Submitted to the Annals of Statistics

CONVERGENCE RATE OF MARKOV CHAIN METHODS FOR GENOMIC MOTIF DISCOVERY*

By Dawn B. Woodard
and
Jeffrey S. Rosenthal

Cornell University and University of Toronto

We analyze the convergence rate of a simplified version of a popular Gibbs sampling method used for statistical discovery of gene regulatory binding motifs in DNA sequences. This sampler satisfies a very strong form of ergodicity (uniform). However, we show that, due to multimodality of the posterior distribution, the rate of convergence often decreases exponentially as a function of the length of the DNA sequence. Specifically, we show that this occurs whenever there is more than one true repeating pattern in the data. In practice there are typically multiple such patterns in biological data, the goal being to detect the most well-conserved and frequently-occurring of these. Our findings match empirical results, in which the motif-discovery Gibbs sampler has exhibited such poor convergence that it is used only for finding modes of the posterior distribution (candidate motifs) rather than for obtaining samples from that distribution. Ours are some of the first meaningful bounds on the convergence rate of a Markov chain method for sampling from a multimodal posterior distribution, as a function of statistical quantities like the number of observations.

1. Introduction. Gene regulatory binding motifs are short DNA sequences that control gene expression. The identification of these regulatory motifs poses several challenges: they are only 6-15 base pairs in length, and do not contain clear start and stop codons; a regulatory motif is indistinguishable from random sequences of the same length except that it is a particular sequence that occurs more frequently than expected under the background model. Discovery of previously undescribed regulatory motifs in DNA sequences thus involves both finding such a repeating pattern ("motif") and determining where that pattern occurs in the sequences (Kellis et

*The authors would like to thank Krzysztof Latuszynski for assistance with one of the proofs, and the referee and Associate Editor for their excellent feedback. Work partially funded by National Science Foundation award #CMMI-0926814 and by NSERC of Canada.

AMS 2000 subject classifications: Primary 62F15; secondary 60J10

Keywords and phrases: Gibbs sampler, DNA, slow mixing, spectral gap, multimodal

al. 2004); this is illustrated in Figure 1.

One of the most effective methods for identifying new regulatory motifs is based on a statistical model and associated Gibbs sampling computational method (Liu, Neuwald and Lawrence, 1995). This approach has been popularized with the availability of software programs for its use, such as BioProspector (Liu, Brutlag and Liu, 2001) and AlignAce (Roth et al., 1998).

Like most other methods for identifying regulatory motifs, the Gibbs sampling method often yields different answers when starting from different initial configurations. The method is applied by rerunning the Gibbs sampler many times, using randomly generated initial positions. The resulting candidate motifs are sorted according to some goodness-of-fit measure and then the highest-scoring motifs are reported (Lawrence et al., 1993; Liu, Brutlag and Liu, 2001; Jensen et al., 2004). This fact contrasts with the theoretical properties and traditional use of a Gibbs sampler, namely to be simulated until it has some claim of having converged to the posterior distribution, at which point the answer should be the same regardless of initialization.

We address a particular model and Gibbs sampler that are representative of this family of methods. We analyze the convergence rate of a simplified version of the Gibbs sampler and show that, due to multimodality of the posterior distribution, the convergence rate typically decreases exponentially as a function of the DNA sequence length (Theorem 3.2). Specifically this occurs when there is more than one true repeating pattern in the data, meaning that the DNA is made up of short subsequences, each of which is either equal to one of several motifs or is generated from the background model. In practice there are typically multiple distinct repeating patterns in biological data, corresponding to multiple gene regulatory binding motifs or to repeating patterns that have other biological significance, such as "determinants of mRNA stability or even sites for regulation by antisense transcripts" (Roth et al., 1998). The goal is to detect the most frequently-occurring and well-conserved motif or motifs (Neuwald, Liu and Lawrence, 1995). So in practice we can expect the sampler convergence rate to decay exponentially; this is equivalent to the run time of the algorithm growing exponentially in the sequence length, for a fixed accuracy. The multimodality of the posterior and resulting poor convergence are illustrated in Figure 2, which shows posterior density estimates of a particular function of the parameter vector, from two different Gibbs sampling chains. Initialized with distinct parameter values, the two chains have become trapped in different modes of the posterior density and thus have not yet individually converged to the posterior distribution.

The multimodality of the posterior distribution arises due to a contradic-

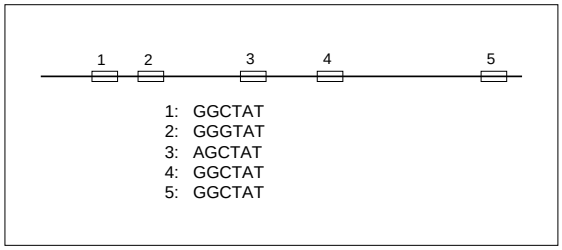

FIG 1. *Illustration of motif discovery: finding an unknown repeating pattern in a long DNA sequence. The pattern can vary slightly between instances.*

tion between the data, which typically have multiple true repeating patterns, and the model assumption of a single such pattern. Practitioners use the model not because it is believed to precisely capture the true process that generated the data (which is extremely complex) but because it captures several important features of that process (Neuwald, Liu and Lawrence, 1995; Roth et al., 1998). Our results show that the presence of multiple motifs, even if some occur very infrequently, causes slow convergence. Recognizing that there can be multiple true motifs, a variant on the Gibbs sampler has been proposed that allows for a fixed number of motifs greater than one (Neuwald, Liu and Lawrence, 1995). This approach is only likely to fix the slow convergence if the number of motifs in the model is at least as large as the number of true motifs in the data. This is only a practical solution if the number of true motifs is small.

Our simplification of the model and associated Gibbs sampler assumes that motifs can only end at locations in the sequence that are divisible by the motif length, instead of at arbitrary locations (Section 2.2). This is done to facilitate analysis, by avoiding the "phase shift" issue that occurs in the original sampler (Lawrence et al., 1993; Liu, 1994). Since phase shift slows convergence of the chain, it is likely (but unproven) that our results on slow convergence of the simplified chain also hold for the original chain.

We also have progress towards a result that the convergence rate decreases polynomially if there is no more than one true (and identifiable) motif in the data. We give empirical support for this conjecture, and prove polynomial decay of the convergence rate for the case of length-one motifs. In this case any true motifs are non-identifiable; see Theorem 3.3.

Ours are some of the few meaningful bounds on the convergence rate of a Bayesian statistics Markov chain method, as a function of statistical quantities such as the number of observations or number of groups. Such results are particularly rare for multimodal posterior densities. Roberts and Sahu

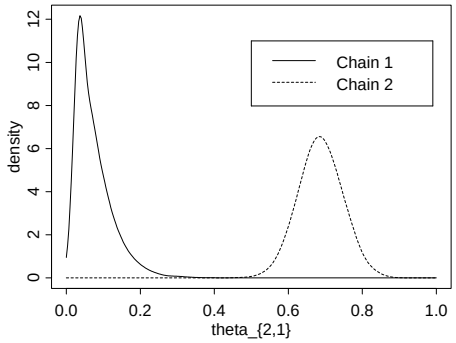

FIG 2. *The posterior density estimates of* $\hat{\theta}_{2,1}(\mathbf{A})$ *from two different Gibbs sampling chains, in the case of two true motifs.*

(2001) show that the convergence rate of a Gibbs sampler for a unimodal posterior density in $\mathbb{R}^d$ approaches a constant as the number of observations increases. Belloni and Chernozhukov (2009) show that if the posterior density converges uniformly to a normal density, then a Metropolis-Hastings chain restricted to a neighborhood of the true parameter value has polynomially decaying convergence rate. Jones and Hobert (2001,2004) and other authors (e.g. Rosenthal 1995,1996) obtain bounds on the time to be within distance $\epsilon > 0$ of convergence for various hierarchical random effect models having unimodal posterior densities, as a function of the initial values, data, and hyperparameters. Mossel and Vigoda (2006) show that the convergence rate of a Bayesian phylogenetic Markov chain method can decrease exponentially in the number of samples in the dataset. We also learned after completing this article that Dr. Scott Schmidler at Duke University has independently obtained some convergence results in the motif-discovery context (personal communication).

Showing that a Markov chain method used in statistical practice is "well-behaved" usually consists of proving geometric ergodicity (Liu, Wong and Kong, 1995; Jarner and Hansen, 2000; Fort et al., 2003; Johnson and Jones, 2010), meaning that the chain converges to the posterior distribution at a geometric rate. The Gibbs sampler we analyze satisfies the even stronger property of uniform ergodicity; despite this, it is so poorly-behaved as to be unusable for obtaining samples from the posterior distribution for long DNA sequences.

Characterizing the dependence of the convergence rate on statistical quantities like the number of observations or the number of parameters is critical in justifying the use of a Markov chain method. However, there are several difficulties in doing so. First, the posterior distribution of a statistical

model has a much more complex form than the stylized distributions for which Markov chain convergence rates are typically obtained (Borgs et al. 1999; Bhatnagar and Randall 2004; Woodard, Schmidler, and Huber 2009). Second, the data, and thus the convergence rate of the Markov chain, are stochastic and depend on the data-generating mechanism.

We address these challenges by utilizing Bayesian asymptotic theory, which characterizes the behavior of the posterior distribution as the number of observations grows. This is complicated by the fact that Bayesian asymptotic theory is most well-developed in the case of a continuous parameter space, but the motif Gibbs sampler is defined on a discrete parameter space. We solve this by applying the asymptotic results on an alternative continuous parameterization of the motif model and then mapping those results to the discrete parameterization. Due to these technical challenges our main theorem requires sufficiently long motifs, and restricts to the case where each true motif corresponds to a fixed sequence of nucleotides (disallowing the small variations seen in Figure 1). We give an additional argument and simulation results suggesting that slow mixing holds even for very short motifs, and when the true motifs are allowed to vary between instances.

The motif discovery example provides insights into the dynamics of standard Markov chain methods applied to statistical models with highly multimodal posterior distributions. Other examples that may have the same exponential-time property include Markov chains for model search in the context of regression with a large number of predictors (Liang and Wong, 2000; Hans, Dobra and West, 2007) and Markov chains for spatial mixture models based on random fields (Geman and Geman, 1984; Green and Richardson, 2002). Our example also provides a test case for the use of more sophisticated Markov chain methods that are designed to handle multimodality (Del Moral, Doucet and Jasra, 2006; Andrieu, Doucet and Holenstein, 2010): if a method can be shown to sample from the posterior distribution of the motif-discovery model in polynomial time, then it would be dramatically more efficient than the Gibbs sampling approach.

Background on the Gibbs sampling method for motif discovery and on Markov chain convergence rates is in Section 2. Our convergence results are in Section 3, and a simulation study is given in Section 4. The proof of our main result is in Section 5, and we draw conclusions in Section 6. Additional proofs are in an online appendix.

2. Background.

2.1. *Statistical Motif Discovery.* The goal of motif discovery is to find short sub-sequences of nucleotides (length 6-15 base pairs) that occur multi-

ple times (more often than could be explained under the background model) in one or more long DNA sequences. Neither the nucleotide pattern nor the sub-sequence locations are known. This goal is illustrated in Figure 1.

We address one of the two main variants of Gibbs sampler used in motif discovery. The variant we analyze takes the number of motif instances per sequence to be unknown, while the other variant fixes the number of instances per sequence (Jensen et al., 2004); the two approaches are closely related and should have similar properties. Programs such as BioProspector are based on the method we analyze, and build in a number of additional features, such as a prior distribution on the motif frequency and handling of gapped motifs; however, by adding parameters and complexity to the model these enhancements probably make the Gibbs sampler slower to converge, and so are unlikely to affect our slow-mixing results.

We focus further on the case of a single DNA sequence (having an unknown number of motif instances). The case of multiple sequences can be addressed by concatenating to obtain a single sequence.

The motif instances are not necessarily identical. Taking the length w of the motif to be known, one can describe the nucleotide pattern by a position-specific frequency matrix, which contains the probability of occurrence of each nucleotide at each position in the motif. Call this matrix $\boldsymbol{\theta}_{1:w} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_w)$, where $\boldsymbol{\theta}_k$ is the unknown probability vector for the kth position. Let the nucleotides be labeled $1, \ldots, M$, so that $\boldsymbol{\theta}_k$ has length M; for DNA data $M = 4$. For each instance of the motif, the nucleotide in position k is assumed to be drawn independently from a discrete distribution with parameters $\boldsymbol{\theta}_k$. The positions in the full sequence that are not part of a motif instance are assumed to have nucleotide drawn independently from a discrete distribution with unknown probability vector $\boldsymbol{\theta}_0$.

Let $\mathbf{S} = (S_1, \ldots, S_L) \in \{1, \ldots, M\}^L$ be the observed sequence, having length L. In the original model of e.g. Liu, Neuwald and Lawrence (1995), a motif is allowed to start at any index $i \in \{1, \ldots, L - w + 1\}$, but we will analyze a simplified version that only allows a motif to start at indices $wi - w + 1$ for $i \in \{1, \ldots, L/w\}$ where L is divisible by w. This choice is explained in Section 2.2. Let A_i be the (unknown) indicator of whether a motif begins at index $wi - w + 1$, for $i \in \{1, \ldots, L/w\}$, and define $\mathbf{A} = (A_1, \ldots, A_{L/w})$. Let $\mathbf{N}(\mathbf{A}^{(k)})$ be the length-M vector of counts of the occurrence of each nucleotide at position k of all motif instances, conditional on $\mathbf{A}$. Similarly, $\mathbf{N}(\mathbf{A}^c)$ is defined to be the length-M vector of counts for each nucleotide in the background locations, i.e. the locations that are not part of any motif

instance:

(2.1)
$$N(\mathbf{A}^{(k)})_m \triangleq \sum_{i=1}^{L/w} \mathbf{1}_{\{A_i=1,\, S_{wi-w+k}=m\}} \qquad k \in \{1,\ldots,w\}, m \in \{1,\ldots,M\}$$
$$N(\mathbf{A}^c)_m \triangleq N(\mathbf{S})_m - \sum_{k=1}^{w} N(\mathbf{A}^{(k)})_m$$
$$N(\mathbf{S})_m \triangleq \sum_{i=1}^{L} \mathbf{1}_{\{S_i=m\}}.$$

For any two equal-length vectors $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_K)$ and $\mathbf{n} = (n_1,\ldots,n_K)$ define the notation

$$\boldsymbol{\beta}^{\mathbf{n}} = \prod_{k=1}^{K} \beta_k^{n_k} \qquad \Gamma(\mathbf{n}) = \prod_{k=1}^{K} \Gamma(n_k) \qquad |\mathbf{n}| = \sum_{k=1}^{K} n_k \tag{2.2}$$

where Γ is the gamma function. Using this notation, the full-data likelihood can be written as:

$$\begin{aligned}\pi(\mathbf{S}|\mathbf{A},\boldsymbol{\theta}_{0:w}) &= \prod_{i=1}^{L/w} \left[\prod_{k=1}^{w} \theta_{k,S_{wi-w+k}}\right]^{\mathbf{1}_{\{A_i=1\}}} \left[\prod_{k=1}^{w} \theta_{0,S_{wi-w+k}}\right]^{\mathbf{1}_{\{A_i=0\}}} \\ &= \boldsymbol{\theta}_0^{\mathbf{N}(\mathbf{A}^c)} \times \prod_{k=1}^{w} \boldsymbol{\theta}_k^{\mathbf{N}(\mathbf{A}^{(k)})}\end{aligned} \tag{2.3}$$

where $\boldsymbol{\theta}_{0:w} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_w)$ denotes all model parameters. We will use π to indicate the likelihood, prior, or the full, marginal, or conditional posterior distributions as distinguished by its arguments.

The prior distributions for the unknown quantities are $\pi(\boldsymbol{\theta}_k) = \text{Dirichlet}(\boldsymbol{\beta}_k)$ for $k \in \{0,\ldots,w\}$ and $\pi(A_i = 1) = p_0$ independently. Here $p_0 \in (0,1)$ is a known constant and $\boldsymbol{\beta}_0,\ldots,\boldsymbol{\beta}_w$ are fixed length-M vectors with $\beta_{k,m} > 0$. The corresponding posterior distribution is (Jensen et al., 2004):

$$\begin{aligned}\pi(\mathbf{A},\boldsymbol{\theta}_{0:w}|\mathbf{S}) &\propto \pi(\mathbf{A},\boldsymbol{\theta}_{0:w},\mathbf{S}) \\ &= p_0^{|\mathbf{A}|}(1-p_0)^{L/w-|\mathbf{A}|} \times \boldsymbol{\theta}_0^{\mathbf{N}(\mathbf{A}^c)+\boldsymbol{\beta}_0-1} \times \prod_{k=1}^{w} \boldsymbol{\theta}_k^{\mathbf{N}(\mathbf{A}^{(k)})+\boldsymbol{\beta}_k-1}.\end{aligned} \tag{2.4}$$

Liu, Neuwald and Lawrence (1995) integrate out the parameters $\boldsymbol{\theta}_{0:w}$ in the above formula to yield a posterior distribution on $\mathbf{A}$. Using the notation

from (2.2),

$$\pi(\mathbf{A}|\mathbf{S}) \propto p_0^{|\mathbf{A}|}(1-p_0)^{L/w-|\mathbf{A}|}\frac{\Gamma(\mathbf{N}(\mathbf{A}^c)+\boldsymbol{\beta}_0)}{\Gamma(|\mathbf{N}(\mathbf{A}^c)|+|\boldsymbol{\beta}_0|)}\prod_{k=1}^{w}\frac{\Gamma(\mathbf{N}(\mathbf{A}^{(k)})+\boldsymbol{\beta}_k)}{\Gamma(|\mathbf{N}(\mathbf{A}^{(k)})|+|\boldsymbol{\beta}_k|)}. \tag{2.5}$$

Liu (1994) gives theoretical results supporting faster convergence of a Gibbs sampler for the reduced posterior $\pi(\mathbf{A}|\mathbf{S})$ relative to a Gibbs sampler for $\pi(\mathbf{A}, \boldsymbol{\theta}_{0:w}|\mathbf{S})$. So Liu, Neuwald and Lawrence (1995) propose to use a Gibbs sampler to draw from $\pi(\mathbf{A}|\mathbf{S})$, having state vector $\mathbf{A} \in \mathcal{X} \triangleq \{0,1\}^{L/w}$. This sampler iteratively updates each A_i according to its conditional posterior distribution; details are given in Section 5.1.

Although this Gibbs sampler has both systematic-scan and random-scan versions, we expect that the mixing properties (defined in Section 2.3) of the two versions are identical. For this reason we focus attention on the random-scan Gibbs sampler, which is easier to analyze. We also make the transition matrix of the Markov chain nonnegative definite by including a holding probability of 1/2 at every state; this is a common technique when analyzing the mixing properties of Markov chains (Madras and Zheng, 2003; Woodard, Schmidler and Huber, 2009a). It only increases the mixing time (Section 2.3) by a factor of two, so does not affect results on the order of the run time as a function of L. Let $\mathbf{A}_{[-i]}$ indicate the vector $\mathbf{A}$ excluding the ith element, and $\pi(A_i|\mathbf{A}_{[-i]}, \mathbf{S}) \propto \pi(\mathbf{A}|\mathbf{S})$ indicate the conditional posterior distribution of A_i given $\mathbf{A}_{[-i]}$. With these definitions we can write the transition matrix T of the Gibbs sampler as follows for any $\mathbf{A}, \mathbf{A}' \in \mathcal{X}$.

$$T(\mathbf{A}, \mathbf{A}') \triangleq \frac{1}{L/w}\sum_{i=1}^{L/w}\mathbf{1}_{\{\mathbf{A}'_{[-i]}=\mathbf{A}_{[-i]}\}}\times \frac{1}{2}\left[\mathbf{1}_{\{A'_i=A_i\}} + \pi(A_i=0|\mathbf{A}_{[-i]},\mathbf{S})\mathbf{1}_{\{A'_i=0\}} + \pi(A_i=1|\mathbf{A}_{[-i]},\mathbf{S})\mathbf{1}_{\{A'_i=1\}}\right]. \tag{2.6}$$

Expressions for $\pi(A_i|\mathbf{A}_{[-i]}, \mathbf{S})$ are given in Section 5.1.

We also need an expression for the likelihood marginalized over $\mathbf{A}$. Define

the vector notation $\mathbf{S}_{n:m} = (S_n, \ldots, S_m)$ for any $n \leq m$. Then

$$\pi(\mathbf{S}|\boldsymbol{\theta}_{0:w}) = \prod_{i=1}^{L/w} f(\mathbf{S}_{wi-w+1:wi}|\boldsymbol{\theta}_{0:w}) \qquad \text{where} \tag{2.7}$$

$$f(\mathbf{s}|\boldsymbol{\theta}_{0:w}) \triangleq p_0 \prod_{k=1}^{w} \theta_{k,s_k} + (1-p_0) \prod_{k=1}^{w} \theta_{0,s_k} \qquad \mathbf{s} \in \{1, \ldots, M\}^w. \tag{2.8}$$

Under the model (2.7)-(2.8) each subsequence $\mathbf{S}_{wi-w+1:wi}$ is either generated from the motif $\boldsymbol{\theta}_{1:w}$ with probability p_0, or generated from the background $\boldsymbol{\theta}_0$ with probability $(1-p_0)$. So we have i.i.d. observations $\mathbf{S}_{wi-w+1:wi}$ for $i \in \{1, \ldots, L/w\}$, which will allow us to use Bayesian asymptotic theory for i.i.d. parametric models.

2.2. *Reason for the Simplification.* As stated in Section 2.1, while the original model allows a motif to start at any index $i \in \{1, \ldots, L-w+1\}$, we analyze a simplification of the model that assumes that motifs can only start at indices $wi - w + 1$ for $i \in \{1, \ldots, L/w\}$. This simplification is done to facilitate analysis; however, we believe that our results are likely to hold for the original model as well.

First, our rapid mixing result (Theorem 3.3) immediately holds for the original model and associated Gibbs sampler. This is because it is for the case of $w = 1$, where the original and simplified models are identical.

Second, the proof of our slow mixing result (Theorem 3.2) can be extended to the case where the model allows motifs to start at indices that are fixed distance $\geq w$ apart. However, Theorem 3.2 does not easily extend to the case where motifs can start in locations that are less than w distance apart (including the original model). This is due to the following "phase shift" issue, which complicates analysis. For illustration consider the case where $M = 4$ (there are four possible nucleotides) and $w = 5$ (motifs are five nucleotides long), and the true motif is (deterministically) the sequence $(1, 4, 2, 2, 3)$. Phase shift means that it is possible to estimate that a motif begins or ends in the middle of one of the $(1, 4, 2, 2, 3)$ subsequences that exist in the data. For example, if the DNA sequence $\mathbf{S}$ satisfies $\mathbf{S}_{22:26} = (1, 4, 2, 2, 3)$, corresponding to a true motif beginning at position 22, then the original model also allows for the possibility that $A_{23} = 1$, meaning that a motif could be estimated to instead start at position 23 with the sequence $4, 2, 2, 3$.

While phase shift complicates analysis of the original Gibbs sampler, it should also make the original Gibbs sampler converge more slowly than the simplified Gibbs sampler. The effect of phase shift on convergence of the original Gibbs sampler is that it can become trapped in a local mode of

the posterior distribution that corresponds to a shifted version of the true motif. This effect is described in Lawrence et al. (1993) and Liu (1994). To illustrate, take the above example where the true motif is $(1, 4, 2, 2, 3)$. There is a local mode of the posterior distribution for which the inferred motif starts with the sequence $4, 2, 2, 3$, another for which the inferred motif ends with the sequence $1, 4, 2, 2$, and so on. This posterior multimodality slows convergence of the original Gibbs sampler. It also suggests that our slow mixing result (Theorem 3.2) for the simplified Gibbs sampler holds for the original sampler.

Analysis of the original sampler should be possible using the same general approach taken here, but a number of the technical details would need to change. We leave this to future work.

2.3. *Markov Chain Convergence Rates.* Consider a Markov chain with transition matrix T and stationary distribution π on a discrete state space $\mathcal{X}$. For $x \in \mathcal{X}$ and $D \subset \mathcal{X}$ let $T(x, D) = \sum_{y \in D} T(x, y)$. If the chain is initialized at $x \in \mathcal{X}$ then the total variation distance to stationarity after n iterations is

$$\|T^n(x, \cdot) - \pi(\cdot)\|_{TV} \triangleq \max_{D \subset \mathcal{X}} |T^n(x, D) - \pi(D)|.$$

The mixing time of the chain is the number of iterations required to be within distance $\epsilon > 0$ of stationarity:

$$\tau_\epsilon \triangleq \max_{x \in \mathcal{X}} \min\{n : \|T^m(x, \cdot) - \pi(\cdot)\|_{TV} \leq \epsilon \text{ for all } m \geq n\}$$

(cf. Sinclair 1992).

Consider T irreducible, aperiodic, reversible and nonnegative definite, which holds for the random-scan Gibbs sampler in Section 2.1. Then τ_ϵ is finite and closely related to the spectral gap $\mathbf{Gap}(T) \triangleq 1 - \lambda_2$, where $\lambda_2 \in [0, 1)$ is the second-largest eigenvalue of T. Since the state space $\mathcal{X}$ is finite, $\mathbf{Gap}(T) > 0$ and the chain is called *uniformly ergodic* (Roberts and Rosenthal, 2004). The quantities τ_ϵ and $\mathbf{Gap}(T)$ are related via (Sinclair, 1992)

$$\tau_\epsilon \leq \mathbf{Gap}(T)^{-1}\big(-\ln[\min_{x \in \mathcal{X}} \pi(x)] - \ln \epsilon\big) \tag{2.9}$$

$$\tau_\epsilon \geq \frac{1}{2}(1 - \mathbf{Gap}(T))\mathbf{Gap}(T)^{-1}\big(-\ln(2\epsilon)\big).$$

The efficiency of the Markov chain can be measured by how quickly τ_ϵ increases as a function of the problem difficulty, for instance the dimension

of the parameter space. In our case we are interested in the dependence of τ_ϵ on the length L of the DNA sequence, since in practice one analyzes very long sequences. We would certainly hope that τ_ϵ grows at most polynomially in L for any fixed ϵ; this property is called *rapid mixing*. *Slow mixing* means that τ_ϵ increases exponentially for some ϵ. By (2.9) rapid mixing is equivalent to $\mathbf{Gap}(T)$ decreasing at most polynomially towards zero, and slow mixing is equivalent to $\mathbf{Gap}(T)$ decreasing exponentially towards zero, if $-\log[\min_{x\in\mathcal{X}} \pi(x)]$ increases polynomially in L. The latter property holds for the random-scan Gibbs sampler in Section 2.1. The rapid/slow mixing distinction is a measure of the computational tractability of an algorithm; polynomial factors are expected to be eventually dominated by increases in computing power due to Moore's Law, while exponential factors cause a persistent computational problem.

3. Convergence Results. We consider the mixing time (equivalently, spectral gap) of the Gibbs sampler when the data are drawn from a generalization of the model given in Section 2.1 that allows multiple true motifs. First we give negative results for the case of multiple true motifs, then we give a positive result for a case with no true motifs.

3.1. *Slow Mixing for Multiple True Motifs.* In this section, we show that if the data actually contain multiple true motifs, then the Gibbs sampler is slowly mixing: that $\mathbf{Gap}(T)$ is $\mathcal{O}(\alpha^{-L})$ for $\alpha > 1$ where $-\log[\min_{x\in\mathcal{X}} \pi(x)]$ is $\mathcal{O}(L^q)$ for some $q > 0$. To make this statement precise in the presence of random data, we need to make assumptions about the model by which the data are generated. Our convergence results are obtained using Bayesian asymptotics based on this generative model; for other connections between Markov chain convergence and Bayesian asymptotics see Kamatani (2011) and Belloni and Chernozhukov (2009).

For a concrete example to keep in mind, consider the case where $M = 4$ (there are four possible nucleotides) and $w = 5$ (motifs are five nucleotides long). Then let the DNA sequence $\mathbf{S}$ be generated as the concatenation of many length-five subsequences, each of which is either (motif one) equal to $(1, 4, 2, 2, 3)$ with probability 0.005, or (motif two) equal to $(4, 2, 4, 1, 3)$ with probability 0.001, or generated as i.i.d. noise where each nucleotide has equal probability. Theorem 3.2 below says that the Gibbs sampler T is slowly mixing for data generated in this way.

When analyzing the Gibbs sampler we do *not* assume that the data $\mathbf{S}$ are generated according to the inference model (2.7)-(2.8). Our most general result only assumes that the subsequences $\mathbf{S}_{wi-w+1:wi}$ are i.i.d..

ASSUMPTION 3.1. *The data subsequences* $\mathbf{S}_{wi-w+1:wi}$ *indexed by* $i \in \{1, \ldots, L/w\}$ *are independent and identically distributed according to some probability mass function* $g(\mathbf{s}) > 0 : \mathbf{s} \in \{1, \ldots, M\}^w$*; i.e.*

$$\mathbf{S} \sim \prod_{i=1}^{L/w} g(\mathbf{S}_{wi-w+1:wi}).$$

Under Assumption 3.1, we give a simple sufficient condition for slow mixing that relates the generative model $g(\mathbf{s})$ to the inference model $f(\mathbf{s}|\boldsymbol{\theta}_{0:w})$ via the quantity $E \log f(\mathbf{s}|\boldsymbol{\theta}_{0:w}) = \sum_{\mathbf{s}} g(\mathbf{s}) \log f(\mathbf{s}|\boldsymbol{\theta}_{0:w})$. Since $\boldsymbol{\theta}_k$ for $k \in \{0, \ldots, w\}$ is defined on the simplex $\Psi \triangleq \{\boldsymbol{\theta}_k : \sum_{m=1}^{M} \theta_{k,m} = 1, \theta_{k,m} \geq 0\}$, the quantity $E \log f(\mathbf{s}|\boldsymbol{\theta}_{0:w}) \in [-\infty, 0)$ is a function of $\boldsymbol{\theta}_{0:w} \in \Psi^{w+1}$. It is continuous in $\boldsymbol{\theta}_{0:w}$, because it is a linear combination of a finite number of the continuous functions $\log f(\mathbf{s}|\boldsymbol{\theta}_{0:w})$. We call $\eta(\boldsymbol{\theta}_{0:w}) \triangleq E \log f(\mathbf{s}|\boldsymbol{\theta}_{0:w})$ *multimodal* if there exist $\boldsymbol{\theta}_{0:w}^{(1)}, \boldsymbol{\theta}_{0:w}^{(2)} \in \Psi^{w+1}$ and bounded sets $F_1 \ni \boldsymbol{\theta}_{0:w}^{(1)}$ and $F_2 \ni \boldsymbol{\theta}_{0:w}^{(2)}$ such that

$$F_1 \cap F_2 = \emptyset \qquad \text{and} \qquad \sup_{\boldsymbol{\theta}_{0:w} \in \partial F_j} \eta(\boldsymbol{\theta}_{0:w}) < \eta(\boldsymbol{\theta}_{0:w}^{(j)}) \qquad j \in \{1, 2\} \tag{3.1}$$

where $\partial F_j \triangleq \mathrm{cl}(F_j) \cap \mathrm{cl}(F_j^c)$ is the boundary of F_j. Equation (3.1) implies that $\boldsymbol{\theta}_{0:w}^{(j)}$ is in the interior of F_j. For a continuous function on a closed, connected subset of $\mathbb{R}^d$ (like Ψ^{w+1}) this definition of multimodality is weaker than the existence of multiple strict local maxima and stronger than the existence of multiple local maxima. We call a function h of $\boldsymbol{\theta}_{0:w}$ a *multiminimum function* if $-h(\boldsymbol{\theta}_{0:w})$ is multimodal.

THEOREM 3.1. *Under Assumption 3.1, if the function* $E \log f(\mathbf{s}|\boldsymbol{\theta}_{0:w})$ *of* $\boldsymbol{\theta}_{0:w}$ *is multimodal then the spectral gap of the Gibbs sampler* T *decreases exponentially in* L*, almost surely.*

Theorem 3.1 will be proven in Section 5. It uses asymptotic results on the behavior of the posterior when the data are not distributed according to the inference model (Berk, 1966). We will see that when $E \log f(\mathbf{s}|\boldsymbol{\theta}_{0:w})$ is multimodal the posterior distribution is also multimodal for large L, and that the heights of the modes relative to the heights of the valleys in between grow exponentially in L, causing the slow mixing. This is due to the fact that, using (2.7), the log-likelihood is

$$\log \pi(\mathbf{S}|\boldsymbol{\theta}_{0:w}) = \sum_{i=1}^{L/w} \log f(\mathbf{S}_{wi-w+1:wi}|\boldsymbol{\theta}_{0:w})$$

which satisfies the following for any $\boldsymbol{\theta}_{0:w}$.

$$\frac{1}{L/w}\sum_{i=1}^{L/w}\log f(\mathbf{S}_{wi-w+1:wi}|\boldsymbol{\theta}_{0:w}) \overset{L\to\infty}{\longrightarrow} E\log f(\mathbf{s}|\boldsymbol{\theta}_{0:w}) \qquad \text{a.s.}$$

by the Strong Law of Large Numbers. This effect leads to the likelihood function being multimodal for large L. Statistically, these correspond to multiple values of $\boldsymbol{\theta}_{0:w}$ that explain the data well.

Another way of stating Theorem 3.1 is via the Kullback-Leibler divergence between $f(\mathbf{s}|\boldsymbol{\theta}_{0:w})$ and $g(\mathbf{s})$. The divergence measures the degree of difference between $f(\mathbf{s}|\boldsymbol{\theta}_{0:w})$ and $g(\mathbf{s})$ and is defined as

$$\begin{aligned}(3.2)\qquad &\sum_{\mathbf{s}} g(\mathbf{s})\log\frac{g(\mathbf{s})}{f(\mathbf{s}|\boldsymbol{\theta}_{0:w})}\\ &=\sum_{\mathbf{s}} g(\mathbf{s})\log g(\mathbf{s}) - \sum_{\mathbf{s}} g(\mathbf{s})\log f(\mathbf{s}|\boldsymbol{\theta}_{0:w}).\end{aligned}$$

Since $\sum_{\mathbf{s}} g(\mathbf{s})\log g(\mathbf{s})$ does not depend on $\boldsymbol{\theta}_{0:w}$, the divergence is a multi-minimum function iff $E\log f(\mathbf{s}|\boldsymbol{\theta}_{0:w})$ is multimodal.

COROLLARY 3.1. *Under Assumption 3.1, if the Kullback-Leibler divergence (3.2) is a multiminimum function of $\boldsymbol{\theta}_{0:w}$ then the spectral gap of the Gibbs sampler T decreases exponentially in L, almost surely.*

Next we show that multimodality of $E\log f(\mathbf{s}|\boldsymbol{\theta}_{0:w})$ occurs when the generative model $g(\mathbf{s})$ includes $J>1$ true motifs, described by position-specific frequency matrices $\boldsymbol{\theta}^{j*}_{1:w}$ for $j\in\{1,\dots,J\}$. Assumption 3.2 says that $g(\mathbf{s})$ is obtained by extending the inference model (2.8) to the case of $J>1$ motifs.

ASSUMPTION 3.2. *The p.m.f. $g(\mathbf{s})$ from Assumption 3.1 satisfies*

$$(3.3)\qquad g_{\boldsymbol{\theta}^*}(\mathbf{s}) = \sum_{j=1}^{J} p_j \prod_{k=1}^{w}\theta^{j*}_{k,s_k} + \left(1-\sum_{j=1}^{J}p_j\right)\prod_{k=1}^{w}\theta^{*}_{0,s_k}$$

where $J>1$ and

1. *$p_j>0$ for $j\in\{1,\dots,J\}$ are the motif frequencies, where $\sum_{j=1}^{J}p_j<1$*
2. *$\boldsymbol{\theta}^*_0$ is a background probability vector with $\sum_{m=1}^{M}\theta^*_{0,m}=1$ and $\theta^*_{0,m}>0$*
3. *$\boldsymbol{\theta}^{j*}_{1:w}$ for $j\in\{1,\dots,J\}$ are position-specific frequency matrices.*

Due to the complex form of $E \log f(\mathbf{s}|\boldsymbol{\theta}_{0:w}) = \sum_{\mathbf{s}} g_{\boldsymbol{\theta}^*}(\mathbf{s}) \log f(\mathbf{s}|\boldsymbol{\theta}_{0:w})$ under Assumptions 3.1-3.2 it is difficult to characterize the number of modes without making any additional assumption. We restrict here to the case where each true motif is a fixed length-w sequence of nucleotides, for example where $w = 5$ and $M = 4$ and the first and second motifs correspond to the deterministic sequences $(1, 4, 2, 2, 3)$ and $(4, 2, 4, 1, 3)$, respectively. Mathematically this means that for each $j \in \{1, \ldots, J\}$ and $k \in \{1, \ldots, w\}$ there is some $m \in \{1, \ldots, M\}$ for which $\theta^{j*}_{k,m} = 1$. The case without this restriction is discussed below.

ASSUMPTION 3.3. *For each $j \in \{1, \ldots, J\}$ and each $k \in \{1, \ldots, w\}$ there is some $t^j_k \in \{1, \ldots, M\}$ for which $\theta^{j*}_{k,t^j_k} = 1$. Also,*

$$a \triangleq \min_{j \neq j'} \liminf_{w \to \infty} \frac{1}{w} \sum_{k=1}^{w} \mathbf{1}_{\left\{t^j_k \neq t^{j'}_k\right\}} > 0. \tag{3.4}$$

Assumption 3.3 says that the motifs are deterministic in the above sense, and that for any two motifs $j \neq j'$ the proportion of differences between the motif sequences does not decay to zero as w grows. This ensures that the motifs are different enough from one another for large w to cause a mixing problem in the Markov chain. With these assumptions, we have multimodality of $E \log f(\mathbf{s}|\boldsymbol{\theta}_{0:w})$ for large enough w.

LEMMA 3.1. *Under Assumptions 3.1-3.3 there exists $w^* < \infty$ that depends on p_0, $\boldsymbol{\theta}^*_0$, J, $\{p_j\}_{j=1}^J$, and a such that the following holds. If $w \geq w^*$ then $E \log f(\mathbf{s}|\boldsymbol{\theta}_{0:w})$ is multimodal with at least $J > 1$ local maxima.*

Lemma 3.1 is proven in the Web Appendix. Combining Theorem 3.1 and Lemma 3.1 immediately yields our main result: slow mixing for the case of multiple true motifs.

THEOREM 3.2. *Under Assumptions 3.1-3.3, there exists $w^* < \infty$ such that whenever $w \geq w^*$ the spectral gap of the Gibbs sampler T decreases exponentially in L, almost surely.*

While Theorem 3.2 is stated for w large enough and assumes deterministic true motifs, the simulation results in Section 4 suggest that slow mixing occurs even for non-deterministic true motifs and for w as low as six.

Theorem 3.2 says that the presence of multiple motifs in the generative model contradicts the inference model assumption of a single motif, causing slow mixing. In realistic biological situations there are frequently multiple

motifs, corresponding to multiple gene regulatory binding motifs or to repeating patterns that have other biological significance (Neuwald, Liu and Lawrence, 1995; Roth et al., 1998). Theorem 3.2 says that these patterns do not have to occur often in order to cause slow mixing (that slow mixing occurs even when some of the p_j are very small). So when L is large the Gibbs sampler should be used only as a tool for generating candidate motifs, and the results cannot be interpreted as obtaining samples from the posterior distribution, or used for Monte Carlo estimation.

Theorem 3.2 assumes that each true motif is a deterministic sequence; now consider the case of variable true motifs, i.e. where Assumptions 3.1-3.2 hold but not Assumption 3.3. We give an informal argument suggesting that the function $E \log f(\mathbf{s}|\boldsymbol{\theta}_{0:w})$ is still multimodal.

Consider the case where $\sum_{j=1}^{J} p_j = p_0$. Then, using (2.8) and (3.3),

$$\begin{aligned} g_{\boldsymbol{\theta}^*}(\mathbf{s}) &= \sum_{j=1}^{J} \frac{p_j}{p_0} \left[p_0 \prod_{k=1}^{w} \theta^{j*}_{k,s_k} + (1-p_0) \prod_{k=1}^{w} \theta^{*}_{0,s_k} \right] \\ &= \sum_{j=1}^{J} \frac{p_j}{p_0} f(\mathbf{s}|(\boldsymbol{\theta}^*_0, \boldsymbol{\theta}^{j*}_{1:w})). \end{aligned}$$

So $E \log f(\mathbf{s}|\boldsymbol{\theta}_{0:w})$ can be written as

$$\sum_{\mathbf{s}} g_{\boldsymbol{\theta}^*}(\mathbf{s}) \log f(\mathbf{s}|\boldsymbol{\theta}_{0:w}) = \sum_{j=1}^{J} \frac{p_j}{p_0} \sum_{\mathbf{s}} f(\mathbf{s}|(\boldsymbol{\theta}^*_0, \boldsymbol{\theta}^{j*}_{1:w})) \log f(\mathbf{s}|\boldsymbol{\theta}_{0:w}). \tag{3.5}$$

By standard information-theoretic results (Kullback, 1959; Berk, 1966), for each j the function $\sum_{\mathbf{s}} f(\mathbf{s}|(\boldsymbol{\theta}^*_0, \boldsymbol{\theta}^{j*}_{1:w})) \log f(\mathbf{s}|\boldsymbol{\theta}_{0:w})$ has a unique global maximum at $\boldsymbol{\theta}_{0:w} = (\boldsymbol{\theta}^*_0, \boldsymbol{\theta}^{j*}_{1:w})$. Since (3.5) is the weighted sum of continuous functions that have global maxima occurring at distinct locations $(\boldsymbol{\theta}^*_0, \boldsymbol{\theta}^{j*}_{1:w})$, it seems likely that (3.5) is multimodal when $J > 1$.

3.2. *Rapid Mixing for ≤ 1 True Motif.* Simulations suggest (Section 4) that when there is no more than one true motif the Gibbs sampler is rapidly mixing, i.e. $\mathbf{Gap}(T)^{-1} = \mathcal{O}(L^q)$ for some $q > 0$. We have one theoretical result in this direction, showing rapid mixing for the case $w = 1$. In this case any true motif is indistinguishable from the background signal, so there are effectively zero true motifs.

THEOREM 3.3. *If $w = 1$ then the Gibbs sampler T has spectral gap that decreases polynomially in L, uniformly over $\mathbf{S} \in \{1, \ldots, M\}^L$. Specifically,*

for $M = 2$

$$\sup_{\mathbf{S} \in \{1,\ldots,M\}^L} \mathbf{Gap}(T)^{-1} = \mathcal{O}(L^{14})$$

and for fixed $M > 2$ *the same result holds for a larger-degree polynomial.*

Theorem 3.3 (proven in the Web Appendix) shows rapid mixing in the worst case over possible datasets $\mathbf{S}$. Contrast with Theorems 3.1 and 3.2, which show slow mixing almost surely under a particular generative model $g(\mathbf{s})$.

It is likely that the spectral gap bound given in Theorem 3.3 is very loose as a function of L, since the tools that we use to obtain it (Theorem B.4 in particular) can be imprecise. However, obtaining a tighter bound would require substantially longer arguments, so we leave this to future work. Additionally, one could assume a particular distribution for $\mathbf{S}$ and use an average-case analysis to obtain a tighter bound, but also one that would have narrower interpretation.

4. Simulation Study. We simulate data with either $J = 1$ or $J = 2$ true motifs, and measure the convergence of the Gibbs sampler. The data are simulated as follows; to emulate DNA data we take $M = 4$. The true position-specific frequency matrix $\boldsymbol{\theta}^{j*}_{1:w}$ for each motif j is obtained by drawing its columns $\boldsymbol{\theta}^{j*}_k$ independently for $k \in \{1, \ldots, w\}$ from a Dirichlet distribution with parameter vector $\boldsymbol{\alpha}_1$ (chosen as described below). The background frequency vector $\boldsymbol{\theta}^*_0$ is drawn from a Dirichlet distribution with parameter vector $\boldsymbol{\alpha}_0$. We also define the motif frequency to be $p_j = 0.005$ for each motif j (a typical value in practice). With these definitions, the data vector $\mathbf{S}$ is obtained by drawing each subsequence $\mathbf{S}_{wi-w+1:wi}$ for $i \in \{1, \ldots, L/w\}$ from (3.3), using various combinations of w and L. Unlike Assumption 3.3, here we use variable motifs, meaning that $\theta^{j*}_{k,m} \neq 1$ for all k and m. We have also done experiments with other values of p_j ($p_j = 0.003$ and $p_j = 0.02$), which gave qualitatively the same results.

We choose $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_1$ so that the distribution of $\boldsymbol{\theta}^*_0$ and that of $\boldsymbol{\theta}^{j*}_k$ is symmetric in the four nucleotides; this means that we must have $\boldsymbol{\alpha}_0 = a_0 \times (1, 1, 1, 1)$ and $\boldsymbol{\alpha}_1 = a_1 \times (1, 1, 1, 1)$ for some $a_0, a_1 > 0$. Since motifs are by definition fairly well-conserved, we choose a_1 so that the median of $\max_{m \in \{1,\ldots,4\}} \theta^{j*}_{k,m}$ is 0.95 (a_1 is found numerically). Since background data are typically more balanced among the four nucleotides, we choose a_0 so that the median of $\max_{m \in \{1,\ldots,4\}} \theta^*_{0,m}$ is 0.3.

For each simulated data vector $\mathbf{S}$ we run a systematic-scan Gibbs sampler five times from different initial values and use the Gelman-Rubin scale factor

(Gelman and Rubin, 1992) to detect whether the chains have converged to different parts of the parameter space, corresponding to different local modes of the posterior density. Since the slow mixing in Theorem 3.2 is caused by multimodality of the posterior distribution, this approach should detect the problem effectively. If different runs of the Markov chain explore different parts of the parameter space, the Gelman-Rubin scale factor should be large (typically much larger than 2), while if they are drawing from the same distribution the scale factor should be close to 1.

In order to detect the worst-case behavior, we take the initial vector $\mathbf{A} = (A_1, \ldots, A_{L/w})$ for the first chain to be the vector of indicators of whether each subsequence $\mathbf{S}_{wi-w+1:wi}$ was generated from motif one. If applicable we initialize the second chain at the vector of indicators of whether each subsequence $\mathbf{S}_{wi-w+1:wi}$ was generated from motif two. The initial vector $\mathbf{A}$ in all other cases is generated randomly according to $A_i \overset{\text{iid}}{\sim} \text{Bernoulli}(p_0)$. Although in practice one would not know the true motif locations, we do this to ensure that we detect even very narrow and hard-to-find modes corresponding to the true motifs. We run each Gibbs sampler for a burn-in period of $1,000$ updates of the entire vector $\mathbf{A}$, and then a sampling period of $10,000$ updates of $\mathbf{A}$. With these choices standard convergence diagnostics (cf. Geweke 1992) that evaluate the convergence of the chains individually do not detect a convergence problem.

When we run the Gibbs sampler we specify the inference model motif frequency p_0 as $p_0 = \sum_{j=1}^{J} p_j$ (other choices are investigated below). We specify the prior hyperparameters as $\beta_{k,m} = 1$ for $k \in \{0, \ldots, w\}$ and $m \in \{1, \ldots, 4\}$; this is the standard choice.

Having simulated the chains, we calculate the Gelman-Rubin scale factor for the following parameter summaries:

$$\hat{\theta}_{k,m}(\mathbf{A}) \triangleq \frac{N(\mathbf{A}^{(k)})_m + \beta_{k,m}}{|\mathbf{N}(\mathbf{A}^{(k)})| + |\boldsymbol{\beta}_k|} \qquad k \in \{1, \ldots, w\}, m \in \{1, \ldots, 4\}$$

$$\hat{\theta}_{0,m}(\mathbf{A}) \triangleq \frac{N(\mathbf{A}^{c})_m + \beta_{0,m}}{|\mathbf{N}(\mathbf{A}^{c})| + |\boldsymbol{\beta}_0|}$$

as well as for $|\mathbf{A}|$, recalling the notation (2.1) and (2.2). The values $\hat{\theta}_{k,m}(\mathbf{A})$ and $\hat{\theta}_{0,m}(\mathbf{A})$ are relevant because they are the posterior means of $\theta_{k,m}$ and $\theta_{0,m}$ given $\mathbf{A}$. The posterior density estimates of $\hat{\theta}_{2,1}(\mathbf{A})$ from two different Markov chains in the case $J = 2$ are shown in Figure 2. The Gelman-Rubin scale factor for these chains is 10.9, accurately reflecting the fact that the two chains have converged to different parts of the parameter space.

The top display in Table 1 addresses the case of one true motif ($J = 1$) and various combinations of w and L. For each combination it reports the

$J = 1$	$w = 6$	$w = 10$	$w = 15$
$L/w = 2,000$	0	0	0
$L/w = 3,000$	0	0	0
$L/w = 4,000$	0	0	0
$L/w = 8,000$	0	0	0
$J = 2$	$w = 6$	$w = 10$	$w = 15$
$L/w = 2,000$	0	20	70
$L/w = 3,000$	10	70	100
$L/w = 4,000$	20	80	100
$L/w = 8,000$	80	90	100

TABLE 1

For the cases of one true motif (top) and two true motifs (bottom), the percentage of simulated datasets for which the Gelman-Rubin scale factor from five Gibbs sampling chains is > 1.5.

percentage out of 20 simulated datasets for which the maximum Gelman-Rubin scale factor (over the different parameter summaries) is greater than 1.5. The bottom display in Table 1 reports the same quantities for the case of two true motifs ($J = 2$). For one motif no convergence problem is detected for any of the simulated datasets. For two motifs, regardless of the value of w, there is a severe convergence problem for large values of L.

Finally, we investigate the effect of other choices for p_0. Specifying $p_0 = 0.005$, $p_0 = 0.002$, or $p_0 = 0.02$ yields results that are qualitatively the same as those in Table 1.

5. Proof of Theorem 3.1.

5.1. *Specification of the Gibbs Sampler.* Here we give the details of the Gibbs sampler T. Recalling the notation of Section 2.1, the sampler iteratively updates each A_i according to its conditional posterior distribution, given as follows where $\mathbf{A}_{[-i]}$ refers to the vector $\mathbf{A}$ excluding the ith element, where $\mathbf{A}_{[i,0]}$ is the vector $\mathbf{A}$ with the ith element replaced by 0, and where

$\mathbf{A}_{[i,1]}$ is $\mathbf{A}$ with the ith element replaced by 1. Using (2.5),

$$\begin{aligned}
&\frac{\pi(A_i=1|\mathbf{A}_{[-i]},\mathbf{S})}{\pi(A_i=0|\mathbf{A}_{[-i]},\mathbf{S})} \qquad\qquad i\in\{1,\dots L/w\}\\
&= \frac{p_0}{1-p_0}\left(\frac{\Gamma(\mathbf{N}(\mathbf{A}^c_{[i,1]})+\boldsymbol{\beta}_0)}{\Gamma(\mathbf{N}(\mathbf{A}^c_{[i,0]})+\boldsymbol{\beta}_0)}\right)\frac{\Gamma(|\mathbf{N}(\mathbf{A}^c_{[i,0]})|+|\boldsymbol{\beta}_0|)}{\Gamma(|\mathbf{N}(\mathbf{A}^c_{[i,1]})|+|\boldsymbol{\beta}_0|)}\times \\
&\qquad \prod_{k=1}^{w}\frac{N(\mathbf{A}^{(k)}_{[i,0]})_{S_{wi-w+k}}+\beta_{k,S_{wi-w+k}}}{|\mathbf{N}(\mathbf{A}^{(k)}_{[i,0]})|+|\boldsymbol{\beta}_k|} \qquad (5.1)\\
&= \frac{p_0}{1-p_0}\left(\frac{\Gamma(\mathbf{N}(\mathbf{A}^c_{[i,1]})+\boldsymbol{\beta}_0)}{\Gamma(\mathbf{N}(\mathbf{A}^c_{[i,0]})+\boldsymbol{\beta}_0)}\right)\frac{\Gamma(|\mathbf{N}(\mathbf{A}^c_{[i,0]})|+|\boldsymbol{\beta}_0|)}{\Gamma(|\mathbf{N}(\mathbf{A}^c_{[i,1]})|+|\boldsymbol{\beta}_0|)}\prod_{k=1}^{w}\check{\theta}_{k,S_{wi-w+k}}
\end{aligned}$$

where the elements of the vector $\check{\boldsymbol{\theta}}_k$ are the current estimates of the frequency of each nucleotide in position k of the motif, i.e.:

$$\check{\theta}_{k,m} \triangleq \frac{N(\mathbf{A}^{(k)}_{[i,0]})_m+\beta_{k,m}}{|\mathbf{N}(\mathbf{A}^{(k)}_{[i,0]})|+|\boldsymbol{\beta}_k|} \qquad k\in\{1,\dots,w\}, m\in\{1,\dots,M\}. \qquad (5.2)$$

For details see Liu, Neuwald and Lawrence (1995) and Jensen et al. (2004).

5.2. *Outline of Proof of Thm. 3.1.* Informally, the proof of Theorem 3.1 proceeds by showing the following results.

Step 1. The spectral gap of the Gibbs sampler is determined by the unimodality or multimodality of the marginal posterior distribution of a particular summary vector $\mathbf{C}(\mathbf{A})$ of $\mathbf{A}$, denoted by $\bar{\pi}(\mathbf{C}(\mathbf{A})|\mathbf{S})$. If $\bar{\pi}(\mathbf{C}(\mathbf{A})|\mathbf{S})$ is multimodal the spectral gap is determined by the heights of the modes relative to the heights of the valleys between the modes.

Step 2. When $E\log f(\mathbf{s}|\boldsymbol{\theta}_{0:w})$ is multimodal, the marginal posterior distribution $\pi(\boldsymbol{\theta}_{0:w}|\mathbf{S})$ of the continuous parameters $\boldsymbol{\theta}_{0:w}$ is also multimodal, with height of the modes increasing exponentially in L, relative to the height of the valleys between the modes.

Step 3. The result of Step 2 can be mapped to $\bar{\pi}(\mathbf{C}(\mathbf{A})|\mathbf{S})$, showing that the posterior distribution of $\mathbf{C}(\mathbf{A})$ has multiple modes with height that grows exponentially in L (relative to the valleys in between).

For simplicity of notation we consider the case $M=2$ (two nucleotides), although the proof is analogous for any fixed M.

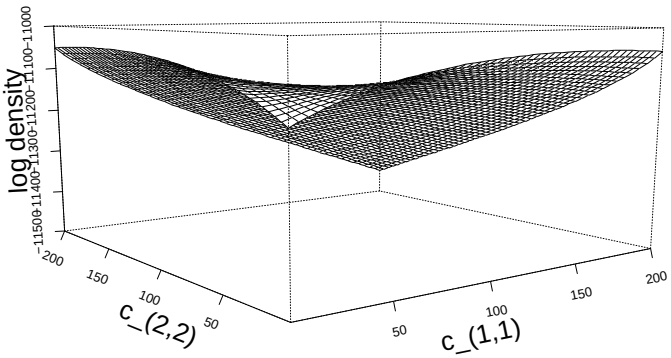

FIG 3. *The log-density* $\log \bar{\pi}$ *as a function of* $c_{(1,1)}$ *and* $c_{(2,2)}$*, fixing* $c_{(1,2)} = c_{(2,1)} = 0$ *and for the case* $w = 2$.

Formally, for $\mathbf{s} \in \{1,2\}^w$ any length-w vector of nucleotides define

$$C(\mathbf{A})_{\mathbf{s}} \triangleq |\{i : A_i = 1, \mathbf{S}_{wi-w+1:wi} = \mathbf{s}\}| \tag{5.3}$$

to be the number of instances of motif $\mathbf{s}$ (where $|\{\ldots\}|$ indicates the cardinality of a set). Similarly, let

$$C(\mathbf{S})_{\mathbf{s}} \triangleq |\{i : \mathbf{S}_{wi-w+1:wi} = \mathbf{s}\}| \tag{5.4}$$

be the number of times that the sequence of nucleotides $\mathbf{s}$ occurs in the data. Then we must have $C(\mathbf{A})_{\mathbf{s}} \leq C(\mathbf{S})_{\mathbf{s}}$ for each $\mathbf{s}$, i.e. $\mathbf{C}(\mathbf{A})$ lies in the space

$$\bar{\mathcal{X}} \triangleq \prod_{\mathbf{s} \in \{1,2\}^w} \{0, \ldots, C(\mathbf{S})_{\mathbf{s}}\}. \tag{5.5}$$

The posterior distribution $\pi(\mathbf{A}|\mathbf{S})$ only depends on $\mathbf{A}$ through $\mathbf{C}(\mathbf{A})$, which can be seen as follows. Using (2.5), $\pi(\mathbf{A}|\mathbf{S})$ depends on $\mathbf{A}$ via the quantities $|\mathbf{A}|$, $\mathbf{N}(\mathbf{A}^{(k)})$, and $\mathbf{N}(\mathbf{A}^c)$. These in turn only depend on $\mathbf{C}(\mathbf{A})$, since (using (2.1)-(2.2) and (5.3))

$$\begin{aligned} |\mathbf{A}| &= |\mathbf{C}(\mathbf{A})| \\ N(\mathbf{A}^{(k)})_m &= \sum_{\mathbf{s}} C(\mathbf{A})_{\mathbf{s}} \mathbf{1}_{\{s_k = m\}} \qquad k \in \{1, \ldots, w\}, m \in \{1,2\} \\ N(\mathbf{A}^c)_m &= N(\mathbf{S})_m - \sum_{k=1}^{w} N(\mathbf{A}^{(k)})_m. \end{aligned} \tag{5.6}$$

The marginal posterior distribution of $\mathbf{C}(\mathbf{A})$ is denoted by

$$\bar{\pi}(\mathbf{c}|\mathbf{S}) \triangleq \sum_{\mathbf{A}:\mathbf{C}(\mathbf{A})=\mathbf{c}} \pi(\mathbf{A}|\mathbf{S}) \qquad \mathbf{c} \in \bar{\mathcal{X}}. \tag{5.7}$$

Figure 3 illustrates multimodality of $\bar{\pi}(\mathbf{c}|\mathbf{S})$ for the case $w = 2$. For $w = 2$ the arguments to the function $\bar{\pi}$ are the quantities $c_{(1,1)}$, $c_{(1,2)}$, $c_{(2,1)}$, and $c_{(2,2)}$. The data $\mathbf{S}$ used to create Figure 3 were generated with two true motifs, yielding the two visible modes of $\bar{\pi}$.

We will use Theorem B.1 to bound $\mathbf{Gap}(T)$. Partition the state space $\mathcal{X}$ of T according to the value of $\mathbf{C}(\mathbf{A})$.

$$D_{\mathbf{c}} \triangleq \{\mathbf{A} \in \mathcal{X} : \mathbf{C}(\mathbf{A}) = \mathbf{c}\} \qquad \mathbf{c} \in \bar{\mathcal{X}}. \tag{5.8}$$

Define the projection matrix (Appendix B) for T with respect to this partition:

$$\begin{aligned} \bar{T}(\mathbf{c}_1, \mathbf{c}_2) &\triangleq \frac{\sum_{\mathbf{A}:\mathbf{C}(\mathbf{A})=\mathbf{c}_1} \pi(\mathbf{A}|\mathbf{S}) T(\mathbf{A}, D_{\mathbf{c}_2})}{\sum_{\mathbf{A}:\mathbf{C}(\mathbf{A})=\mathbf{c}_1} \pi(\mathbf{A}|\mathbf{S})} \qquad \mathbf{c}_1, \mathbf{c}_2 \in \bar{\mathcal{X}} \\ &= \sum_{\mathbf{A}:\mathbf{C}(\mathbf{A})=\mathbf{c}_1} \frac{1}{|D_{\mathbf{c}_1}|} T(\mathbf{A}, D_{\mathbf{c}_2}) \end{aligned} \tag{5.9}$$

since $\pi(\mathbf{A}|\mathbf{S})$ depends on $\mathbf{A}$ only via $\mathbf{C}(\mathbf{A})$, so that $\pi(\mathbf{A}|\mathbf{S})$ is equal for all $\mathbf{A} \in D_{\mathbf{c}_1}$. The matrix $\bar{T}$ is reversible with respect to $\bar{\pi}$ (Appendix B).

It is easier to obtain useful bounds on $\mathbf{Gap}(T)$ indirectly by relating T to $\bar{T}$ and bounding $\mathbf{Gap}(\bar{T})$ than it is to obtain useful bounds on $\mathbf{Gap}(T)$ directly. The same technique is utilized in Madras and Zheng (2003) and Woodard, Schmidler and Huber (2009a). The cardinality of the state space $\mathcal{X} = \{0, 1\}^{L/w}$ of T grows exponentially in L for fixed w, while the cardinality of the state space $\bar{\mathcal{X}}$ of $\bar{T}$ grows only polynomially in L, since (using (5.5))

$$|\bar{\mathcal{X}}| \leq (L/w + 1)^{2^w}. \tag{5.10}$$

We obtain an upper bound on $\mathbf{Gap}(\bar{T})$ using conductance (Theorem B.2) and a lower bound on $\mathbf{Gap}(\bar{T})$ using path bounds (Theorem B.4). Bounds obtained using these tools can easily be inaccurate by a factor equal to the cardinality of the space. If we were to obtain these bounds directly for $\mathbf{Gap}(T)$ they would be loose by an exponential factor in L, and thus unusable for our purposes.

5.3. *Step 1 of Proof of Thm. 3.1.* Consider the graph associated with the reversible matrix $\bar{T}$, with vertices corresponding to $\mathbf{c} \in \bar{\mathcal{X}}$ and edges corresponding to pairs $\mathbf{c}_1, \mathbf{c}_2 \in \bar{\mathcal{X}}$ having $\bar{T}(\mathbf{c}_1, \mathbf{c}_2) > 0$. For any $\mathbf{c}_1, \mathbf{c}_2 \in \bar{\mathcal{X}}$ let $\Gamma_{\mathbf{c}_1,\mathbf{c}_2}$ denote the set of paths between $\mathbf{c}_1$ and $\mathbf{c}_2$ in the graph that do not have repeated vertices. Also let $\mathbf{c} \in \gamma$ indicate that the state $\mathbf{c} \in \bar{\mathcal{X}}$ is a vertex in the path γ. Then Theorem 5.1 formalizes Step 1 from Section 5.2.

THEOREM 5.1. $\mathbf{Gap}(T)$ *decreases exponentially in* L *iff*

$$d \triangleq \min_{\mathbf{c}_1,\mathbf{c}_2 \in \bar{\mathcal{X}}} \max_{\gamma \in \Gamma_{\mathbf{c}_1,\mathbf{c}_2}} \min_{\mathbf{c} \in \gamma} \frac{\bar{\pi}(\mathbf{c}|\mathbf{S})}{\bar{\pi}(\mathbf{c}_1|\mathbf{S})\bar{\pi}(\mathbf{c}_2|\mathbf{S})} \tag{5.11}$$

decreases exponentially in L.

The quantity d measures the multimodality of $\bar{\pi}$. Roughly, think of $\mathbf{c}_1, \mathbf{c}_2$ as being local modes of $\bar{\pi}$; if all paths between $\mathbf{c}_1$ and $\mathbf{c}_2$ contain a state with low probability then d is small.

PROOF. The transition matrix T is nonnegative definite and reversible with respect to $\pi(\mathbf{A}|\mathbf{S})$. Notice that T^2 is also reversible w.r.t. $\pi(\mathbf{A}|\mathbf{S})$. Using (5.8), let $T^2|_{D_\mathbf{c}}$ be the restriction of T^2 to $D_\mathbf{c}$ as defined in Appendix B. Then by Lemma B.1 and Theorem B.1,

$$\begin{aligned} \mathbf{Gap}(T) &\geq \frac{1}{3}\mathbf{Gap}(T^3) = \frac{1}{3}\mathbf{Gap}(T^{1/2}T^2T^{1/2}) \\ &\geq \frac{1}{3}\mathbf{Gap}(\bar{T}) \min_{\mathbf{c} \in \bar{\mathcal{X}}} \mathbf{Gap}(T^2|_{D_\mathbf{c}}). \end{aligned} \tag{5.12}$$

Combining with Proposition 5.1 below, we have that

$$\mathbf{Gap}(T)^{-1} = \mathcal{O}(\mathbf{Gap}(\bar{T})^{-1} \times L^{2w+3}). \tag{5.13}$$

Theorem B.1 also gives the bound $\mathbf{Gap}(T) \leq \mathbf{Gap}(\bar{T})$. So $\mathbf{Gap}(T)$ is within a polynomial (in L) factor of $\mathbf{Gap}(\bar{T})$. Theorem 5.1 then follows from Proposition 5.2 below. □

Finally we give several results used in the proof of Theorem 5.1.

PROPOSITION 5.1. *We have* $\left[\min_{\mathbf{c} \in \bar{\mathcal{X}}} \mathbf{Gap}(T^2|_{D_\mathbf{c}})\right]^{-1} = \mathcal{O}(L^{2w+3})$.

PROOF. Take any $\mathbf{c} \in \bar{\mathcal{X}}$. For $\mathbf{s} \in \{1,2\}^w$ such that $C(\mathbf{S})_\mathbf{s} > 0$ let $\mathcal{X}_\mathbf{s} \triangleq \{\mathbf{z} \in \{0,1\}^{C(\mathbf{S})_\mathbf{s}} : \sum_i z_i = c_\mathbf{s}\}$. Using (5.4) and (5.8) the subvector of $\mathbf{A} \in D_\mathbf{c}$

defined by $(A_i : \mathbf{S}_{wi-w+1:wi} = \mathbf{s})$ takes values in the space $\mathcal{X}_\mathbf{s}$. So there is some bijective map h such that

$$\{h(\mathbf{A}) : \mathbf{A} \in D_\mathbf{c}\} = \prod_{\mathbf{s}\in\{1,2\}^w : C(\mathbf{S})_\mathbf{s} > 0} \mathcal{X}_\mathbf{s}. \tag{5.14}$$

Define a transition matrix $\tilde{T}$ having state space $\prod_{\mathbf{s}\in\{1,2\}^w:C(\mathbf{S})_\mathbf{s}>0} \mathcal{X}_\mathbf{s}$ and elements $\tilde{T}(h(\mathbf{A}), h(\mathbf{A}')) \triangleq T^2|_{D_\mathbf{c}}(\mathbf{A}, \mathbf{A}')$ for all $\mathbf{A}, \mathbf{A}' \in D_\mathbf{c}$. Then we have

$$\mathbf{Gap}(\tilde{T}) = \mathbf{Gap}(T^2|_{D_\mathbf{c}}). \tag{5.15}$$

Using (B.2) and the fact that T^2 is reversible w.r.t. $\pi(\mathbf{A}|\mathbf{S})$, $T^2|_{D_\mathbf{c}}$ is reversible w.r.t. $\pi|_{D_\mathbf{c}}(\mathbf{A}|\mathbf{S})$. Since $\pi(\mathbf{A}|\mathbf{S})$ is equal for all $\mathbf{A} \in D_\mathbf{c}$, $\pi|_{D_\mathbf{c}}(\mathbf{A}|\mathbf{S})$ is uniform on $D_\mathbf{c}$, and thus $\tilde{T}$ is also reversible w.r.t. the uniform distribution.

We will compare $\tilde{T}$ to a product chain as defined in Theorem B.3, denoted by T^*. Define T^* to have component chains indexed by $\mathbf{s} \in \{1,2\}^w : C(\mathbf{S})_\mathbf{s} > 0$. The component chains are denoted by $T^*_\mathbf{s}$, have state space $\mathcal{X}_\mathbf{s}$, and are combined using weights $b_\mathbf{s} = \frac{C(\mathbf{S})_\mathbf{s}}{L/w}$ to form T^*. Define $T^*_\mathbf{s}$ to be the exclusion process with $c_\mathbf{s}$ particles on the complete graph of $\{1, \ldots, C(\mathbf{S})_\mathbf{s}\}$ (Diaconis and Saloff-Coste, 1993). This chain is defined informally as follows, where $\mathbf{z} \in \mathcal{X}_\mathbf{s}$ is the current state. If $c_\mathbf{s} = 0$ or $c_\mathbf{s} = C(\mathbf{S})_\mathbf{s}$ then $\mathcal{X}_\mathbf{s}$ has a single state and the transition matrix $T^*_\mathbf{s}$ is trivially defined. Otherwise, $T^*_\mathbf{s}$ chooses an index j uniformly at random from $\{i \in \{1, \ldots, C(\mathbf{S})_\mathbf{s}\} : z_i = 1\}$. Then it chooses an index $\ell \in \{1, \ldots, C(\mathbf{S})_\mathbf{s}\}$ uniformly at random. If $z_\ell = 0$ then z_j is changed to 0 and z_ℓ is changed to 1; otherwise, the state $\mathbf{z}$ does not change. The matrix $T^*_\mathbf{s}$ is reversible with respect to the distribution $\mu_\mathbf{s}$ that is uniform on $\mathcal{X}_\mathbf{s}$ (Diaconis and Saloff-Coste, 1993). So by Theorem B.3, T^* is reversible with respect to the uniform distribution on $\prod_{\mathbf{s}\in\{1,2\}^w:C(\mathbf{S})_\mathbf{s}>0} \mathcal{X}_\mathbf{s}$.

By Theorem 3.1 of Diaconis and Saloff-Coste (1993), for $c_\mathbf{s} > 0$ we have $\mathbf{Gap}(T^*_\mathbf{s}) \geq 1/c_\mathbf{s}$, while for $c_\mathbf{s} = 0$, $\mathbf{Gap}(T^*_\mathbf{s}) = 1$. Then Theorem B.3 together with $c_\mathbf{s} \leq C(\mathbf{S})_\mathbf{s}$ yields

$$\begin{aligned} \mathbf{Gap}(T^*) &= \min_{\mathbf{s}\in\{1,2\}^w:C(\mathbf{S})_\mathbf{s}>0} b_\mathbf{s}\mathbf{Gap}(T^*_\mathbf{s}) \\ &\geq \min_{\mathbf{s}\in\{1,2\}^w:C(\mathbf{S})_\mathbf{s}>0} \left\{ \frac{C(\mathbf{S})_\mathbf{s}}{L/w} \left(1 \wedge \frac{1}{c_\mathbf{s}}\right)\right\} \qquad \geq \frac{w}{L}. \end{aligned} \tag{5.16}$$

By (2.6) and the fact that $\mathbf{C}(\mathbf{A}) = \mathbf{C}(\mathbf{A}') = \mathbf{c}$ for $\mathbf{A}, \mathbf{A}' \in D_\mathbf{c}$, $T^2(\mathbf{A}, \mathbf{A}') > 0$ iff $\exists \mathbf{s} \in \{1,2\}^w : C(\mathbf{S})_\mathbf{s} > 0$ such that $\mathbf{A}'$ differs from $\mathbf{A}$ only by swapping two elements of the subvector $(A_i : \mathbf{S}_{wi-w+1:wi} = \mathbf{s})$. Swapping two elements of a vector in $\mathcal{X}_\mathbf{s}$ is precisely the move made by the transition matrix $T^*_\mathbf{s}$. So

$$T^*(h(\mathbf{A}), h(\mathbf{A}')) > 0 \quad \text{iff} \quad T^2(\mathbf{A}, \mathbf{A}') > 0 \qquad \text{for } \mathbf{A}, \mathbf{A}' \in D_\mathbf{c} : \mathbf{A} \neq \mathbf{A}'. \tag{5.17}$$

Using (B.1),

$$T^2|_{D_{\mathbf{c}}}(\mathbf{A},\mathbf{A}') = T^2(\mathbf{A},\mathbf{A}') \qquad \forall \mathbf{A},\mathbf{A}' \in D_{\mathbf{c}} : \mathbf{A} \neq \mathbf{A}'. \tag{5.18}$$

Also, by Lemma 5.1 below,

$$d_1 \triangleq \left[\min_{\mathbf{A},\mathbf{A}' \in \mathcal{X}: T^2(\mathbf{A},\mathbf{A}')>0} T^2(\mathbf{A},\mathbf{A}')\right]^{-1} = \mathcal{O}(L^{2w+2}). \tag{5.19}$$

Combining with (5.17)-(5.18), for any $\mathbf{A},\mathbf{A}' \in D_{\mathbf{c}}$ such that $\mathbf{A} \neq \mathbf{A}'$ and $T^*(h(\mathbf{A}),h(\mathbf{A}')) > 0$,

$$\tilde{T}(h(\mathbf{A}),h(\mathbf{A}')) = T^2(\mathbf{A},\mathbf{A}') \geq d_1^{-1} \geq d_1^{-1}\, T^*(h(\mathbf{A}),h(\mathbf{A}')).$$

For $\mathbf{A},\mathbf{A}' \in D_{\mathbf{c}}$ such that $\mathbf{A} \neq \mathbf{A}'$ and $T^*(h(\mathbf{A}),h(\mathbf{A}')) = 0$ the same inequality holds trivially. So

$$\tilde{T}(h(\mathbf{A}),h(\mathbf{A}')) \geq d_1^{-1} T^*(h(\mathbf{A}),h(\mathbf{A}')) \qquad \forall h(\mathbf{A}) \neq h(\mathbf{A}') \in \prod_{\mathbf{s}\in\{1,2\}^w : C(\mathbf{S})_{\mathbf{s}}>0} \mathcal{X}_{\mathbf{s}}.$$

By Lemma B.2 and the fact that both $\tilde{T}$ and T^* are reversible with respect to the uniform distribution on $\prod_{\mathbf{s}\in\{1,2\}^w : C(\mathbf{S})_{\mathbf{s}}>0} \mathcal{X}_{\mathbf{s}}$, we then have $\mathbf{Gap}(\tilde{T}) \geq d_1^{-1}\mathbf{Gap}(T^*)$. Combining with (5.15) and (5.16),

$$\mathbf{Gap}(T^2|_{D_{\mathbf{c}}})^{-1} = \mathbf{Gap}(\tilde{T})^{-1} \leq d_1 \mathbf{Gap}(T^*)^{-1} \leq \frac{d_1 L}{w}$$

regardless of the value of $\mathbf{c}$. By (5.19) d_1 does not depend on $\mathbf{c}$, so $\max_{\mathbf{c}\in\bar{\mathcal{X}}} \mathbf{Gap}(T^2|_{D_{\mathbf{c}}})^{-1} = \mathcal{O}(L^{2w+3})$. □

Lemma 5.1 was used in the proof of Proposition 5.1.

LEMMA 5.1. *We have*

$$\left[\min_{\mathbf{A},\mathbf{A}' \in \mathcal{X}: T(\mathbf{A},\mathbf{A}')>0} T(\mathbf{A},\mathbf{A}')\right]^{-1} = \mathcal{O}(L^{w+1}).$$

PROOF. Recall the definition of $\check{\theta}_{k,m}$ from (5.2). Using (5.1),

$$\begin{aligned}
&\pi(A_i = 1 | \mathbf{A}_{[-i]}, \mathbf{S}) \\
&= \frac{p_0 \prod_{k=1}^w \check{\theta}_{k,S_{wi-w+k}}}{p_0 \prod_{k=1}^w \check{\theta}_{k,S_{wi-w+k}} + (1-p_0)\frac{\Gamma(\mathbf{N}(\mathbf{A}^c_{[i,0]})+\boldsymbol{\beta}_0)\Gamma(|\mathbf{N}(\mathbf{A}^c_{[i,1]})|+|\boldsymbol{\beta}_0|)}{\Gamma(\mathbf{N}(\mathbf{A}^c_{[i,1]})+\boldsymbol{\beta}_0)\Gamma(|\mathbf{N}(\mathbf{A}^c_{[i,0]})|+|\boldsymbol{\beta}_0|)}} \\
&\geq \min\left\{\frac{1}{2}, \frac{p_0 \prod_{k=1}^w \check{\theta}_{k,S_{wi-w+k}}}{2(1-p_0)\frac{\Gamma(\mathbf{N}(\mathbf{A}^c_{[i,0]})+\boldsymbol{\beta}_0)\Gamma(|\mathbf{N}(\mathbf{A}^c_{[i,1]})|+|\boldsymbol{\beta}_0|)}{\Gamma(\mathbf{N}(\mathbf{A}^c_{[i,1]})+\boldsymbol{\beta}_0)\Gamma(|\mathbf{N}(\mathbf{A}^c_{[i,0]})|+|\boldsymbol{\beta}_0|)}}\right\}. \qquad (5.20)
\end{aligned}$$

Also, by (2.1) and the definitions of $\mathbf{A}_{[i,0]}$ and $\mathbf{A}_{[i,1]}$,

$$N(\mathbf{A}^c_{[i,0]})_m = N(\mathbf{A}^c_{[i,1]})_m + \sum_{k=1}^{w} \mathbf{1}_{\{\mathbf{S}_{wi-w+k}=m\}} \qquad m \in \{1,2\}$$
$$|\mathbf{N}(\mathbf{A}^c_{[i,0]})| = |\mathbf{N}(\mathbf{A}^c_{[i,1]})| + w.$$

So

$$\begin{aligned}
&\frac{\Gamma(\mathbf{N}(\mathbf{A}^c_{[i,0]}) + \boldsymbol{\beta}_0)\Gamma(|\mathbf{N}(\mathbf{A}^c_{[i,1]})| + |\boldsymbol{\beta}_0|)}{\Gamma(\mathbf{N}(\mathbf{A}^c_{[i,1]}) + \boldsymbol{\beta}_0)\Gamma(|\mathbf{N}(\mathbf{A}^c_{[i,0]})| + |\boldsymbol{\beta}_0|)} \\
&= \frac{\prod_{m=1}^{2} \left(N(\mathbf{A}^c_{[i,1]})_m + \beta_{0,m}\right)\left(N(\mathbf{A}^c_{[i,1]})_m + \beta_{0,m} + 1\right) \ldots \left(N(\mathbf{A}^c_{[i,1]})_m + \beta_{0,m} + \sum_{k=1}^{w} \mathbf{1}_{\{S_{wi-w+k}=m\}} - 1\right)}{(|\mathbf{N}(\mathbf{A}^c_{[i,1]})| + |\boldsymbol{\beta}_0|)(|\mathbf{N}(\mathbf{A}^c_{[i,1]})| + |\boldsymbol{\beta}_0| + 1) \ldots (|\mathbf{N}(\mathbf{A}^c_{[i,1]})| + |\boldsymbol{\beta}_0| + w - 1)} \\
&\leq \frac{\prod_{m=1}^{2} \left(|\mathbf{N}(\mathbf{A}^c_{[i,1]})| + |\boldsymbol{\beta}_0|\right)\left(|\mathbf{N}(\mathbf{A}^c_{[i,1]})| + |\boldsymbol{\beta}_0| + 1\right) \ldots \left(|\mathbf{N}(\mathbf{A}^c_{[i,1]})| + |\boldsymbol{\beta}_0| + \sum_{k=1}^{w} \mathbf{1}_{\{S_{wi-w+k}=m\}} - 1\right)}{(|\mathbf{N}(\mathbf{A}^c_{[i,1]})| + |\boldsymbol{\beta}_0|)(|\mathbf{N}(\mathbf{A}^c_{[i,1]})| + |\boldsymbol{\beta}_0| + 1) \ldots (|\mathbf{N}(\mathbf{A}^c_{[i,1]})| + |\boldsymbol{\beta}_0| + w - 1)} \\
&\leq 1.
\end{aligned}$$

Combining with (5.2) and (5.20),

$$\pi(A_i = 1 | \mathbf{A}_{[-i]}, \mathbf{S}) \geq \frac{p_0 \prod_{k=1}^{w} \check{\theta}_{k,S_{wi-w+k}}}{2} \qquad \geq \frac{p_0}{2} \prod_{k=1}^{w} \frac{\beta_{k,S_{wi-w+k}}}{L + |\boldsymbol{\beta}_k|}$$

which does not depend on $\mathbf{A}$ or i. So $\left[\min_{\mathbf{A},i} \pi(A_i = 1 | \mathbf{A}_{[-i]}, \mathbf{S})\right]^{-1} = \mathcal{O}(L^w)$. Analogously, $\left[\min_{\mathbf{A},i} \pi(A_i = 0 | \mathbf{A}_{[-i]}, \mathbf{S})\right]^{-1} = \mathcal{O}(L^w)$. Using (2.6) then yields the desired result. □

Proposition 5.2 was used in the proof of Theorem 5.1.

PROPOSITION 5.2. $\mathbf{Gap}(\bar{T})$ *is within a polynomial (in* L*) factor of* d*. Specifically,* $\mathbf{Gap}(\bar{T}) = \mathcal{O}(d \times L^{2^w})$ *and* $\mathbf{Gap}(\bar{T})^{-1} = \mathcal{O}\left(d^{-1} \times L^{w+1+2^{w+1}+2^w}\right)$.

The bounds in Proposition 5.2 only rely on Lemma 5.1 and the fact (5.10) that $|\bar{\mathcal{X}}|$ grows polynomially in L, and could be improved by leveraging additional properties of $\bar{T}$ (at the cost of some technical complexity). However, Proposition 5.2 is sufficient for our purposes.

Proof of Prop. 5.2, Upper Bound. Suppress the dependence of $\bar{\pi}(\mathbf{c}|\mathbf{S})$ on $\mathbf{S}$ for simplicity of notation. Let $\mathbf{c}_1, \mathbf{c}_2 \in \bar{\mathcal{X}}$ be a pair of states that achieve the minimum in the definition (5.11) of d, so that $d = \max_{\gamma \in \Gamma_{\mathbf{c}_1,\mathbf{c}_2}} \min_{\mathbf{c} \in \gamma} \frac{\bar{\pi}(\mathbf{c})}{\bar{\pi}(\mathbf{c}_1)\bar{\pi}(\mathbf{c}_2)}$.

For any $\gamma \in \Gamma_{\mathbf{c}_1,\mathbf{c}_2}$, let $\mathbf{c}_\gamma \triangleq \text{argmin}_{\mathbf{c}\in\gamma} \frac{\bar{\pi}(\mathbf{c})}{\bar{\pi}(\mathbf{c}_1)\bar{\pi}(\mathbf{c}_2)}$ (in the case of a tie choose the state earliest in the path). Defining the set $E = \{\mathbf{c}_\gamma : \gamma \in \Gamma_{\mathbf{c}_1,\mathbf{c}_2}\}$, we have

$$d = \max_{\mathbf{c}_\gamma \in E} \frac{\bar{\pi}(\mathbf{c}_\gamma)}{\bar{\pi}(\mathbf{c}_1)\bar{\pi}(\mathbf{c}_2)}. \tag{5.21}$$

The set E separates $\mathbf{c}_1$ and $\mathbf{c}_2$ in the sense that there is no path $\gamma \in \Gamma_{\mathbf{c}_1,\mathbf{c}_2}$ that does not include some state in E. If $\mathbf{c}_1 \in E$ then there is some $\gamma \in \Gamma_{\mathbf{c}_1,\mathbf{c}_2}$ for which $\mathbf{c}_\gamma = \mathbf{c}_1$, and so $d \geq \frac{1}{\bar{\pi}(\mathbf{c}_2)} \geq 1$. In this case $\mathbf{Gap}(\bar{T}) \leq 2d(L/w+1)^{2^w}$ holds since $\mathbf{Gap}(\bar{T}) \leq 2$.

Now consider the case $\mathbf{c}_1 \notin E$. Let B be the set of states $\mathbf{c} \in \bar{\mathcal{X}}$ that are not reachable from $\mathbf{c}_1$ without going through E:

$$B \triangleq \{\mathbf{c} \in \bar{\mathcal{X}} : \forall \gamma \in \Gamma_{\mathbf{c}_1,\mathbf{c}} \text{ there is some } \mathbf{c}' \in \gamma \text{ s.t. } \mathbf{c}' \in E\}.$$

We have that $\mathbf{c}_2 \in B$, and $\mathbf{c}_1 \in B^c$ since $\mathbf{c}_1 \notin E$. Also, the only states $\mathbf{c} \in B$ for which $\bar{T}(\mathbf{c}, B^c) > 0$ satisfy $\mathbf{c} \in E$, seen as follows. Otherwise, $\exists \mathbf{c} \in B \backslash E$ and $\mathbf{c}_3 \in B^c$ for which $\bar{T}(\mathbf{c}, \mathbf{c}_3) > 0$. Since $\mathbf{c}_3 \in B^c$, there is a path $\gamma \in \Gamma_{\mathbf{c}_1,\mathbf{c}_3}$ that does not go through E. But since $\bar{T}(\mathbf{c}_3, \mathbf{c}) > 0$ ($\bar{T}$ is reversible), there is also a path $\gamma \in \Gamma_{\mathbf{c}_1,\mathbf{c}}$ that does not go through E, which is a contradiction.

Using these facts, (5.10), and (5.21), the conductance of B (Theorem B.2) is:

$$\begin{aligned}
\Phi_{\bar{T}}(B) &= \frac{\sum_{\mathbf{c}\in B} \bar{\pi}(\mathbf{c})\bar{T}(\mathbf{c}, B^c)}{\bar{\pi}(B)\bar{\pi}(B^c)} \leq \frac{\sum_{\mathbf{c}\in E} \bar{\pi}(\mathbf{c})\bar{T}(\mathbf{c}, B^c)}{\bar{\pi}(B)\bar{\pi}(B^c)} \\
&\leq \frac{\sum_{\mathbf{c}\in E} \bar{\pi}(\mathbf{c})}{\bar{\pi}(B)\bar{\pi}(B^c)} \leq \frac{\sum_{\mathbf{c}\in E} \bar{\pi}(\mathbf{c})}{\bar{\pi}(\mathbf{c}_1)\bar{\pi}(\mathbf{c}_2)} \\
&\leq |E| \max_{\mathbf{c}\in E} \frac{\bar{\pi}(\mathbf{c})}{\bar{\pi}(\mathbf{c}_1)\bar{\pi}(\mathbf{c}_2)} = |E|d \leq |\bar{\mathcal{X}}|d \leq d(L/w+1)^{2^w}.
\end{aligned}$$

Using Theorem B.2, $\mathbf{Gap}(\bar{T}) \leq 2\Phi_{\bar{T}}(B) \leq 2d(L/w+1)^{2^w}$ as claimed. □

Proof of Prop. 5.2, Lower Bound. Recall that the transition matrix $\bar{T}$ has state space $\bar{\mathcal{X}}$ and is reversible with respect to $\bar{\pi}$. Using (5.8)-(5.9), for $\mathbf{c}, \mathbf{c}' \in \bar{\mathcal{X}}$ such that $\sum_{\mathbf{s}} |c_{\mathbf{s}} - c'_{\mathbf{s}}| \leq 1$, $\bar{T}(\mathbf{c}, \mathbf{c}') > 0$ and otherwise $\bar{T}(\mathbf{c}, \mathbf{c}') = 0$. Also

$$\bar{T}(\mathbf{c}, \mathbf{c}') \geq \min_{\mathbf{A}\in D_{\mathbf{c}}} T(\mathbf{A}, D_{\mathbf{c}'}) \geq \min_{\mathbf{A}\in D_{\mathbf{c}}} \max_{\mathbf{A}'\in D_{\mathbf{c}'}} T(\mathbf{A}, \mathbf{A}') \qquad \forall \mathbf{c}, \mathbf{c}' \in \bar{\mathcal{X}}. \tag{5.22}$$

If $\bar{T}(\mathbf{c}, \mathbf{c}') > 0$ then for every $\mathbf{A} \in D_{\mathbf{c}}$ there is some $\mathbf{A}' \in D_{\mathbf{c}'}$ with $T(\mathbf{A}, \mathbf{A}') > 0$. By Lemma 5.1 and (5.22),

$$\left[\min_{\mathbf{c},\mathbf{c}'\in\bar{\mathcal{X}}:\bar{T}(\mathbf{c},\mathbf{c}')>0} \bar{T}(\mathbf{c}, \mathbf{c}')\right]^{-1} = \mathcal{O}(L^{w+1}). \tag{5.23}$$

We will use Theorem B.4 to obtain a bound for $\mathbf{Gap}(\bar{T})$; let $\mathcal{E}$ be the set of edges in the graph of $\bar{T}$. For $(z,v) \in \mathcal{E}$ and γ a path in the graph, let $(z,v) \in \gamma$ indicate that the edge (z,v) is in the path γ (as distinct from $z \in \gamma$ which indicates that the vertex z is in γ). To apply Theorem B.4 we need to define a path $\gamma_{x,y}$ for every pair of states $x, y \in \bar{\mathcal{X}}$. Suppressing the dependence of $\bar{\pi}$ on $\mathbf{S}$, choose $\gamma_{x,y}$ to be any path that maximizes $\min_{(z,v)\in\gamma} \frac{\bar{\pi}(z)}{\bar{\pi}(x)\bar{\pi}(y)}$. Letting $|\{\ldots\}|$ denote the cardinality of a set, the path constant ρ defined in Theorem B.4 satisfies

$$
\begin{aligned}
\rho &= \max_{(z,v)\in\mathcal{E}} \frac{1}{\bar{\pi}(z)\bar{T}(z,v)} \sum_{\gamma_{x,y} \ni (z,v)} \bar{\pi}(x)\bar{\pi}(y)\,\mathrm{len}(\gamma_{x,y}) \\
&\le \frac{1}{\min_{(z,v)\in\mathcal{E}} \bar{T}(z,v)} \left[\max_{(z,v)\in\mathcal{E}} \frac{1}{\bar{\pi}(z)} \sum_{\gamma_{x,y} \ni (z,v)} \bar{\pi}(x)\bar{\pi}(y) \right] \max_{x,y} \mathrm{len}(\gamma_{x,y}) \\
&\le \frac{1}{\min_{(z,v)\in\mathcal{E}} \bar{T}(z,v)} \left[\max_{(z,v)\in\mathcal{E}} |\{\gamma_{x,y} \ni (z,v)\}| \right] \left[\max_{(z,v)\in\mathcal{E}} \max_{\gamma_{x,y} \ni (z,v)} \frac{\bar{\pi}(x)\bar{\pi}(y)}{\bar{\pi}(z)} \right] \max_{x,y} \mathrm{len}(\gamma_{x,y}) \\
&= \frac{1}{\min_{(z,v)\in\mathcal{E}} \bar{T}(z,v)} \left[\max_{(z,v)\in\mathcal{E}} |\{\gamma_{x,y} \ni (z,v)\}| \right] \left[\min_{x,y\in\bar{\mathcal{X}}} \min_{(z,v)\in\gamma_{x,y}} \frac{\bar{\pi}(z)}{\bar{\pi}(x)\bar{\pi}(y)} \right]^{-1} \max_{x,y} \mathrm{len}(\gamma_{x,y}) \\
&= \frac{1}{\min_{(z,v)\in\mathcal{E}} \bar{T}(z,v)} \left[\max_{(z,v)\in\mathcal{E}} |\{\gamma_{x,y} \ni (z,v)\}| \right] \left[\min_{x,y\in\bar{\mathcal{X}}} \max_{\gamma\in\Gamma_{x,y}} \min_{(z,v)\in\gamma} \frac{\bar{\pi}(z)}{\bar{\pi}(x)\bar{\pi}(y)} \right]^{-1} \max_{x,y} \mathrm{len}(\gamma_{x,y}) \\
&\le \frac{1}{\min_{(z,v)\in\mathcal{E}} \bar{T}(z,v)} \left[\max_{(z,v)\in\mathcal{E}} |\{\gamma_{x,y} \ni (z,v)\}| \right] \left[\min_{x,y\in\bar{\mathcal{X}}} \max_{\gamma\in\Gamma_{x,y}} \min_{z\in\gamma} \frac{\bar{\pi}(z)}{\bar{\pi}(x)\bar{\pi}(y)} \right]^{-1} \max_{x,y} \mathrm{len}(\gamma_{x,y}).
\end{aligned}
$$

¿From (5.10) $|\bar{\mathcal{X}}| = \mathcal{O}(L^{2^w})$, so the maximum length of paths is $\max_{x,y} \mathrm{len}(\gamma_{x,y}) = \mathcal{O}(L^{2^w})$ and the total number of paths is no more than $|\bar{\mathcal{X}}|^2 = \mathcal{O}(L^{2^{w+1}})$. By Theorem B.4 and (5.23), $\mathbf{Gap}(\bar{T})^{-1} \le \rho = \mathcal{O}\left(d^{-1}\, L^{w+1+2^{w+1}+2^w}\right)$. □

5.4. *Step 2 of Proof of Thm. 3.1.* Recalling that $M = 2$, there are $w+1$ free parameters $\theta_{k,1} \in [0,1]$ for $k \in \{0,\ldots,w\}$, so we write

$$
\boldsymbol{\theta}_{0:w} \in [0,1]^{w+1}. \tag{5.24}
$$

Theorem 5.2 formalizes Step 2, using Theorem A.1.

THEOREM 5.2. *Under Assumption 3.1, if $E \log f(\mathbf{s}|\boldsymbol{\theta}_{0:w})$ is multimodal then there exist $\epsilon > 0$ and two sets $B_1, B_2 \subset [0,1]^{w+1}$ separated by Euclidean distance ϵ such that*

$$
\frac{\pi\left(\boldsymbol{\theta}_{0:w} \notin B_1 \cup B_2 | \mathbf{S}\right)}{\pi\left(\boldsymbol{\theta}_{0:w} \in B_1 | \mathbf{S}\right)} \qquad \textit{and} \qquad \frac{\pi\left(\boldsymbol{\theta}_{0:w} \notin B_1 \cup B_2 | \mathbf{S}\right)}{\pi\left(\boldsymbol{\theta}_{0:w} \in B_2 | \mathbf{S}\right)} \tag{5.25}
$$

decrease exponentially in L, almost surely.

PROOF. The inference model assumes $\mathbf{S}_{wi-w+1:wi} \overset{\text{iid}}{\sim} f(\mathbf{s}|\boldsymbol{\theta}_{0:w})$ for $i \in \{1, \ldots, L/w\}$. By Assumption 3.1, the generative model assumes $\mathbf{S}_{wi-w+1:wi} \overset{\text{iid}}{\sim} g(\mathbf{s})$, fitting into the framework of Theorem A.1. Using the notation of that Theorem, consider the case where $\eta(\boldsymbol{\theta}_{0:w}) = E \log f(\mathbf{s}|\boldsymbol{\theta}_{0:w})$ is multimodal. Then there are $\boldsymbol{\theta}^{(j)}_{0:w} \in [0,1]^{w+1}$ and $F_j \subset [0,1]^{w+1}$ for $j \in \{1,2\}$ such that $\boldsymbol{\theta}^{(j)}_{0:w} \in F_j$ and (3.1) holds. Let $\phi_j \triangleq \sup_{\partial F_j} \eta(\boldsymbol{\theta}_{0:w}) < \eta(\boldsymbol{\theta}^{(j)}_{0:w})$ for $j \in \{1,2\}$, and take any ξ for which

$$\xi \in \left(\min_{j \in \{1,2\}} \phi_j, \ \min_{j \in \{1,2\}} \eta(\boldsymbol{\theta}^{(j)}_{0:w}) \right). \tag{5.26}$$

So $\xi > \phi_j$ for some $j \in \{1,2\}$; assume WLOG that $\xi > \phi_1$.

Define the sets

$$\begin{aligned} V &\triangleq \{\boldsymbol{\theta}_{0:w} \in [0,1]^{w+1} : \eta(\boldsymbol{\theta}_{0:w}) \geq \xi\} \\ \tilde{B}_1 &\triangleq F_1 \cap V \qquad \tilde{B}_2 \triangleq V \backslash F_1. \end{aligned} \tag{5.27}$$

By (5.26), $\eta(\boldsymbol{\theta}^{(1)}_{0:w}) > \xi$ and $\boldsymbol{\theta}^{(1)}_{0:w} \in \tilde{B}_1$. Also using (3.1), $\eta(\boldsymbol{\theta}^{(2)}_{0:w}) > \xi$ and $\boldsymbol{\theta}^{(2)}_{0:w} \in F_2 \cap V \subset \tilde{B}_2$, so

$$V = \tilde{B}_1 \cup \tilde{B}_2 \qquad \text{and} \qquad \sup_{\boldsymbol{\theta}_{0:w} \in \tilde{B}_j} \eta(\boldsymbol{\theta}_{0:w}) > \xi \qquad j \in \{1,2\}. \tag{5.28}$$

If $\nexists \epsilon > 0$ such that $\tilde{B}_1$ and $\tilde{B}_2$ are separated by distance ϵ then (since $[0,1]^{w+1}$ is compact) $\text{cl}(\tilde{B}_1) \cap \text{cl}(\tilde{B}_2) \neq \emptyset$. By (5.27) $\tilde{B}_1 \subset F_1$ and $\tilde{B}_2 \subset [0,1]^{w+1} \backslash F_1$ so $\text{cl}(\tilde{B}_1) \cap \text{cl}(\tilde{B}_2) \subset \partial F_1$. This is a contradiction since, due to (5.27)-(5.28) and the continuity of η,

$$\inf_{\text{cl}(\tilde{B}_1) \cap \text{cl}(\tilde{B}_2)} \eta \geq \inf_{\text{cl}(\tilde{B}_1) \cup \text{cl}(\tilde{B}_2)} \eta = \inf_{\tilde{B}_1 \cup \tilde{B}_2} \eta \geq \xi > \phi_1 = \sup_{\partial F_1} \eta.$$

So $\exists \epsilon > 0$ such that $\tilde{B}_1$ and $\tilde{B}_2$ are separated by distance ϵ.

In order to satisfy Assumption 2 of Theorem A.1 we remove points from the space $[0,1]^{w+1}$ of $\boldsymbol{\theta}_{0:w}$ for which $\exists \mathbf{s} \in \{1,2\}^w : f(\mathbf{s}|\boldsymbol{\theta}_{0:w}) = 0$. This results in the space

$$\Lambda = ((0,1) \times [0,1]^w) \cup ([0,1] \times (0,1)^w). \tag{5.29}$$

The fact that $f(\mathbf{s}|\boldsymbol{\theta}_{0:w}) > 0$ for all $\mathbf{s}$ and all $\boldsymbol{\theta}_{0:w} \in \Lambda$ is a consequence of (2.8): if $\theta_{0,1} \in (0,1)$ then $f(\mathbf{s}|\boldsymbol{\theta}_{0:w}) > 0$ for all $\mathbf{s}$, and the same holds when $\theta_{k,1} \in (0,1)$ for all $k \in \{1, \ldots, w\}$.

Taking

$$B_j \triangleq \tilde{B}_j \cap \Lambda \qquad j \in \{1,2\} \tag{5.30}$$

B_1 and B_2 are separated by distance ϵ. Due to (3.1), $\boldsymbol{\theta}_{0:w}^{(j)}$ is not a limit point of $[0,1]^{w+1}\backslash F_j$ for $j \in \{1,2\}$. Using (5.29) there are points $\boldsymbol{\theta}_{0:w} \in \Lambda$ arbitrarily close to $\boldsymbol{\theta}_{0:w}^{(j)} \in [0,1]^{w+1}$. From (5.26) and the continuity of η, all such points $\boldsymbol{\theta}_{0:w}$ close enough to $\boldsymbol{\theta}_{0:w}^{(j)}$ have $\boldsymbol{\theta}_{0:w} \in F_j \cap \Lambda$ and $\eta(\boldsymbol{\theta}_{0:w}) > \xi$. Using (3.1), (5.27), and (5.30) these points are in B_j, so

$$\sup_{\boldsymbol{\theta}_{0:w} \in B_j} \eta(\boldsymbol{\theta}_{0:w}) > \xi \qquad j \in \{1,2\}. \tag{5.31}$$

Let $\mathrm{Int}(\cdot)$ denote set interior with respect to the space Λ. We claim that

$$\{\boldsymbol{\theta}_{0:w} \in \Lambda : \eta(\boldsymbol{\theta}_{0:w}) > \xi\} \;\subset\; \mathrm{Int}(B_1) \cup \mathrm{Int}(B_2) \tag{5.32}$$

which can be seen as follows. Take any $\boldsymbol{\theta}_{0:w} \in \Lambda$ such that $\eta(\boldsymbol{\theta}_{0:w}) > \xi$. By (5.27), (5.28), (5.30) and since η is continuous, $\boldsymbol{\theta}_{0:w} \in \mathrm{Int}(B_1 \cup B_2)$. Because B_1 and B_2 are separated by distance $\epsilon > 0$, $\boldsymbol{\theta}_{0:w} \in \mathrm{Int}(B_1) \cup \mathrm{Int}(B_2)$.

Define the alternative parameter spaces $\Lambda_1 \triangleq \Lambda\backslash\mathrm{Int}(B_2)$ and $\Lambda_2 \triangleq \Lambda\backslash\mathrm{Int}(B_1)$. By (5.31), $\sup_{\Lambda_j} \eta > \xi$ for $j \in \{1,2\}$. So $\delta_j \triangleq \frac{\sup_{\Lambda_j} \eta - \xi}{2} > 0$ for $j \in \{1,2\}$. Combining (5.32) with the fact that $\sup_{\Lambda_j} \eta - \delta_j > \xi$,

$$U_{\delta_j}^j \triangleq \left\{ \boldsymbol{\theta}_{0:w} \in \Lambda_j : \eta(\boldsymbol{\theta}_{0:w}) \geq \sup_{\Lambda_j} \eta - \delta_j \right\} \qquad \subset \mathrm{Int}(B_j) \subset B_j \qquad j \in \{1,2\}. \tag{5.33}$$

By (5.33),

$$\Lambda\backslash(B_1 \cup B_2) \;\subset\; \Lambda\backslash(B_1 \cup \mathrm{Int}(B_2)) \;\subset\; \Lambda\backslash(U_{\delta_1}^1 \cup \mathrm{Int}(B_2)) \;=\; \Lambda_1\backslash U_{\delta_1}^1. \tag{5.34}$$

Analogously, $\Lambda\backslash(B_1 \cup B_2) \subset \Lambda_2\backslash U_{\delta_2}^2$.

The regularity conditions of Theorem A.1 are verified in the Web Appendix for each of the parameter spaces Λ_1 and Λ_2. We apply that Theorem for each of $j \in \{1,2\}$, with parameter space Λ_j, using $\delta = \delta_j$ and taking $n = L/w$. This yields

$$\limsup_{n\to\infty} \left(\frac{P_n(\Lambda_j\backslash U_{\delta_j}^j)}{P_n(U_{\delta_j}^j)} \right)^{1/n} \leq e^{-\delta_j} \qquad \text{a.s.} \qquad j \in \{1,2\}.$$

Combining with (5.33)-(5.34) and the fact that $[0,1]^{w+1}\backslash\Lambda$ has probability zero under $\pi(\boldsymbol{\theta}_{0:w}|\mathbf{S})$, for $j \in \{1,2\}$

$$\limsup_{L\to\infty}\left(\frac{\pi\big(\boldsymbol{\theta}_{0:w} \in [0,1]^{w+1}\backslash(B_1\cup B_2) \,\big|\, \mathbf{S}\big)}{\pi(\boldsymbol{\theta}_{0:w} \in B_j \,|\, \mathbf{S})}\right)^{\frac{1}{(L/w)}}$$

$$= \limsup_{L\to\infty}\left(\frac{\pi\big(\boldsymbol{\theta}_{0:w} \in \Lambda\backslash(B_1\cup B_2) \,\big|\, \mathbf{S}\big)}{\pi(\boldsymbol{\theta}_{0:w} \in B_j \,|\, \mathbf{S})}\right)^{\frac{1}{(L/w)}}$$

$$\leq \limsup_{L\to\infty}\left(\frac{\pi(\boldsymbol{\theta}_{0:w} \in \Lambda_j\backslash U^j_{\delta_j} \,|\, \mathbf{S})}{\pi(\boldsymbol{\theta}_{0:w} \in U^j_{\delta_j} \,|\, \mathbf{S})}\right)^{\frac{1}{(L/w)}} = \limsup_{n\to\infty}\left(\frac{P_n(\Lambda_j\backslash U^j_{\delta_j})}{P_n(U^j_{\delta_j})}\right)^{1/n} \leq e^{-\delta_j}$$

almost surely. □

5.5. *Step 3 of Proof of Thm. 3.1.* Finally we formalize Step 3.

THEOREM 5.3. *If there exist $\epsilon > 0$ and two sets $B_1, B_2 \subset [0,1]^{w+1}$ separated by Euclidean distance ϵ such that the ratios in (5.25) decrease exponentially in L, then the quantity d in (5.11) decreases exponentially in L.*

Theorem 5.3 is proven in the Web Appendix. Theorems 5.1-5.3 together imply Theorem 3.1. □

6. Conclusions. The Gibbs sampling method is a popular approach to finding gene regulatory binding motifs, but its poor convergence in practice means that it can only be used to generate candidate motifs that must be ranked using a secondary criterion. If one could efficiently obtain samples from the posterior distribution, these samples could be used to directly find the "best," i.e. most probable, motifs, obviating the need for secondary analysis. We have obtained theoretical and empirical results showing that the convergence of the Gibbs sampler is even worse than previously realized. Our results reinforce the need to convey the limitations of any estimates obtained using the Gibbs sampler, and the need to develop more efficient Markov chain methods for motif discovery.

Although our main result (Theorem 3.1) is phrased in terms of a specific model, the methods used to prove this result are very widely applicable to situations with i.i.d. data, where the data are not necessarily generated according to the model, and where the function $E \log f(X|\theta)$ is multimodal. The extent to which slow mixing holds in other contexts will be determined by how generally this multimodality condition holds, so we are currently investigating this condition in detail.

APPENDIX A: BAYESIAN ASYMPTOTICS

We quote a result from Berk (1966) on Bayesian asymptotics for i.i.d. observations. Let $f(x|\theta)$ be the density (with respect to some σ-finite measure on a space $\mathcal{Y}$) of each observation X_i under the inference model, parameterized by $\theta \in \Lambda$ where Λ is a Borel subset of a complete separable metric space. Let the true distribution of the observations be denoted by G. Define:

1. The "carrier" of a distribution P: the smallest relatively closed set having probability one under P.
2. P_n: the posterior distribution of θ with n observations, with respect to a prior P having carrier Λ.
3. $\eta(\theta) \triangleq E \log f(X|\theta)$ where the expectation is taken with respect to $X \sim G$.
4. $\eta^* \triangleq \sup\{\eta(\theta) : \theta \in \Lambda\}$.
5. $U_\delta \triangleq \{\theta \in \Lambda : \eta(\theta) \geq \eta^* - \delta\}$ for $\delta \geq 0$.

Assume that

1. $f(x|\theta)$ is measurable jointly in x and θ; for G-almost every x, $f(x|\theta)$ is continuous in θ.
2. For all $\theta \in \Lambda$, $G\{x : f(x|\theta) > 0\} = 1$.
3. For any compact $F \subset \Lambda$, $E \sup_{\theta \in F} |\log f(X|\theta)| < \infty$.
4. $\eta(\theta)$ is continuous.
5. For any real number r there is a co-compact set $D \subset \Lambda$ ($D^c = \Lambda \backslash D$ is compact) and a cover $D_1, \ldots, D_K$ of D such that

$$E \sup_{\theta \in D_k} \log f(X|\theta) \leq r \qquad k \in \{1, \ldots, K\}. \tag{A.1}$$

With these assumptions, we have Theorem A.1.

THEOREM A.1. *(Berk, 1966) For G-almost every sequence of observations $\{x_i : i \in \mathbb{N}\}$ and any $\delta > 0$,*

$$\limsup_{n \to \infty} \left(\frac{P_n(\Lambda \backslash U_\delta)}{P_n(U_\delta)} \right)^{1/n} \leq e^{-\delta}.$$

Theorem A.1 is a sub-result given in the proof of Berk's main theorem. It says that the posterior probability of U_δ^c decreases exponentially in n. Here we have stated the result slightly more generally than Berk (1966) in the sense that we replace his Assumption iii with the only two relevant consequences of that Assumption: our Assumptions 3-4. Also, in our Assumption 5 we have allowed a cover of D, whereas Berk's Assumption iv takes $K = 1$. The extension to the case of general K is immediate from his proof.

APPENDIX B: TOOLS FOR BOUNDING SPECTRAL GAPS

Let P and Q be transition kernels that are reversible with respect to distributions μ_P and μ_Q on a (general) state space $\mathcal{X}$ with countably-generated σ-algebra. Let $P|_B$ for $B \subset \mathcal{X}$ indicate the restriction of P to B, which is defined to have state space B and transition probabilities identical to P except that any move to B^c is rejected:

$$P|_B(x, D) = P(x, D) + \mathbf{1}_{\{x \in D\}} P(x, B^c) \qquad x \in B, D \subset B. \tag{B.1}$$

Also let $\mu_P|_B$ be the restriction of μ_P to B, i.e.

$$\mu_P|_B(dx) \triangleq \mu_P(dx)/\mu_P(B) \qquad x \in B. \tag{B.2}$$

Then $P|_B$ is reversible w.r.t. $\mu_P|_B$.

For a partition $\{B_j\}_{j=1}^J$ of $\mathcal{X}$, let $\bar{P}$ be the *projection matrix* of P with respect to $\{B_j\}_{j=1}^J$, defined to have state space $\{1, \ldots, J\}$ and i, j element equal to the probability that P transitions to B_j, given that the current state is in B_i. I.e.,

$$\bar{P}(i, j) \triangleq \int \mu_P|_{B_i}(dx) P(x, B_j) \qquad i, j \in \{1, \ldots, J\}.$$

The matrix $\bar{P}$ is reversible w.r.t. $\bar{\mu}$, where $\bar{\mu}(j) \triangleq \mu_P(B_j)$.

LEMMA B.1. *(Madras and Zheng 2003) For any $N \in \mathbb{N}$ we have* $\mathbf{Gap}(P) \geq \frac{1}{N}\mathbf{Gap}(P^N)$.

Although Madras and Zheng (2003) state this result for finite state spaces, their proof also holds for general state spaces.

THEOREM B.1. *(Madras and Randall 2002) Let $\mu_P = \mu_Q$ and let $\{B_j\}_{j=1}^J$ be any partition of $\mathcal{X}$. Assume that P is nonnegative definite and let $P^{1/2}$ be its nonnegative square root. Then*

$$\mathbf{Gap}(P^{1/2} Q P^{1/2}) \geq \mathbf{Gap}(\bar{P}) \min_j \mathbf{Gap}(Q|_{B_j})$$
$$\mathbf{Gap}(P) \leq \mathbf{Gap}(\bar{P})$$

where $\bar{P}$ is the projection matrix of P with respect to $\{B_j\}_{j=1}^J$.

THEOREM B.2. *(E.g. Sinclair 1992) For $\mathcal{X}$ finite define*

$$\Phi_P \triangleq \min_{B \subset \mathcal{X} : 0 < \mu_P(B) < 1} \Phi_P(B) \qquad \Phi_P(B) \triangleq \frac{\sum_{x \in B} \mu_P(x) P(x, B^c)}{\mu_P(B)\mu_P(B^c)}.$$

Here Φ_P is called the "conductance", and $\Phi_P(B)$ is referred to as the conductance of the set B. Then $\mathbf{Gap}(P) \leq 2\Phi_P$.

THEOREM B.3. *(Diaconis and Saloff-Coste 1996) Take any $N \in \mathbb{N}$ and let P_k, $k = 0, \ldots, N$, be μ_k-reversible transition kernels on state spaces $\mathcal{X}_k$. Let P be the transition kernel with state $\mathbf{x} = (x_0, \ldots, x_N)$ in the space $\mathcal{X} = \prod_k \mathcal{X}_k$, given by*

$$P(\mathbf{x}, d\mathbf{y}) = \sum_{k=0}^{N} b_k P_k(x_k, dy_k)\, \delta_{\mathbf{x}_{[-k]}}(\mathbf{y}_{[-k]}) d\mathbf{y}_{[-k]} \qquad \mathbf{x}, \mathbf{y} \in \mathcal{X}$$

for some set of $b_k > 0$ such that $\sum_k b_k = 1$, where δ is Dirac's delta function, and where $\mathbf{x}_{[-k]}$ indicates the vector $\mathbf{x}$ excluding x_k. P is called a product chain with "component" chains P_k. It is reversible with respect to $\mu_P(dx) = \prod_k \mu_k(dx_k)$, and

$$\mathbf{Gap}(P) = \min_{k=0,\ldots,N} b_k \mathbf{Gap}(P_k).$$

Lemma 3.2 of Diaconis and Saloff-Coste (1996) states Theorem B.3 for finite state spaces; however, the proof holds in the general case.

LEMMA B.2. *Take finite $\mathcal{X}$ and $\mu_P = \mu_Q$. If $\exists b > 0$ such that $bQ(x, y) \leq P(x, y)$ for every $x, y \in \mathcal{X}$ such that $x \neq y$, then $b\mathbf{Gap}(Q) \leq \mathbf{Gap}(P)$.*

PROOF. The proof is nearly identical to that of Lemma 5.1 in Woodard, Schmidler and Huber (2009a). □

Lemma B.2 is closely related to Peskun ordering results; cf. Peskun (1973); Tierney (1998); Mira (2001).

THEOREM B.4. *(Sinclair, 1992; Diaconis and Stroock, 1991) For $\mathcal{X}$ finite, define a simple path $\gamma_{x,y}$ between every ordered pair $x, y \in \mathcal{X}$ in the graph of the Markov chain with transition matrix P. A simple path is a sequence of connected edges with no repeated vertices. Define the quantity*

$$\rho \triangleq \max_{(z,v) \in \mathcal{E}} \frac{1}{\mu_P(z) P(z, v)} \sum_{\gamma_{x,y} \ni (z,v)} \mu_P(x) \mu_P(y) len(\gamma_{x,y})$$

where $\mathcal{E}$ is the set of edges, where $\gamma_{x,y} \ni (z, v)$ is a path using the edge (z, v), and where $len(\gamma_{x,y})$ is the number of edges in $\gamma_{x,y}$. Then $\mathbf{Gap}(P) \geq \rho^{-1}$.

REFERENCES

Andrieu, C., Doucet, A. and Holenstein, R. (2010). Particle Markov chain Monte Carlo (with discussion). *Journal of the Royal Statistical Society Series B* **72** 269–342.

Belloni, A. and Chernozhukov, V. (2009). On the computational complexity of MCMC-based estimators in large samples. *Annals of Statistics* **37** 2011–2055.

Berk, R. H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *Annals of Mathematical Statistics* **37** 51–58.

Bhatnagar, N. and Randall, D. (2004). Torpid mixing of simulated tempering on the Potts model. In *Proceedings of the 15th ACM/SIAM Symposium on Discrete Algorithms* 478–487. Association for Computing Machinery, New York.

Borgs, C., Chayes, J. T., Frieze, A., Kim, J. H., Tetali, P., Vigoda, E. and Vu, V. H. (1999). Torpid mixing of some MCMC algorithms in statistical physics. In *Proceedings of the 40th IEEE Symposium on Foundations of Computer Science* 218–229. IEEE, New York.

Del Moral, P., Doucet, A. and Jasra, A. (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society, Series B* **68** 411–436.

Diaconis, P. and Saloff-Coste, L. (1993). Comparison theorems for reversible Markov chains. *Annals of Applied Probability* **3** 696–730.

Diaconis, P. and Saloff-Coste, L. (1996). Logarithmic Sobolev inequalities for finite Markov chains. *Annals of Applied Probability* **6** 695–750.

Diaconis, P. and Stroock, D. (1991). Geometric bounds for eigenvalues of Markov chains. *Annals of Applied Probability* **1** 36-61.

Fort, G., Moulines, E., Roberts, G. O. and Rosenthal, J. S. (2003). On the geometric ergodicity of hybrid samplers. *Journal of Applied Probability* **40** 123–146.

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

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6** 721–741.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford University Press, Oxford.

Green, P. J. and Richardson, S. (2002). Hidden Markov models and disease mapping. *Journal of the American Statistical Association* **97** 1055-1070.

Hans, C., Dobra, A. and West, M. (2007). Shotgun stochastic search for regression with many candidate predictors. *Journal of the American Statistical Association* **102** 507–516.

Jarner, S. F. and Hansen, E. (2000). Geometric ergodicity of Metropolis algorithms. *Stochastic Processes and their Applications* **85** 341–361.

Jensen, S. T., Liu, X. S., Zhou, Q. and Liu, J. S. (2004). Computational discovery of gene regulatory binding motifs: A Bayesian perspective. *Statistical Science* **19** 188–204.

Johnson, A. A. and Jones, G. L. (2010). Gibbs sampling for a Bayesian hierarchical general linear model. *Electronic Journal of Statistics* **4** 313–333.

Jones, G. L. and Hobert, J. P. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statistical Science* **16** 312–334.

Jones, G. L. and Hobert, J. P. (2004). Sufficient burn-in for Gibbs samplers for a hierarchical random effects model. *Annals of Statistics* **32** 784–817.

Kamatani, K. (2011). Weak consistency of Markov chain Monte Carlo methods. Technical report, available at http://arxiv.org/abs/1103.5679.

Kellis, M., Patterson, N., Birren, B., Berger, B. and Lander, E. S. (2004). Methods in comparative genomics: Genome correspondence, gene identification, and regulatory motif discovery. *Journal of Computational Biology* **11** 319–355.

Kullback, S. (1959). *Information Theory and Statistics*. Wiley, New York.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F. and Wootton, J. C. (1993). Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science* **262** 208–214.

Liang, F. and Wong, W. H. (2000). Evolutionary Monte Carlo: Applications to C_p model sampling and change point problem. *Statistica Sinica* **10** 317–342.

Liu, J. S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association* **89** 958–966.

Liu, X., Brutlag, D. L. and Liu, J. S. (2001). BioProspector: Discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pacific Symposium on Biocomputing* **6** 127–138.

Liu, J. S., Neuwald, A. F. and Lawrence, C. E. (1995). Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *Journal of the American Statistical Association* **90** 1156–1170.

Liu, J. S., Wong, W. H. and Kong, A. (1995). Covariance structure and convergence rate of the Gibbs sampler with various scans. *Journal of the Royal Statistical Society, Series B* **57** 157–169.

Madras, N. and Randall, D. (2002). Markov chain decomposition for convergence rate analysis. *Annals of Applied Probability* **12** 581–606.

Madras, N. and Zheng, Z. (2003). On the swapping algorithm. *Random Structures and Algorithms* **1** 66-97.

Mira, A. (2001). Ordering and improving the performance of Monte Carlo Markov chains. *Statistical Science* **16** 340–350.

Mossel, E. and Vigoda, E. (2006). Limitations of Markov chain Monte Carlo algorithms for Bayesian inference of phylogeny. *Annals of Applied Probability* **16** 2215–2234.

Neuwald, A. F., Liu, J. S. and Lawrence, C. E. (1995). Gibbs motif sampling: Detection of bacterial outer membrane protein repeats. *Protein Science* **4** 1618–1632.

Peskun, P. H. (1973). Optimum Monte Carlo sampling using Markov chains. *Biometrika* **60** 607–612.

Roberts, G. O. and Rosenthal, J. S. (2004). General state space Markov chains and MCMC algorithms. *Probability Surveys* **1** 20–71.

Roberts, G. O. and Sahu, S. K. (2001). Approximate pre-determined convergence properties of the Gibbs sampler. *Journal of Computational and Graphical Statistics* **10** 216–229.

Rosenthal, J. S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of the American Statistical Association* **90** 558–566.

Rosenthal, J. S. (1996). Analysis of the Gibbs sampler for a model related to James-Stein estimators. *Statistics and Computing* **6** 269–275.

Roth, F. P., Hughes, J. D., Estep, P. W. and Church, G. M. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnology* **16** 939–945.

Sinclair, A. (1992). Improved bounds for mixing rates of Markov chains and multicommodity flow. *Combinatorics, Probability, and Computing* **1** 351–370.

Tierney, L. (1998). A note on Metropolis-Hastings kernels for general state spaces. *Annals of Applied Probability* **8** 1–9.

Woodard, D. B., Schmidler, S. C. and Huber, M. (2009a). Conditions for rapid

mixing of parallel and simulated tempering on multimodal distributions. *Annals of Applied Probability* **19** 617–640.

Woodard, D. B., Schmidler, S. C. and Huber, M. (2009b). Sufficient conditions for torpid mixing of parallel and simulated tempering. *Electronic Journal of Probability* **14** 780–804.

Dawn B. Woodard, Assistant Professor
School of Operations Research
and Information Engineering,
Cornell University
206 Rhodes Hall, Ithaca, NY 14850.
URL: http://people.orie.cornell.edu/woodard

Jeffrey S. Rosenthal, Professor
Department of Statistics,
University of Toronto
100 St. George St., Rm. 6018,
Toronto, Ontario M5S 3G3.
URL: http://www.probability.ca/jeff