(T|x)$ is exactly the posterior probability that node $s^{(i)}$ is a leaf, namely, γ_i as in (4.10). □

5 Experimental results

Being able to obtain exact i.i.d. samples from the posterior is generally more desirable and typically leads to more efficient estimation than using approximate MCMC samples. In Section 5.1 we offer empirical evidence justifying this statement in the present setting through a simple simulation example. Then in Section 5.2 we present the results of a careful empirical study of the natural entropy estimator induced by the BCT framework, compared against a number of the most common alternative estimators, on three simulated and three real-world data sets.

5.1 Comparison with MCMC

Consider $n = 1000$ observations generated from a 5th order, ternary chain, with model given by the context tree of Figure 1 in Section 2 (the values of the parameters $\theta = \{\theta_s; s \in T\}$ are given in Section C of the Supplementary Material). A simple and effective convergence diagnostic here (which can also be viewed as an example of an estimation problem) is the examination of the frequency with which the MAP model, T_1^*, appears in the i.i.d. or the MCMC sample trajectory. The model T_1^* can be identified by the BCT algorithm and its posterior probability $\pi(T_1^*|x)$ can be computed, as in Kontoyiannis et al. (2022).

As shown in Figure 2, the estimates based on the random-walk MCMC sampler of Kontoyiannis et al. (2022) and on the i.i.d. sampler of Section 3.3 both appear to converge quite quickly, with the corresponding MCMC estimates converging significantly more slowly. In 50 independent repetitions of the same experiment (with $N = 1000$ simulated samples in each run), the estimated variance of the MCMC estimates (0.0084) was found to be larger than that for the i.i.d. estimates (1.4×10^{-4}), by a factor of around 60.

Figure 3 shows the trace plots (Roy, 2020) obtained from $N = 10000$ simulated samples from the MCMC and i.i.d. samplers, which can be used to monitor the log-posterior in each case. It is immediately evident that the i.i.d. sampler is more efficient in exploring the effective support of the posterior.

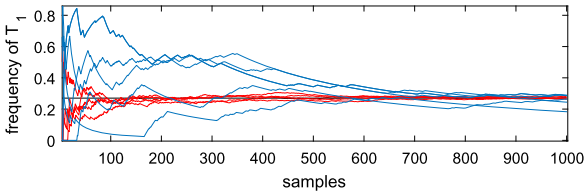


Figure 2: Frequency of T_1^*. Blue: MCMC estimates. Red: i.i.d. estimates. In each case, the five graphs correspond to five independent repetitions of the experiment with $N = 1000$ simulated samples. The horizontal line is the limiting frequency, $\pi(T_1^*|x)$.

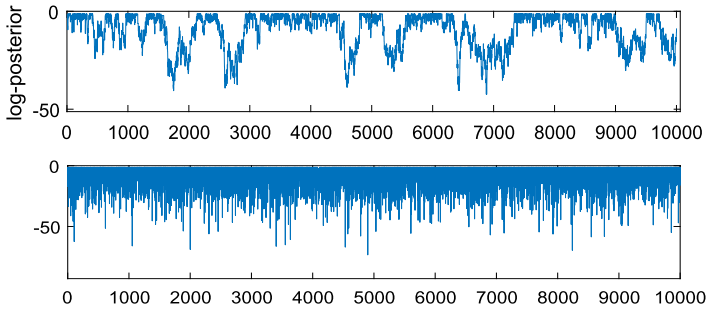


Figure 3: Trace plots showing the log-posterior, $\log \pi(T^{(i)}|x)$, at each iteration. Top: MCMC samples. Bottom: i.i.d. samples.

As expected, the i.i.d. sampler has superior performance compared to the MCMC sampler, both in terms of estimation and in terms of mixing. Also, although the two types of samplers have comparable complexity in terms of computation time and memory requirements, the structure of the i.i.d. sampler is much simpler, giving a much easier implementation. In view of these observations, in the following section we only employ the i.i.d. sampler for the purposes of entropy estimation.

5.2 Entropy estimation

Estimating the entropy rate from empirical data – in this case, a discrete time series – is an important and timely problem that has received a lot of attention in the recent literature, in connection with questions in many areas including neuroscience (Timme and Lapish, 2018), natural language modelling (Willems et al., 2016), animal communication (Kershenbaum, 2014), and cryptography (Simion, 2020), among others; see, e.g., the recent literature reviews by Verdú (2019) and Feutrill and Roughan (2021). The well-known difficulties of entropy estimation stemming from the nonlinear nature of the entropy rate functional and its dependence on the entire process distribution are discussed in the references listed above.

For a general process $\{X_n\}$ on a finite alphabet, the *entropy rate* $\bar{H}$ is defined as $\bar{H} = \lim_{n\to\infty}(1/n)H(X_1^n)$, whenever the limit exists, where $H(X_1^n)$ denotes the usual Shannon entropy (in nats rather than bits, as we take logarithms to the base e) of the discrete random vector X_1^n. For an ergodic, first-order Markov chain $\{X_n\}$, $\bar{H}$ can be expressed as,

$$\bar{H} = -\sum_{i,j\in S} \pi(i)P_{ij}\log P_{ij}, \tag{5.1}$$

where S is the state space of $\{X_n\}$, and (P_{ij}) and $(\pi(i))$ denote its transition matrix and its stationary distribution, respectively. An analogous formula can be written for the entropy rate of any ergodic variable-memory chain with model $T \in \mathcal{T}(D)$, by viewing it as a full Dth order chain and considering blocks of length $(D+1)$, as usual; cf. Cover and Thomas (2012, Ch. 4). This means that $\bar{H}$ can be expressed as an explicit function $\bar{H} = H(T,\theta)$ of the model and the parameters.

Therefore, given a time series x, using the MC sampler of Section 3.3 to produce i.i.d. samples $(T^{(i)},\theta^{(i)})$ from $\pi(T,\theta|x)$, we can obtain i.i.d. samples $H^{(i)} = H(T^{(i)},\theta^{(i)})$ from the posterior $\pi(\bar{H}|x)$ of the entropy rate. The calculation of each $H^{(i)} = H(T^{(i)},\theta^{(i)})$ is straightforward and only requires the computation of the stationary distribution π of the induced first-order chain that corresponds to taking blocks of size $[\mathrm{depth}(T^{(i)})+1]$. The only potential difficulty is if either the depth of $T^{(i)}$ or the alphabet size m are so large that the computation of π becomes computationally expensive. In such cases, $H^{(i)}$ can be computed approximately by including an additional Monte Carlo step: Generate a sufficiently long random sample Y_{-D+1}^M from the chain $(T^{(i)},\theta^{(i)})$, and calculate:

$$H^{(i)} \approx -\frac{1}{M}\log P(Y_1^M|Y_{-D+1}^0, T^{(i)},\theta^{(i)}). \tag{5.2}$$

The ergodic theorem and the central limit theorem for Markov chains (Chung, 1967; Meyn and Tweedie, 2012) then guarantee the accuracy of (5.2).

In the remainder of this section, the BCT estimator (with maximum model depth $D = 10$) is compared with the state-of-the-art approaches, as identified by Gao et al. (2008) and Verdú (2019) and summarised below. The BCT estimator is found to generally give the most reliable estimates on a variety of different types of simulated and real-world data. Moreover, compared to most existing approaches that give simple point estimates (sometimes accompanied by confidence intervals), the BCT estimator has the additional advantage that it provides the entire posterior distribution $\pi(\bar{H}|x)$.

Plug-in estimator Motivated by the definition of the entropy rate, the simplest and one of the most commonly used estimators of the entropy rate is the per-sample entropy of the empirical distribution of k-blocks. Letting $\widehat{p}_k(y_1^k)$, $y_1^k \in A^k$, denote the empirical distribution of k-blocks induced by the data on A^k, the *plug-in* or *maximum-likelihood* estimator is simply, $\widehat{H}_k = (1/k)H(\widehat{p}_k)$. The main advantage of this estimator is its simplicity. Well-known drawbacks include its high variance due to undersampling, and the difficulty in choosing appropriate block-lengths k effectively.

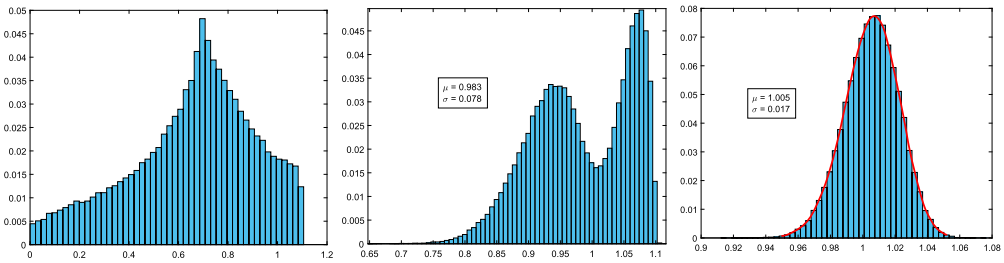
Lempel-Ziv estimator Among the numerous match-length-based entropy estimators that have been derived from the Lempel-Ziv family of data compression algorithms, we consider the increasing-window estimator of Gao et al. (2008), identified there as the most effective one. For every position i in the observed data, let ℓ_i denote the length of the longest segment $x_i^{i+\ell_i-1}$ starting at i which also appears somewhere in the window x_0^{i-1} preceding i. Writing $L_i = 1 + \ell_i$ for each i, the relevant estimator is,

$$\widehat{H}_{\mathrm{LZ}} = \frac{1}{n} \sum_{i=2}^{n} \frac{\log i}{L_i}.$$

CTW estimator This uses the prior predictive likelihood $P(x)$ computed by the CTW algorithm, to define $\widehat{H}_{\mathrm{CTW}} = -(1/n) \log P(x_1^n)$. This estimator was found by Gao et al. (2008) and Verdú (2019) to achieve the best performance in practice. Its consistency and asymptotic normality follow easily from standard results, and its (always positive) bias is of $O((\log n)/n)$, which can be shown to be in a minimax sense as small as possible. In all experiments we take the maximum depth of CTW to be $D = 10$.

PPM estimator Using a different adaptive probability assignment, $Q(x)$, this method forms an estimate of the same type as the CTW estimator, $\widehat{H}_{\mathrm{PPM}} = -(1/n) \log Q(x_1^n)$, where prediction by partial matching (PPM) (Cleary and Witten, 1984) is used to fit the model that leads to $Q(x_1^n)$. We use the interpolated smoothing variant of PPM introduced by Bunton (1996), which is implemented in the R package available at: https://rdrr.io/github/pmcharrison/ppm/.

A ternary chain We consider the same $n = 1000$ observations generated from the 5th order, ternary chain examined in Section 5.1. The entropy rate of this chain is $\bar{H} = 1.02$. In Figure 4 we show MC estimates of the prior distribution $\pi(\bar{H})$, and of the posterior $\pi(\bar{H}|x)$ based on $n = 100$ and on $n = 1000$ observations from the chain. After $n = 1000$ observations, the posterior is close to a Gaussian with mean $\mu = 1.005$ and standard deviation $\sigma = 0.017$. For each histogram $N = 10^5$ i.i.d. samples were used,



(a) prior (b) $n = 100$ (c) $n = 1000$

Figure 4: Prior $\pi(\bar{H})$ and posterior $\pi(\bar{H}|x)$ of the entropy rate $\bar{H}$ with $n = 100$ and $n = 1000$ observations x.

and in each case (and in all subsequent examples), the vertical axis of the histograms shows the frequency of the bins in the Monte Carlo sample.

Figure 5 shows the performance of the BCT estimator compared with the other four estimators described above, as a function of the length n of the available observations x. For BCT we plot the posterior mean. For the plug-in we plot estimates with block-lengths $k = 5, 6, 7$. It is easily observed that the BCT estimator outperforms all the alternatives, and converges faster and closer to the true value of $\bar{H}$.

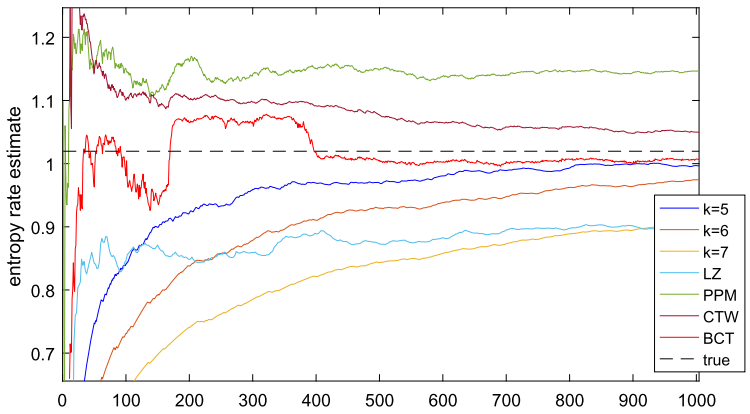


Figure 5: Entropy rate estimates for the 5th order ternary chain, as the number of observations increases.

A third order binary chain Here, we consider $n = 1000$ observations generated from an example of a third order binary chain from Berchtold and Raftery (2002). The underlying model is the complete binary tree of depth 3 pruned at node $s = 11$; the tree model T and the parameter values $\theta = \{\theta_s; s \in T\}$ are given in Section C of the Supplementary Material. The entropy rate of this chain is $\bar{H} = 0.4815$. Figure 6 shows the performance of all five estimators, where the BCT estimator (which uses the posterior mean again) is found to have the best performance. The histogram of the BCT posterior after $n = 1000$ observations, shown in Section C of the Supplementary Material, is close to a Gaussian with mean $\mu = 0.4806$ and standard deviation $\sigma = 0.0405$.

A bimodal posterior We re-examine a simulated time series x from Kontoyiannis et al. (2022), which consists of $n = 1450$ observations generated from a 3rd order chain $\{X_n\}$ with alphabet size $m = 6$ and with the property that each X_n depends on past observations only via X_{n-3}. The complete specification of the chain is given in Section C of the Supplementary Material. Its entropy rate is $\bar{H} = 1.355$. An interesting aspect of this data set is that the model posterior is bimodal, with one mode corresponding to the empty tree (describing i.i.d. observations) and the other consisting of tree models of depth 3.

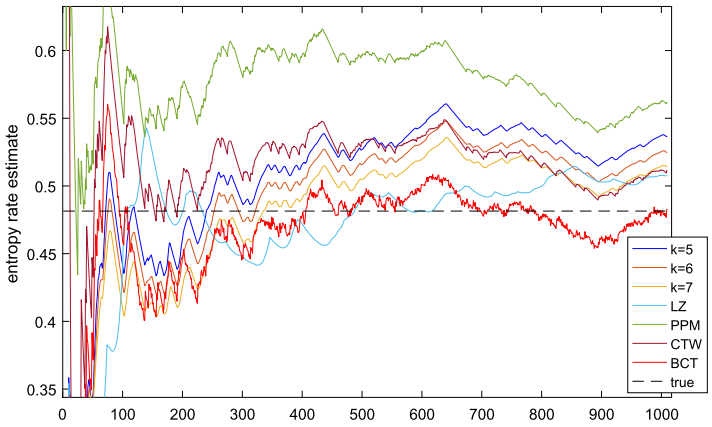
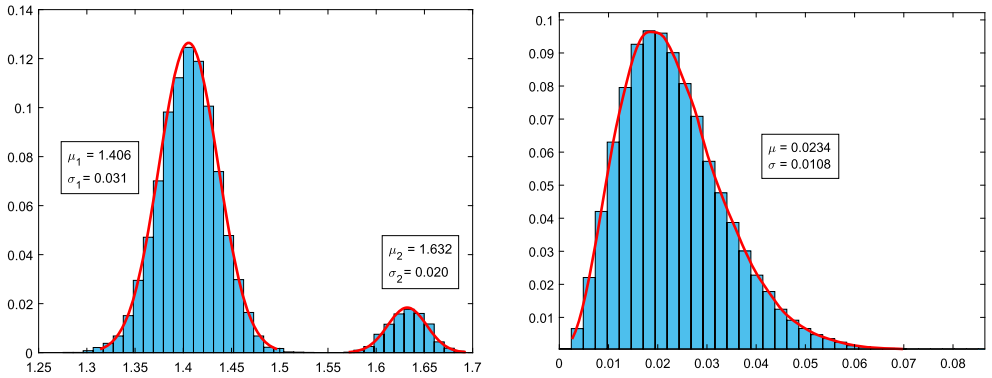


Figure 6: Entropy rate estimates for the third order binary chain, as the number of observations increases.

As shown in Figure 7a, the posterior of the entropy rate is also bimodal here, with two separated approximately-Gaussian modes corresponding to each of the modes of the model posterior. The dominant mode is the one corresponding to models of depth 3; it has mean $\mu_1 = 1.406$, standard deviation $\sigma_1 = 0.031$, and relative weight $w_1 = 0.91$. The second mode corresponding to the empty tree has mean $\mu_2 = 1.632$, standard deviation $\sigma_2 = 0.020$, and a much smaller weight $w_2 = 1 - w_1 = 0.09$. In this case, the mode of $\pi(\bar{H}|x)$ gives a more reasonable choice for a point estimate than the posterior mean. Like in the previous two examples, the BCT estimator performs better than most benchmarks, as illustrated in Section C of the Supplementary Material.



(a) bimodal example (b) spike train

Figure 7: Histograms of the posterior distribution $\pi(\bar{H}|x)$ of the entropy rate, constructed from $N = 10^5$ i.i.d. samples in each case.

Neural spike trains We consider $n = 1000$ binary observations from a spike train recorded from a single neuron in region V4 of a monkey's brain. The BCT posterior is shown in Figure 7b: Its mean is $\mu = 0.0234$, its standard deviation is $\sigma = 0.0108$, and is skewed to the right. This dataset is the first part of a long spike train of length $n = 3{,}919{,}361$ from Gregoriou et al. (2009, 2012). Although there is no "true" value of the entropy rate here, for the purposes of comparison we use the estimate obtained by the CTW estimator (identified as the most effective method by Gao et al. (2008) and Verdú (2019)) when all $n = 3{,}919{,}361$ samples are used, giving $\bar{H} = 0.0241$. The resulting estimates for all five methods (with the posterior mean given for BCT) are summarised in Table 1, verifying again that BCT outperforms all the other methods.

	"True"	BCT	CTW	PPM	LZ	$k = 2$	$k = 5$	$k = 10$	$k = 15$
$\widehat{H}$	0.0241	**0.0234**	0.0249	0.0360	0.0559	0.0204	0.0204	0.0198	0.0187

Table 1: Entropy rate estimates for the neural spike train.

Financial data Here, we consider $n = 2000$ observations from the financial dataset F.2 of Kontoyiannis et al. (2022). This consists of tick-by-tick price changes of the Facebook stock price, quantised to three values: $x_i = 0$ if the price goes down, $x_i = 1$ if it stays the same, and $x_i = 2$ if it goes up. The BCT entropy-rate posterior is shown in Figure 8a: It has mean $\mu = 0.921$, and standard deviation $\sigma = 0.028$.

Once again, as the "true" value of the entropy rate we take the estimate produced by the CTW estimator on a longer sequence with $n = 10^4$ observations, giving $\bar{H} = 0.916$. The results of all five estimators are summarised in Table 2, where for the BCT estimator we once again give the posterior mean.

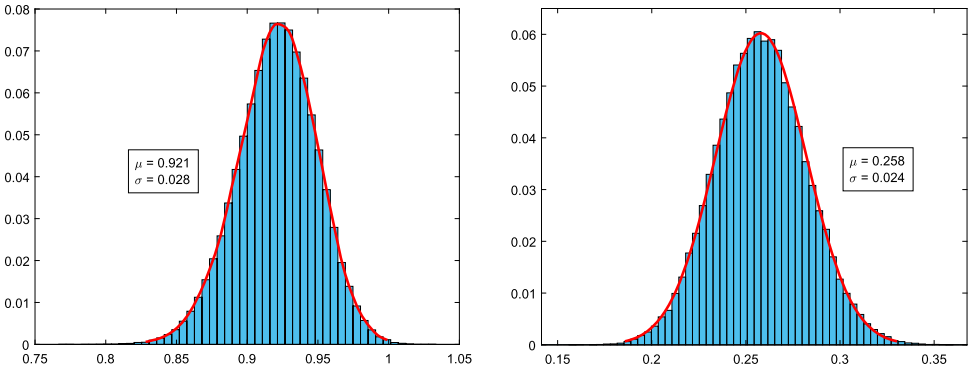
	"True"	BCT	CTW	PPM	LZ	$k = 5$	$k = 6$	$k = 7$	$k = 10$
$\widehat{H}$	0.916	**0.921**	0.939	1.049	0.846	0.930	0.907	0.870	0.713

Table 2: Entropy rate estimates for the financial data set.

Pewee birdsong The last data set examined is a time series x describing the twilight song of the wood pewee bird (Craig, 1943; Sarkar and Dunson, 2016). It consists of $n = 1327$ observations from an alphabet of size $m = 3$. The BCT posterior is shown in Figure 8b: It is approximately Gaussian with mean $\mu = 0.258$ and standard deviation $\sigma = 0.024$. The fact that the standard deviation is small is important as it suggests "confidence" in the resulting estimates, which is important because here (as in most real applications) there is no knowledge of a "true" underlying value. Table 3 shows all the resulting estimates; the posterior mean is shown for the BCT estimator.

	BCT	CTW	PPM	LZ	$k = 2$	$k = 5$	$k = 10$	$k = 15$
$\widehat{H}$	0.258	0.278	0.318	0.275	0.776	0.467	0.336	0.272

Table 3: Entropy rate estimates for the pewee song data.



(a) financial dataset (b) pewee birdsong

Figure 8: Histograms of the posterior distribution $\pi(\bar{H}|x)$ of the entropy rate, constructed from $N = 10^5$ i.i.d. samples in each case.

Summary The main conclusion from the results on the six data sets examined in this section is that the BCT estimator gives the most accurate and reliable results among the five estimators considered. In addition to the fact that the BCT point estimates typically outperform those produced by other methods, the BCT estimator is accompanied by the entire posterior distribution $\pi(\bar{H}|x)$ of the entropy rate, induced by the observations x. As usual, this distribution can be used to quantify the uncertainty in estimating $\bar{H}$, and it contains significantly more information than simple point estimates and their associated confidence intervals.

6 Concluding remarks

In this work, we revisited the Bayesian Context Trees (BCT) modelling framework, which was recently found to be very effective for a range of statistical tasks in the analysis of discrete time series. We showed that the prior and posterior distributions on model space admit simple and easily interpretable representations in terms of branching processes, and we demonstrated their utility both in theory and in practice. The branching process representation was first employed to develop an efficient Monte Carlo sampler that provides i.i.d. samples from the joint posterior on models and parameters, thus facilitating effective Bayesian inference with empirical time series data. Then, it was used to establish strong theoretical results on the asymptotic consistency of the BCT posterior on model space, which provide important theoretical justifications for the use of the BCT framework in practice. Finally, the performance of the proposed Monte Carlo sampler was examined extensively in the context of entropy estimation. The resulting fully-Bayesian entropy estimator was found to outperform several of the state-of-the-art approaches, on both simulated and real-world data.

Although the BCT framework was originally developed for modelling and inference of discrete-valued time series, it was recently used to develop general mixture models

for real-valued time series, along with a collection of associated algorithmic tools for inference (Papageorgiou and Kontoyiannis, 2022). Extending the results presented in this work to that setting presents an interesting direction of further research, motivated by important and timely practical applications.

Supplementary Material

The Supplementary Material include complete proofs of all theoretical results, and also additional experimental results, details and discussion (DOI: 10.1214/23-BA1362SUPP; .pdf).

References

Athreya, K. and Ney, P. (2004). *Branching processes*. Courier Corporation. MR2047480. 502, 506, 507

Bacallado, S. (2011). "Bayesian analysis of variable-order, reversible Markov chains." *The Annals of Statistics*, 39(2): 838–864. MR2816340. doi: https://doi.org/10.1214/10-AOS857. 503

Bacallado, S., Favaro, S., and Trippa, L. (2013). "Bayesian nonparametric analysis of reversible Markov chains." *The Annals of Statistics*, 870–896. MR3099124. doi: https://doi.org/10.1214/13-AOS1102. 503

Bacallado, S., Pande, V., Favaro, S., and Trippa, L. (2016). "Bayesian regularization of the length of memory in reversible sequences." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(4): 933–946. MR3534357. doi: https://doi.org/10.1111/rssb.12140. 503

Berchtold, A. and Raftery, A. (2002). "The mixture transition distribution model for high-order Markov chains and non-Gaussian time series." *Statistical Science*, 17(3): 328–356. MR1962488. doi: https://doi.org/10.1214/ss/1042727943. 501, 520

Bernardo, J. and Smith, A. (2009). *Bayesian theory*, volume 405. John Wiley & Sons. MR1274699. doi: https://doi.org/10.1002/9780470316870. 502

Blackwell, D. (1947). "Conditional expectation and unbiased sequential estimation." *The Annals of Mathematical Statistics*, 105–110. MR0019903. doi: https://doi.org/10.1214/aoms/1177730497. 509

Breiman, L. (1992). *Probability*, volume 7 of *Classics in Applied Mathematics*. Philadelphia, PA: SIAM. MR1163370. doi: https://doi.org/10.1137/1.9781611971286. 514

Bühlmann, P. (2000). "Model selection for variable length Markov chains and tuning the context algorithm." *Annals of the Institute of Statistical Mathematics*, 52(2): 287–315. MR1763564. doi: https://doi.org/10.1023/A:1004165822461. 501

Bühlmann, P. and Wyner, A. (1999). "Variable length Markov chains." *The An-*

nals of Statistics, 27(2): 480–513. MR1714720. doi: https://doi.org/10.1214/aos/1018031204. 501

Bunton, S. (1996). *On-line stochastic processes in data compression*. Ph.D. thesis, University of Washington. MR2695334. 519

Cai, H., Kulkarni, S., and Verdú, S. (2004). "Universal entropy estimation via block sorting." *IEEE Transactions on Information Theory*, 50(7): 1551–1561. MR2095858. doi: https://doi.org/10.1109/TIT.2004.830771. 503

Catoni, O. (2004). *Statistical learning theory and stochastic optimization*, volume 1851 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001. MR2163920. doi: https://doi.org/10.1007/b99352. 511

Chipman, H., George, E., McCulloch, R., Clyde, M., Foster, D., and Stine, R. (2001). "The practical implementation of Bayesian model selection." In *Model selection*, volume 38 of *IMS Lecture Notes Monogr. Ser.*, 65–134. Inst. Math. Statist., Beachwood, OH. With discussion by M. Clyde, Dean P. Foster, and Robert A. Stine, and a rejoinder by the authors. MR2000752. doi: https://doi.org/10.1214/lnms/1215540964. 502

Chung, K. (1967). "Markov chains with stationary transition probabilities." *Springer-Verlag, New York*. MR0217872. 514, 518

Cleary, J. and Witten, I. (1984). "Data compression using adaptive coding and partial string matching." *IEEE Transactions on Communications*, 32(4): 396–402. 503, 519

Cover, T. and Thomas, J. (2012). *Elements of information theory*. New York: J. Wiley & Sons, second edition. MR2239987. 511, 518

Cowles, M. and Carlin, B. (1996). "Markov chain Monte Carlo convergence diagnostics: A comparative review." *Journal of the American Statistical Association*, 91(434): 883–904. MR1395755. doi: https://doi.org/10.2307/2291683. 509

Craig, W. (1943). *The song of the wood pewee (Myiochanes virens Linnaeus): A study of bird music*. New York State Museum Bulletin No. 334. Albany, NY: University of the State of New York. 522

Csiszár, I. and Talata, Z. (2006). "Context tree estimation for not necessarily finite memory processes, via BIC and MDL." *IEEE Transactions on Information Theory*, 52(3): 1007–1016. MR2238067. doi: https://doi.org/10.1109/TIT.2005.864431. 515

Feutrill, A. and Roughan, M. (2021). "A review of Shannon and differential entropy rate estimation." *Entropy*, 23(8). MR4303838. doi: https://doi.org/10.3390/e23081046. 517

Fokianos, K. and Kedem, B. (2003). "Regression theory for categorical time series." *Statistical Science*, 18(3): 357–376. MR2061915. doi: https://doi.org/10.1214/ss/1076102425. 501

Gao, Y., Kontoyiannis, I., and Bienenstock, E. (2008). "Estimating the entropy of binary time series: Methodology, some theory and a simulation study." *Entropy*, 10(2): 71–99. MR2425433. doi: https://doi.org/10.3390/entropy-e10020071. 503, 518, 519, 522

Gelfand, A. and Smith, A. (1990). "Sampling-based approaches to calculating marginal densities." *Journal of the American Statistical Association*, 85(410): 398–409. MR1141740. 509

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (1995). *Bayesian data analysis*. Chapman and Hall/CRC. MR1385925. 502

Gelman, A. and Rubin, D. (1992). "Inference from iterative simulation using multiple sequences." *Statistical Science*, 7(4): 457–472. 509

Gregoriou, G., Gotts, S., and Desimone, R. (2012). "Cell-type-specific synchronization of neural activity in FEF with V4 during attention." *Neuron*, 73(3): 581–594. 522

Gregoriou, G., Gotts, S., Zhou, H., and Desimone, R. (2009). "High-frequency, long-range coupling between prefrontal and visual cortex during attention." *Science*, 324(5931): 1207–1210. 522

Harris, T. (1963). *The theory of branching processes*, volume 6. Springer Berlin. MR0163361. 502, 506, 507

Heiner, M. and Kottas, A. (2022). "Estimation and selection for high-order Markov chains with Bayesian mixture transition distribution models." *Journal of Computational and Graphical Statistics*, 31(1): 100–112. MR4387214. doi: https://doi.org/10.1080/10618600.2021.1979565. 503

Heiner, M., Kottas, A., and Munch, S. (2019). "Structured priors for sparse probability vectors with application to model selection in Markov chains." *Statistics and Computing*, 29(5): 1077–1093. MR3994618. doi: https://doi.org/10.1007/s11222-019-09856-2. 503

Ibragimov, I. (1962). "Some limit theorems for stationary processes." *Theory of Probability and its Applications*, 7: 349–382. MR0148125. 513

Jääskinen, V., Xiong, J., Corander, J., and Koski, T. (2014). "Sparse Markov chains for sequence data." *Scandinavian Journal of Statistics*, 41(3): 639–655. MR3249421. doi: https://doi.org/10.1111/sjos.12053. 501

Jiao, J., Permuter, H., Zhao, L., Kim, Y., and Weissman, T. (2013). "Universal estimation of directed information." *IEEE Transactions on Information Theory*, 59(10): 6220–6242. MR3106825. doi: https://doi.org/10.1109/TIT.2013.2267934. 511

Kershenbaum, A. (2014). "Entropy rate as a measure of animal vocal complexity." *Bioacoustics*, 23(3): 195–208. 517

Kontoyiannis, I., Mertzanis, L., Panotopoulou, A., Papageorgiou, I., and Skoularidou, M. (2022). "Bayesian Context Trees: Modelling and exact inference for discrete time series." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*,

84(4): 1287–1323. MR4494161. doi: https://doi.org/10.1111/rssb.12511. 502, 503, 504, 505, 506, 507, 509, 514, 516, 520, 522

Krichevsky, R. and Trofimov, V. (1981). "The performance of universal encoding." *IEEE Transactions on Information Theory*, 27(2): 199–207. MR0633417. doi: https://doi.org/10.1109/TIT.1981.1056331. 511

London, M., Schreibman, A., Häusser, M., Larkum, M., and Segev, I. (2002). "The information efficacy of a synapse." *Nature Neuroscience*, 5(4): 332–340. 503

Lungu, V., Papageorgiou, I., and Kontoyiannis, I. (2022a). "Bayesian Change-Point Detection via Context-Tree Weighting." In *2022 IEEE Information Theory Workshop (ITW)*, 125–130. IEEE. doi: https://doi.org/10.1109/ITW54588.2022.9965823. 502

Lungu, V., Papageorgiou, I., and Kontoyiannis, I. (2022b). "Change-point Detection and Segmentation of Discrete Data using Bayesian Context Trees." *arXiv preprint arXiv:2203.04341*. 502

Mächler, M. and Bühlmann, P. (2004). "Variable length Markov chains: methodology, computing, and software." *Journal of Computational and Graphical Statistics*, 13(2): 435–455. MR2063994. doi: https://doi.org/10.1198/1061860043524. 501

Meyn, S. and Tweedie, R. (2012). *Markov chains and stochastic stability*. Springer Science & Business Media. MR1287609. doi: https://doi.org/10.1007/978-1-4471-3267-7. 518

Nemenman, I., Bialek, W., and Van Steveninck, R. (2004). "Entropy and information in neural spike trains: Progress on the sampling problem." *Physical Review E*, 69(5): 056111. 503

Paninski, L. (2003). "Estimation of entropy and mutual information." *Neural Computation*, 15(6): 1191–1253. 503

Papageorgiou, I. and Kontoyiannis, I. (2022). "The Bayesian Context Trees State Space Model: Interpretable mixture models for time series." *arXiv preprint arXiv:2106.03023*. 524

Papageorgiou, I. and Kontoyiannis, I. (2023). "Supplementary Material for "Posterior representations for Bayesian Context Trees: Sampling, estimation and convergence"." *Bayesian Analysis*. doi: https://doi.org/10.1214/23-BA1362SUPP. 508

Papageorgiou, I., Kontoyiannis, I., Mertzanis, L., Panotopoulou, A., and Skoularidou, M. (2021). "Revisiting context-tree weighting for Bayesian inference." In *2021 IEEE International Symposium on Information Theory (ISIT)*, 2906–2911. doi: https://doi.org/10.1109/ISIT45174.2021.9518189. 502

Papageorgiou, I., Lungu, V., and Kontoyiannis, I. (2020). *BCT: Bayesian Context Trees for Discrete Time Series*. R package version 1.1. https://CRAN.R-project.org/package=BCT. 502

Philipp, W. and Stout, W. (1975). *Almost sure invariance principles for partial sums of*

weakly dependent random variables, volume 161. Memoirs of the AMS. MR0433597. doi: https://doi.org/10.1090/memo/0161. 513, 514

Raftery, A. (1985). "A model for high-order Markov chains." *Journal of the Royal Statistical Society: Series B (Methodological)*, 47(3): 528–539. MR0844484. 501

Rio, E. (1995). "The functional law of the iterated logarithm for stationary strongly mixing sequences." *The Annals of Probability*, 1188–1203. MR1349167. 514

Rissanen, J. (1983a). "A universal data compression system." *IEEE Transactions on Information Theory*, 29(5): 656–664. MR0730903. doi: https://doi.org/10.1109/TIT.1983.1056741. 501

Rissanen, J. (1983b). "A universal prior for integers and estimation by minimum description length." *Annals of Statistics*, 11(2): 416–431. 501

Rissanen, J. (1986). "Complexity of strings in the class of Markov sources." *IEEE Transactions on Information Theory*, 32(4): 526–532. MR0852122. doi: https://doi.org/10.1109/TIT.1986.1057210. 501

Robert, C. and Casella, G. (2004). *Monte Carlo statistical methods*, volume 2. Springer. MR2080278. doi: https://doi.org/10.1007/978-1-4757-4145-2. 509

Roy, V. (2020). "Convergence diagnostics for Markov chain Monte Carlo." *Annual Review of Statistics and Its Application*, 7: 387–412. MR4104198. doi: https://doi.org/10.1146/annurev-statistics-031219-041300. 516

Sarkar, A. and Dunson, D. (2016). "Bayesian nonparametric modeling of higher order Markov chains." *Journal of the American Statistical Association*, 111(516): 1791–1803. MR3601736. doi: https://doi.org/10.1080/01621459.2015.1115763. 501, 522

Shannon, C. (1951). "Prediction and entropy of printed English." *Bell System Technical Journal*, 30(1): 50–64. 503

Simion, E. (2020). "Entropy and randomness: From analogic to quantum world." *IEEE Access*, 8: 74553–74561. 517

Strong, S., Koberle, R., Van Steveninck, R., and Bialek, W. (1998). "Entropy and information in neural spike trains." *Physical Review Letters*, 80(1): 197. MR1983010. 503

Timme, N. and Lapish, C. (2018). "A tutorial for information theory in neuroscience." *eneuro*, 5(3): 1–40. 517

Verdú, S. (2019). "Empirical estimation of information measures: A literature guide." *Entropy*, 21(8): 720. MR4003023. doi: https://doi.org/10.3390/e21080720. 503, 517, 518, 519, 522

Weinberger, M., Merhav, N., and Feder, M. (1994). "Optimal sequential probability assignment for individual sequences." *IEEE Transactions on Information Theory*, 40(2): 384–396. MR1294049. doi: https://doi.org/10.1109/18.312161. 502

Willems, F. (1998). "The context-tree weighting method: extensions." *IEEE Transac-*

tions on Information Theory, 44(2): 792–798. MR1607711. doi: https://doi.org/10.1109/18.661523. 502

Willems, F., Shtarkov, Y., and Tjalkens, T. (1993). "Context tree weighting: Basic properties." Unpublished manuscript. Available online at: www.sps.ele.tue.nl/members/F.M.J.Willems/. MR1807734. 510

Willems, F., Shtarkov, Y., and Tjalkens, T. (1995). "The context-tree weighting method: basic properties." *IEEE Transactions on Information Theory*, 41(3): 653–664. 502

Willems, R., Frank, S., Nijhof, A., Hagoort, P., and Van den Bosch, A. (2016). "Prediction during natural language comprehension." *Cerebral Cortex*, 26(6): 2506–2516. 517

Wyner, A. and Ziv, J. (1989). "Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression." *IEEE Transactions on Information Theory*, 35(6): 1250–1258. MR1036629. doi: https://doi.org/10.1109/18.45281. 503

Xie, Q. and Barron, A. (2000). "Asymptotic minimax regret for data compression, gambling, and prediction." *IEEE Transactions on Information Theory*, 46(2): 431–445. MR1748979. doi: https://doi.org/10.1109/18.825803. 511

Xiong, J., Jääskinen, V., and Corander, J. (2016). "Recursive learning for sparse Markov models." *Bayesian Analysis*, 11(1): 247–263. MR3447098. doi: https://doi.org/10.1214/15-BA949. 501

Zeger, S. and Liang, K. (1986). "Longitudinal data analysis for discrete and continuous outcomes." *Biometrics*, 121–130. 501

Zhao, O. and Woodroofe, M. (2008). "Law of the iterated logarithm for stationary processes." *The Annals of Probability*, 36(1): 127–142. MR2370600. doi: https://doi.org/10.1214/009117907000000079. 514

Ziv, J. and Lempel, A. (1977). "A universal algorithm for sequential data compression." *IEEE Transactions on Information Theory*, 23(3): 337–343. MR0530215. doi: https://doi.org/10.1109/tit.1977.1055714. 503

Acknowledgments

We are grateful to Georgia Gregoriou for providing us with the spike train data examined in Section 5.2.