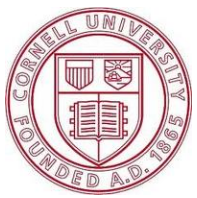
Convergence Rate of Markov Chain Methods for Genomic Motif Discovery

Dawn Woodard
Operations Research and Information Engineering
Cornell University

with Jeffrey Rosenthal, University of Toronto

Workshop on Advances in Monte Carlo, Utah, 2011



Outline

Outline

Motif Discovery

- Regulatory motifs: short DNA subsequences that control gene expression
- Goal: Find these motifs by detecting patterns that occur more often than expected in a long DNA sequence
- Neither pattern nor occurrence locations are known
- Very hard problem since motifs are short, and vary between occurrences
- One of the most popular methods is based on a statistical model and associated Gibbs sampler (Liu, Neuwald, & Lawrence 1995)

Overview

- We **analyze the convergence rate of the Gibbs sampler**
- Show that **if there is more than one true motif** (will define) **the convergence rate decreases exponentially in the length of the DNA sequence**
  - Equivalently, run time increases exponentially in sequence length
  - In practice typically have $>$ 1 true motif
- **Matches empirical results**: sampler gets stuck in local modes and is used only to find candidate motifs

Overview

We also have **progress towards a two-sided result:**

- We give empirical evidence that the Gibbs sampler is efficient if there is no more than one true motif and inefficient if there is more than one
- Conjecture: convergence rate decreases exponentially iff have > 1 true motif
- Supporting this, we prove polynomial decay of the convergence rate for a case with no true motifs

Overview

- **Some of the few meaningful bounds on convergence rate of a Markov chain for a statistical application**, as a function of statistical quantities like # observations, # groups, . . .

- **Previous examples have mainly been for stylized target distributions** like mixtures of normals or Potts models, not for posterior dist'ns from statistical practice

  - Roberts and Sahu (2001): approximate the rate of convergence of Gibbs samplers for unimodal posterior densities in $\mathbb{R}^d$ by approximating with a normal dist'n.

  - Guan and Krone (2007): Bound convergence rate of a particular MC on a mixture of log-concave densities

  - We just learned Scott Schmidler has independently obtained some results for the motif-discovery sampler

Overview

- **Some of the few meaningful bounds on convergence rate of a Markov chain for a statistical application**, as a function of statistical quantities like # observations, # groups, . . .

- **Previous examples have mainly been for stylized target distributions** like mixtures of normals or Potts models, not for posterior dist'ns from statistical practice

  - Roberts and Sahu (2001): approximate the rate of convergence of Gibbs samplers for unimodal posterior densities in $\mathbb{R}^d$ by approximating with a normal dist'n.

  - Guan and Krone (2007): Bound convergence rate of a particular MC on a mixture of log-concave densities

  - We just learned Scott Schmidler has independently obtained some results for the motif-discovery sampler

Overview

Showing that a Markov chain used in statistics is well-behaved typically consists of proving that it is geometrically ergodic

The motif sampler is uniformly ergodic (stronger) but often very inefficient

Numerous other statistical Markov chains may also be exponential-time (based on poor empirical behavior in large problems)

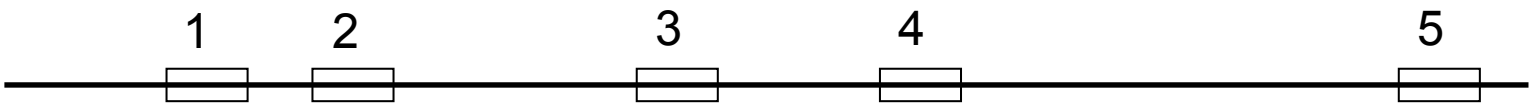
- model averaging in the context of regression with a large number of predictors
- MC for spatial mixture models based on Markov random fields, with many spatial locations

Outline

Outline

Motif Discovery

Goal: Discover repeating pattern in long DNA sequence.



1: GGCTAT
2: GGGTAT
3: AGCTAT
4: GGCTAT
5: GGCTAT

Top: sequence, bottom: highlighted subsequences

Both locations and pattern unknown;
Pattern may vary between occurrences

Motif Discovery

Gibbs sampling method very popular (software packages Bioprospector and AlignACE), but like other methods for motif discovery gives different answers from different starting locations

Actually a family of methods; **we analyze a representative model & associated Gibbs sampler**

Motif Discovery

Model (Liu, Neuwald, & Lawrence 1995):

- **S**: sequence of nucleotides $\mathbf{S} \in \{1, \ldots, M\}^L$ (in practice $M = 4$)
- w: fixed motif length
- **A**: unknown vector of indicators that a motif starts in each possible location in the sequence.
  - Actual model: motif can start in any site
  - We analyze a simplified case where a motif can only end at locations divisible by w, so $\mathbf{A} \in \{0, 1\}^{L/w}$.
  - $\mathbf{A}_i = 1$ means $S_{wi-w+1:wi}$ is a motif occurrence

Motif Discovery

Model:

- θ_j: unknown length-M vector of probabilities for each nucleotide at position j in the motif; $j = 1, \ldots, w$.
- θ_0: unknown length-M vector of probabilities for each nucleotide in non-motif sites

Let Θ be the $w \times M$ matrix having rows θ_j for $j = 1, \ldots, w$

Called the "position-specific frequency matrix"; defines the motif

Motif Discovery

Likelihood:

- $\mathbf{N}(\mathbf{A}_{(j)})$: vector of counts of each nucleotide in position j of all motif occurrences, given $\mathbf{A}$
- $\mathbf{N}(\mathbf{A}^c)$: vector of counts of each nucleotide in all non-motif sites
- For any two vectors $\beta = (\beta_1, \ldots, \beta_K)$ and $\mathbf{N} = (n_1, \ldots, n_K)$, define the notation

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

  where $\Gamma(\cdot)$ is the gamma function.

- $\Rightarrow$ Full-data likelihood:

$$\pi(\mathbf{S}|\Theta, \theta_0, \mathbf{A}) = \theta_0^{\mathbf{N}(\mathbf{A}^c)} \times \prod_{j=1}^{w} \theta_j^{\mathbf{N}(\mathbf{A}_{(j)})}$$

Motif Discovery

- Priors:
  - $\theta_j \sim \text{Dirichlet}(\beta_j)$; $j = 0, \ldots, w$
  - $\mathbf{A}_i \overset{\text{iid}}{\sim} \text{Bernoulli}(p_0)$ for fixed p_0.
- $\Rightarrow$ **Posterior distribution**:

$$\pi(\mathbf{A}, \Theta, \theta_0 | \mathbf{S}) \propto p_0^{|\mathbf{A}|}(1-p_0)^{L/w-|\mathbf{A}|} \times \theta_0^{\mathbf{N}(\mathbf{A}^c)+\beta_0-1} \times \prod_{j=1}^{w} \theta_j^{\mathbf{N}(\mathbf{A}_{(j)})+\beta_j-1}$$

Motif Discovery

Gibbs sampler:

Can integrate out Θ, θ_0 to get a posterior distribution on $\mathbf{A}$:

$$\pi(\mathbf{A}|\mathbf{S}) \propto p_0^{|\mathbf{A}|}(1-p_0)^{L-|\mathbf{A}|} \frac{\Gamma(\mathbf{N}(\mathbf{A}^c)+\beta_0)}{\Gamma(|\mathbf{N}(\mathbf{A}^c)|+|\beta_0|)} \prod_{j=1}^{w} \frac{\Gamma(\mathbf{N}(\mathbf{A}_{(j)})+\beta_j)}{\Gamma(|\mathbf{N}(\mathbf{A}_{(j)})|+|\beta_j|)}.$$

Update each $\mathbf{A}_i$ one-at-a-time according to its conditional distribution.

Motif Discovery

Reason for the simplification (that motif can only end at locations *wi*):

- **Phase shift problem** of original Gibbs sampler: gets stuck in minor modes corresponding to a shifted version of true motif
- **Solution proposed (Liu 1994)**: add a Metropolis step for shifting the motif
- **Our simplified Gibbs sampler captures the dynamics of the original Gibbs sampler, minus the phase-shift**
  - assumes that it is adequately addressed by Liu's fix
  - "best-case" analysis

Outline

Markov Chain Convergence

For a discrete-space Markov chain with transition matrix T, reversible with respect to target distribution π:

- The spectral gap is $\mathbf{Gap}(T) \equiv 1 - \lambda_2$ where λ_2 is the 2nd-largest eigenvalue of T
- # iterations required to obtain n_0 approximately independent samples from π is $O(n_0 \mathbf{Gap}(T)^{-1} \log \|\pi_0 - \pi\|_{L^2})$, where π_0 is the initial dist'n.
- If $\mathbf{Gap}(T)$ *decreases* exponentially in the dimension, the run time *increases* exponentially and the chain is "**slowly mixing**."
- $\mathbf{Gap}(T)$ decreases polynomially in the dimension: "**rapidly mixing**"

Markov Chain Convergence

For a discrete-space Markov chain with transition matrix T, reversible with respect to target distribution π:

- The spectral gap is $\mathbf{Gap}(T) \equiv 1 - \lambda_2$ where λ_2 is the 2nd-largest eigenvalue of T
- # iterations required to obtain n_0 approximately independent samples from π is $O(n_0 \mathbf{Gap}(T)^{-1} \log \|\pi_0 - \pi\|_{L^2})$, where π_0 is the initial dist'n.
- If $\mathbf{Gap}(T)$ *decreases* exponentially in the dimension, the run time *increases* exponentially and the chain is "**slowly mixing**."
- $\mathbf{Gap}(T)$ decreases polynomially in the dimension: "**rapidly mixing**"

Markov Chain Convergence

For a discrete-space Markov chain with transition matrix T, reversible with respect to target distribution π:

- The spectral gap is $\mathbf{Gap}(T) \equiv 1 - \lambda_2$ where λ_2 is the 2nd-largest eigenvalue of T
- # iterations required to obtain n_0 approximately independent samples from π is $O(n_0 \mathbf{Gap}(T)^{-1} \log \|\pi_0 - \pi\|_{L^2})$, where π_0 is the initial dist'n.
- If $\mathbf{Gap}(T)$ *decreases* exponentially in the dimension, the run time *increases* exponentially and the chain is "**slowly mixing**."
- $\mathbf{Gap}(T)$ decreases polynomially in the dimension: "**rapidly mixing**"

Markov Chain Convergence

For a discrete-space Markov chain with transition matrix T, reversible with respect to target distribution π:

- The spectral gap is $\mathbf{Gap}(T) \equiv 1 - \lambda_2$ where λ_2 is the 2nd-largest eigenvalue of T
- # iterations required to obtain n_0 approximately independent samples from π is $O(n_0 \mathbf{Gap}(T)^{-1} \log \|\pi_0 - \pi\|_{L^2})$, where π_0 is the initial dist'n.
- If $\mathbf{Gap}(T)$ *decreases* exponentially in the dimension, the run time *increases* exponentially and the chain is "**slowly mixing**."
- $\mathbf{Gap}(T)$ decreases polynomially in the dimension: "**rapidly mixing**"

Outline

Outline

Slow Mixing Result

Theorem

Defining $J > 0$ position-specific frequency matrices Θ^{j*} for $j = 1, \ldots, J$ and a background frequency vector θ_0^*, assume that:

1. *The data subsequences $\mathbf{S}_{wi-w+1:wi}$ indexed by i are generated i.i.d. from the following distribution G: with probability $p_{0j} > 0$ generated from the motif Θ^{j*}, and otherwise generated according to the background frequencies θ_0^*.*
2. *There is no equivalent data-generating mechanism with smaller J.*

If there are multiple true motifs *($J > 1$), subject to the Condition below, and taking $p_0 = \sum_{j=1}^{J} p_{0j}$* ***the spectral gap of the Gibbs sampler decreases exponentially in the sequence length*** *L, almost surely w.r.t. G.*

Slow Mixing Result

Slow mixing example with $w = 5$ & $M = 4$:

- **S** generated as the concatenation of many length-5 subsequences, each of which is either $(1, 3, 2, 2, 3)$ w.p. 0.003, or $(4, 2, 4, 1, 1)$ w.p. 0.001, or generated as i.i.d. noise.

Slow Mixing Result

In practice there are typically multiple true motifs, corresponding to repeating patterns that have various biological purposes

Goal is to find the most frequently-occurring and well-conserved (Neuwald, Liu, Lawrence 1995, Roth et al. 1998)

Our Thm. says that this contradicts the model assumption of a single motif,

causing slow mixing

Slow Mixing Result

Hard to prove because:

1. Posterior density has a complex form
2. Data, and thus posterior density, are stochastic

Address by using **Bayesian asymptotics** for the case where the data are not drawn from model (Berk 1966).

- Results only available for continuous parameter spaces
- Have to apply to a continuous parametrization, then map to the discrete parameterization **A**.

Slow Mixing Result

Technical Condition

Let $f(s|\Theta, \theta_0)$ be the density of each observation $\mathbf{S}_{wi-w+1:wi}$ under the model. The true density can be written

$$g(s) = \sum_{j=1}^{J} \frac{p_{0j}}{p_0} f(s|\Theta^{j*}, \theta_0^*)$$

Slow Mixing Result

Condition

The Kullback-Leibler divergence between $f(s|\Theta, \theta_0)$ *and* $g(s)$*,*

$$\sum_s g(s) \log \frac{g(s)}{f(s|\Theta, \theta_0)}$$

has multiple local minima (as a function of (Θ, θ_0)*).*

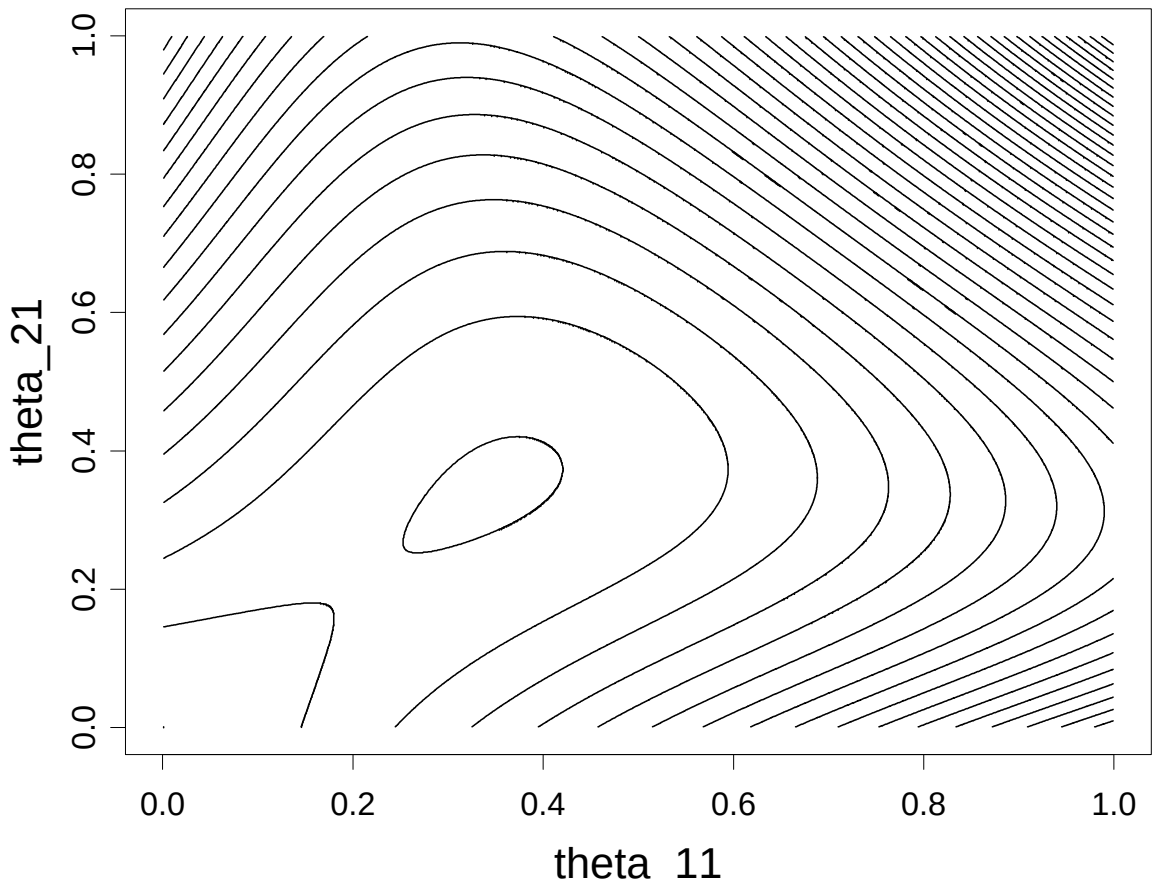
Should usually hold since $g(s)$ is a mixture of the densities $f(s|\Theta^{j*}, \theta_0^*)$; divergence should be smallest when $(\Theta, \theta_0) \approx (\Theta^{j*}, \theta_0^*)$ for some j

Slow Mixing Result

- Have verified this condition for many cases with varying true parameter values.
- Did this for $w = 2$, by plotting K-L divergence as a function of the three parameters $(\theta_{11}, \theta_{21}, \theta_{01})$ and observing multiple local minima.
- Used $J = 2$ and varied the true Θ^{j*}, θ_0^*, p_{0j}. Even for very extreme values the condition holds.
- Multimodality problem should be even worse for longer motifs ($w > 2$).

Slow Mixing Result

K-L divergence for a particular case and a fixed value of θ_{01}:



Outline

Rapid Mixing Result

Simulations suggest that the sampler is rapidly mixing iff $J \leq 1$ (no more than one true motif).

We have one result in this direction, showing rapid mixing for the case where $w = 1$ (in this case any true motif is indistinguishable from the background signal, so $J = 0$)

Rapid Mixing Result

Theorem

For $w = 1$ and any fixed p_0 the spectral gap decreases polynomially in L; in particular,

$$\mathbf{Gap}(T) = \Omega(L^{-14})$$

uniformly over possible values of the data vector $\mathbf{S}$.

Outline

Simulations

Simulate data with varying #s of motifs, according to data-generating mechanism from main Theorem.

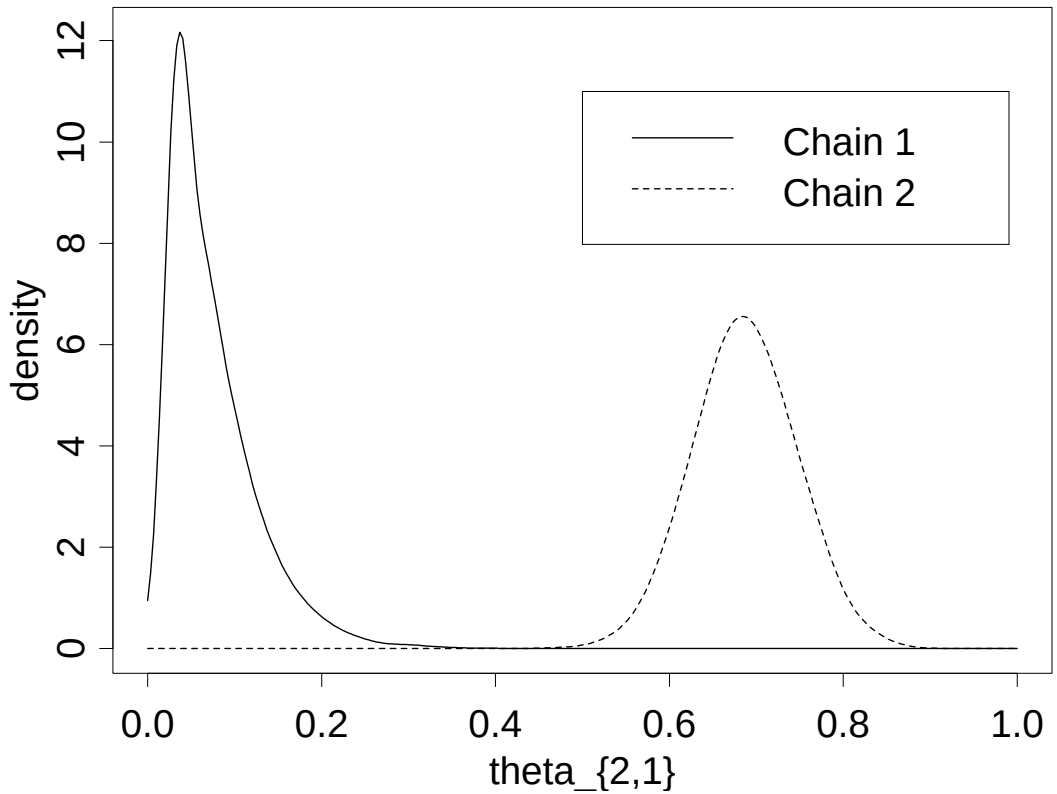
- True Θ^{j*} for each motif sampled from a product Dirichlet dist'n
- Background probabilities θ_0^* from a Dirichlet

For each simulated dataset **S**, **run the Gibbs sampler 5 times starting from different initial vectors A**

- Report the Gelman-Rubin convergence diagnostic
- **Measures whether the chains converged to different modes** (1=good, bigger =bad)

Simulations

Ex: Posterior density estimates of $\theta_{1,2}$ from two Gibbs runs, in the case of 2 true motifs:



Gelman-Rubin scale factor = 10.9

Simulations

% of datasets for which the maximum Gelman-Rubin scale factor is > 1.5:

One motif:

	$w = 6$	$w = 10$	$w = 15$
$L/w = 2,000$	0	0	0
$L/w = 3,000$	0	0	0
$L/w = 4,000$	0	0	0
$L/w = 8,000$	0	0	0

Two motifs:

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

Simulations

One motif: Good Markov chain convergence.

Two motifs: Multiple chains converge to different modes, for L large enough.

Results do not seem to depend on the choice of p_0

Outline

Outline

Alternative Slow Mixing Results

Recall the main Theorem:

Theorem

Defining $J > 0$ position-specific frequency matrices Θ^{j*} for $j = 1, \ldots, J$ and a background frequency vector θ_0^*, assume that:

1. *The data subsequences $\mathbf{S}_{wi-w+1:wi}$ indexed by i are generated i.i.d. from the following distribution G: with probability $p_{0j} > 0$ generated from the motif Θ^{j*}, and otherwise generated according to the background frequencies θ_0^*.*
2. *There is no equivalent data-generating mechanism with smaller J.*

If there are multiple true motifs ($J > 1$), subject to the Condition, and taking $p_0 = \sum_{j=1}^{J} p_{0j}$ the spectral gap of the Gibbs sampler decreases exponentially in the sequence length L, almost surely w.r.t. G.

Alternative Slow Mixing Results

- Simulations suggest that the same result holds for any fixed value of p_0, and that the technical condition holds in general.
- We analyze the closed form of the posterior density to show slow mixing for all p_0 small enough, in specific cases.

Alternative Slow Mixing Results

Case with $w = 2$, $M = 2$, two true motifs (the deterministic sequences $(1,1)$ and $(2,2)$), and no non-motif sites:

Theorem

When $w = 2$, $M = 2$, $p_0 < 1/4$, and **S** *consists of a concatenation of $(1,1)$ and $(2,2)$ subsequences in equal numbers (e.g.* **S** $= 111122112222$ *for $L = 12$),* **Gap**(T) *decreases exponentially in L.*

Alternative Slow Mixing Results

Case with $w = 2$, $M = 2$, two true motifs (the deterministic sequences $(1,1)$ and $(2,2)$), and no non-motif sites:

Theorem

When $w = 2$, $M = 2$, $p_0 < 1/4$, and **S** *consists of a concatenation of $(1,1)$ and $(2,2)$ subsequences in equal numbers (e.g.* $\mathbf{S} = 111122112222$ *for $L = 12$),* $\mathbf{Gap}(T)$ *decreases exponentially in L.*

(Real data would additionally have noise, not considered here)

Alternative Slow Mixing Results

Case with two motifs $(1,1)$ **and** $(2,2)$**, and some non-motif sites:**

Theorem

When $w = 2$*,* $M = 2$*, and* $\mathbf{S}$ *consists of* $L/6$ $(1,1)$ *pairs,* $L/6$ $(2,2)$ *pairs,* $L/12$ $(1,2)$ *pairs, and* $L/12$ $(2,1)$ *pairs in any order, for any* p_0 *small enough* $\mathbf{Gap}(T)$ *decreases exponentially in* L*.*

Outline

Proving Slow Mixing

Recall the main Theorem:

Theorem

Defining $J > 0$ position-specific frequency matrices Θ^{j*} for $j = 1, \ldots, J$ and a background frequency vector θ_0^*, assume that:

1. *The data subsequences $\mathbf{S}_{wi-w+1:wi}$ indexed by i are generated i.i.d. from the following distribution G: with probability $p_{0j} > 0$ generated from the motif Θ^{j*}, and otherwise generated according to the background frequencies θ_0^*.*
2. *There is no equivalent data-generating mechanism with smaller J.*

If there are multiple true motifs *($J > 1$), subject to the Condition, and taking $p_0 = \sum_{j=1}^{J} p_{0j}$* ***the spectral gap of the Gibbs sampler decreases exponentially in the sequence length*** *L, almost surely w.r.t. G.*

Proving Slow Mixing

Intution:

(1) **Gap**(T) is controlled by the unimodality or multimodality of the posterior distribution

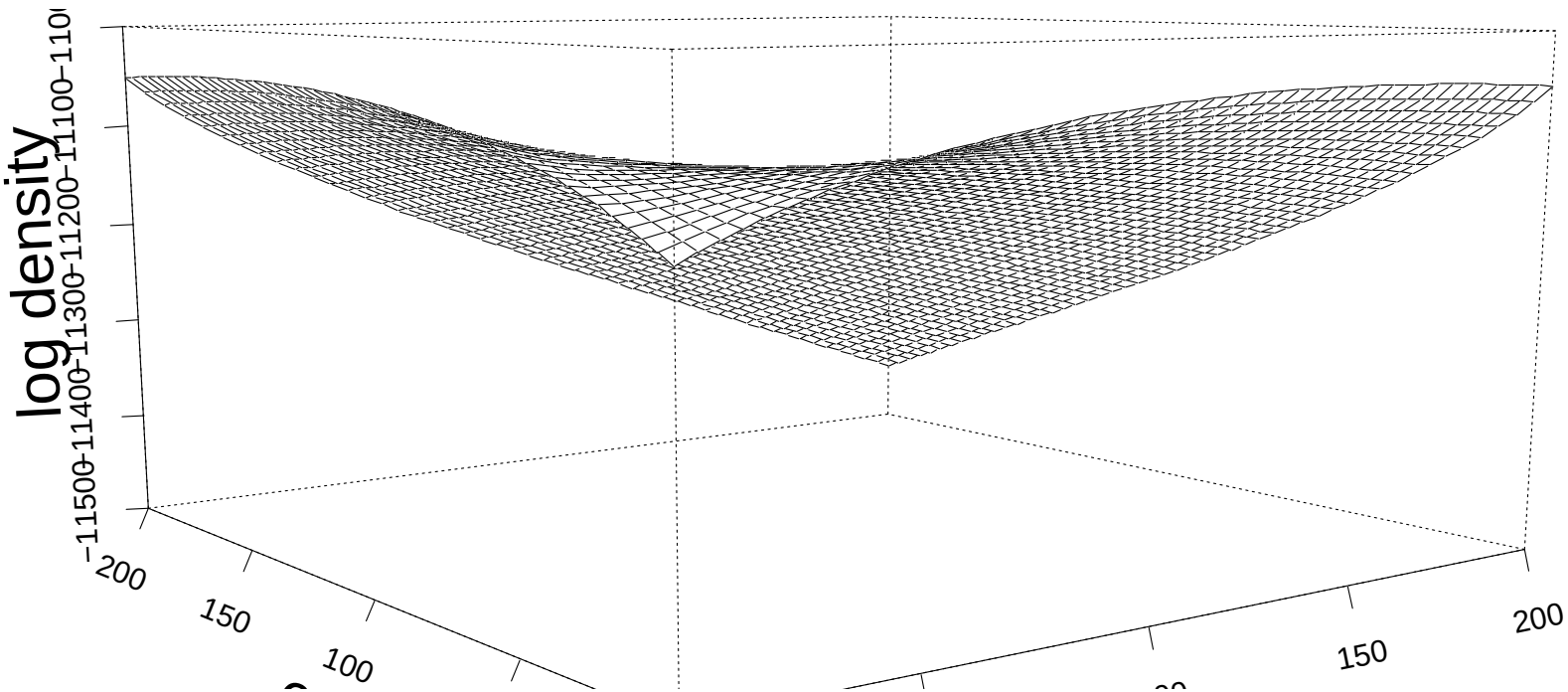
- (Since T makes only local moves)

(2) With multiple true motifs the posterior dist'n has multiple local maxima with height that grows exponentially in L (relative to the height of the valleys in between)

- (Proven using Bayesian asymptotic results for the case of an incorrect model)

Proving Slow Mixing

Plot of posterior density of a 2-D summary of A, for $J = 2$ motifs:



Proving Slow Mixing: Step 2

Proving Step 2: ("With multiple true motifs the posterior dist'n has multiple local maxima with height that grows exponentially in L")

- First we focus on the posterior $\pi(\Theta, \theta_0|\mathbf{S})$ of the continuous parameters (Θ, θ_0).
- We show that $\pi(\Theta, \theta_0|\mathbf{S})$ has multiple local maxima with height that grows exponentially in L
- We map this result to the parametrization on which we simulate the Markov chain: $\pi(\mathbf{A}|\mathbf{S})$

Proving Slow Mixing: Step 2

Proving Step 2: ("With multiple true motifs the posterior dist'n has multiple local maxima with height that grows exponentially in L")

- First we focus on the posterior $\pi(\Theta, \theta_0|\mathbf{S})$ of the continuous parameters (Θ, θ_0).
- We show that $\pi(\Theta, \theta_0|\mathbf{S})$ has multiple local maxima with height that grows exponentially in L
- We map this result to the parametrization on which we simulate the Markov chain: $\pi(\mathbf{A}|\mathbf{S})$

Proving Slow Mixing: Step 2

- We show that $\pi(\Theta, \theta_0|\mathbf{S})$ has multiple modes with height that grows exponentially in L

We do this by applying Bayesian asymptotics

Specifically, results on the asymptotic behavior of the posterior when the model is incorrect (Berk 1966, 1970)

These results are related to the "exponential consistency of posteriors" (Ghosh, Delampady, & Samanta 2006; Choi & Ramamoorthi 2008;...)

Outline

Conclusions

- Gibbs sampling is a **popular method for finding potential gene regulatory binding motifs** in DNA sequences
- It **tends to converge to a local mode**, so it is applied repeatedly with random restarts to generate candidate motifs
  - A better-mixing method could be used to instead find the *best* (most probable) motifs
- We **analyze its convergence rate**, showing exponential decay of the convergence rate when there are multiple true motifs (typically the case)
- We also show polynomial decay of the convergence rate in a case with no identifiable motifs.

Conclusions

- Although it **satisfies a very strong form of ergodicity** (uniform), the MC is typically unusable for obtaining samples from the posterior distribution for long sequences
- **One of the first examples of a Markov chain method that provably fails to obtain samples from the posterior distribution of a statistical model in polynomial time**
- Other stats Markov chains may also have the exponential-time property
  - e.g. model averaging in the context of regression with a large number of predictors

Thanks

Many thanks to NSF for financial support, & to Krzysztof Latuszynski for help with one of the proofs.