Estimating Drift and Minorization Coefficients for Gibbs Sampling Algorithms

David A. Spade

University of Wisconsin–Milwaukee
Department of Mathematical Sciences

Introduction

Bayesian Statistics–often requires sampling from intractable probability distributions.

- Question: How do we sample from these distributions?
- Answer: We often use a Markov chain whose *stationary distribution* is the same as the distribution from which sampling is required.
- Bigger question: How long does this chain need to run so that the states of the chain can be used as approximate samples from the target distribution?
- Where this has come up:
  - Stochastic simulation for evaluating posterior probabilities in Bayesian networks (Hrycej 1990).
  - Differential gene expression analysis (Erkkila, et al., 2010)
  - Educational applications and course redesigns (Sonksen et al., 2013)

Motivating Example

Consider α-mixture of Poisson and Geometric distribution with same mean parameter λ.

- $x_1, \ldots, x_n$: data
- Prior distributions:
  - $\pi(\lambda) \propto \frac{1}{\lambda}$
  - $\pi(\alpha) \propto [\alpha(1-\alpha)]^{\alpha_0 - 1}$, α_0 fixed.
- Likelihood:

$$\mathcal{L}(\mathbf{x}|\alpha, \lambda) \propto \prod_{i=1}^{n} \left\{ \alpha \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} + (1-\alpha)\lambda^{x_i}(1+\lambda)^{-x_i-1} \right\}.$$

- Target Density:

$$p(\alpha, \lambda|\mathbf{x}) \propto \pi(\lambda)\pi(\alpha)\mathcal{L}(\mathbf{x}|\alpha, \lambda).$$

- Very difficult to sample from directly.

Common Difficulties in Bayesian Inference

Primary Difficulty: posterior density can be difficult to sample from directly.

- Can obtain approximate samples in several ways.
- Common way: Markov chain Monte Carlo (MCMC) methods
- Two questions with MCMC
  - How long does the Markov chain have to run before we can use subsequent states as approximate samples from the posterior distribution?
  - How many steps of the chain must we wait between states we can use for sampling?
  - This talk addresses the former of these questions.

Assessing Convergence

The question of how long the chain has to run has been examined in many different ways.

- Many of these are based on the output of the chain.
- Heidelberger and Welch (1983) rely on the theory of the Brownian bridge and spectral analysis.
- Geweke (1992) constructed a method that relies on estimating the spectral density of the chain at time $t = 0$
- Raftery and Lewis (1992) constructed a method that is based on two-state Markov chain theory.
- Gelman and Rubin (1992) use several independent chains to assess convergence by comparing variation between chains and within chains.

Issues with Output-Based Convergence Diagnostics

Methods of convergence assessment are rife with issues.

- They mostly rely on the output of one chain. This gives no indication of how long convergence would take from a chain initiated from any point in the state space.
- The output-based methods are very different, so they can give different diagnoses of whether or not the chain has approximately reached its stationary distribution.
- If the diagnostic method chosen concludes that the chain has not reached its stationary distribution, the chain must be re-run for an even larger number of steps in order to get closer to convergence, and the diagnostic must be re-run as well.
- None of these techniques give a clear idea of what is meant by "closeness" to the stationary distribution.

Other Methods of Assessing Convergence

- Analytical–Rosenthal (1995)–Drift and minorization coefficients (more on this shortly)
- Computational approach–Cowles and Rosenthal (1998)
- Issues:
  1. Theory-based methods can be very difficult to use in practice
  2. Cowles and Rosenthal (1998) becomes computationally intractable in even moderate dimensions
- This talk details remedies to this for common MCMC algorithms.

Goals of the Talk

The goals of this talk are as follows:

- Give the audience some important background information on Markov chains and their convergence behavior.
- Present remedies to common issues in assessing convergence of common MCMC algorithms.
- Provide specific descriptions of doing this for the Gibbs sampler.
- Present mathematical justification for the methods described herein.

Minorization

Definition

A Markov chain $(X_t)_{t\geq 0}$ satisfies a *minorization condition* if there exist $\varepsilon \in (0,1)$ a small set $C \in \mathcal{B}(\mathbb{R}^m)$, a positive integer k, and a probability measure $\nu(\cdot)$ such that for all $\mathbf{x} \in C$ and for all $A \in \mathcal{B}(\mathbb{R}^m)$,

$$K^k(\mathbf{x}, A) \geq \varepsilon \nu(A). \tag{1}$$

Geometric Drift

Definition

A Markov chain $(X_t)_{t\geq 0}$ satisfies a *drift condition* if there exist constants $\lambda \in (0,1)$ and $b < \infty$, a function $V : \mathbb{R}^m \mapsto [1,\infty)$, a positive integer h, and a small set $C \in \mathcal{B}(\mathbb{R}^m)$ such that for all $\mathbf{x} \in \mathbb{R}^m$,

$$K^h V(\mathbf{x}) \leq \lambda V(\mathbf{x}) + b\mathbb{I}_C(\mathbf{x}), \text{ where} \tag{2}$$

$K^h V(\mathbf{x}) = \mathbb{E}[V(X_{t+h})|\mathbf{X}_t = \mathbf{x}]$ and the expectation is taken with respect to the h-step transition kernel.

Rosenthal uses drift and minorization conditions to bound the total variation distance between the n-step transition kernel of a Markov chain and its stationary distribution.

Rosenthal (1995) Theorem

Theorem

Assume that for a function $V : \mathbb{R}^m \mapsto [1, \infty)$, *a positive integer* h, *and constants* $\lambda \in (0, 1)$ *and* $b < \infty$, $(X_t)_{t \geq 0}$ *satisfies*

$$K^h V(\mathbf{x}) \leq \lambda V(\mathbf{x}) + b\mathbb{I}_C(\mathbf{x})$$

for all $\mathbf{x} \in \mathbb{R}^m$, *where* $C = \{\mathbf{x} : V(\mathbf{x}) \leq d\}$ *and* $d > \frac{2b}{1-\lambda} - 1$. *Assume that for some* $\varepsilon > 0$, *some probability measure* $\nu(\cdot)$ *on* $\mathcal{B}(\mathbb{R}^m)$, *and some positive integer* k_0,

$$K^{hk_0}(\mathbf{x}, B) \geq \varepsilon \nu(B)$$

for all $\mathbf{x} \in C$ *and for all* $B \in \mathcal{B}(\mathbb{R}^m)$. *Then for any* $r \in (0, 1)$ *with* $(X_t)_{t \geq 0}$ *beginning in the initial distribution* $\boldsymbol{\Psi}$ *and for any positive integer* n,

$$\begin{aligned}
\delta(K^n, \pi) &\leq (1-\varepsilon)^{\left[\frac{rn}{hk_0}\right]} + (\alpha A)^{-1}\left(\alpha^{-(1-rk_0)}A^r\right)^{\left[\frac{h}{n}\right]} \\
&\times \left(1 + \frac{b}{1-\lambda} + \mathbb{E}_{\boldsymbol{\Psi}}[V(X_0)]\right), \quad \textit{where} \\
\alpha^{-1} &= \frac{1 + 2b + \lambda d}{1 + d}, \\
A &= 1 + 2(\lambda d + b)
\end{aligned}$$

Estimating b Coefficient–Cowles and Rosenthal (1998)

(Z_t)–Markov chain with transition kernel $K(\cdot, \cdot)$.

- Estimating b drift coefficient:
  - Find all points $\mathbf{z} \in \mathbb{R}^m$ such that $V(\mathbf{z}) = 1$.
  - Run N_0 h-step chains from each of these initial values, estimate $\mathbb{E}[V(Z_h)|Z_0 = \mathbf{z}] - 1$.
  - Call estimate $b_{CR}(\mathbf{z})$.
  - $\hat{b}_{CR} = \max_{i=1,2,\ldots,N_1} \hat{b}_{CR}(\mathbf{z}) + 1$.
- N_0 chosen to ensure stability in estimation of b.

Estimating λ

Choose N_1 initial states from all over the state space.

- N_1 chosen to ensure stability in estimating λ.
- Run N_2 h-step chains.
- For $i = 1, 2, \ldots, N_2$, where $\mathbf{z}_i$ is the i^{th} initial value, the quantity $e(\mathbf{z}_i) = \mathbb{E}[V(Z_h)|Z_0 = \mathbf{z}_i]$ is estimated in the same way as it was in the estimation of b.
- $$\hat{\lambda}_{CR} = \max_{i=1,2,\ldots,N_1} \frac{e(\mathbf{z}_i) - \hat{b}_{CR}}{V(\mathbf{z}_i)}.$$
- If $\hat{\lambda}_{CR}$ exceeds 1, then choose a larger value of $\hat{b}_{CR}$ and try again.

Estimating ε

The estimation of ε depends on the fact that if $(X_t)_{t\geq 0}$ satisfies a minorization condition, then by a result used by Cowles and Rosenthal (1998),

$$\varepsilon = \int_{\mathbb{R}^m} \inf_{\mathbf{x}_t \in C} k(\mathbf{x}_{t+1}|\mathbf{x}_t)\, \mathrm{d}\mathbf{x}_{t+1}. \tag{3}$$

Estimating ε

Choose value of d that is comfortably larger than $2\hat{b}_{CR}/(1-\hat{\lambda}_{CR})-1$.

- Divide state space into large number of little bins.
- Choose set of initial values from "corners" or bad parts of the set $V_d := \{\mathbf{z} : V(\mathbf{z}) \leq d\}$.
- Run N_3 hk_0-step chains from each initial value, where k_0 is a positive integer.
- Keep track of fraction of N_3 chains that land in each bin.
- For each bin, take minimum fraction of samples landing in that bin.
- Sum of these minima over all the bins provides estimate $\hat{\varepsilon}_{CR}$ of ε.

Limitations of Cowles and Rosenthal (1998) Approach

- Hard to find all points where $V(\mathbf{x}) = 1$.
- Prohibitively expensive computation.
- Estimation of λ and b may be reasonably fast.
- Estimating ε will not be efficient even in moderate dimensions.
  - Divide state space into a lot of bins.
  - Number of bins increases exponentially in the dimension.
  - Number of chains needed to ensure adequate coverage of the bins to ensure stability in estimation of ε also increases exponentially.
  - Later, I will present an approach that helps to mitigate some of these issues.

Bounding the Mixing Time for the Gibbs Sampler

We begin with the estimation of drift coefficients. To estimate λ, we do the following:

- For drift function $V(\cdot)$, choose $N_{C'}$ points $\mathbf{x}$ outside $C = \{\mathbf{x} : V(\mathbf{x}) \leq d\}$ from which to initialize one-step Gibbs samplers.
- From each initial state $\mathbf{x}_i$, run N_0 one-step chains, and for each one compute $\frac{V(\mathbf{x}^*)}{V(\mathbf{x}_i)}$, where $\mathbf{x}^*$ is the new state of the chain.
- Compute the average value of this ratio over the N_0 chains–gives intermediate estimate $\hat{\lambda}(\mathbf{x}_i)$.
- Estimate of λ is given by $\hat{\lambda} = \max_{i=1,\ldots,N_{C'}} \hat{\lambda}(\mathbf{x}_i)$.

Estimating b Coefficient

- Choose N_C points $\mathbf{x}_i$ inside C from which to initialize one-step Gibbs sampling chains.
- Run N_1 one-step chains from each of these points–each results in new state $\mathbf{x}^*$.
- For each chain, compute $V(\mathbf{x}^*)$ and then average these values over the N_1 chains to get an average $\hat{e}(\mathbf{x}_i)$.
- Remove the component involving λ by taking $\hat{b}(\mathbf{x}_i) = \hat{e}(\mathbf{x}_i) - \hat{\lambda} V(\mathbf{x}_i)$. (Drift function may need to be estimated if marginal needs to be estimated).
- Estimate of b is given by $\hat{b} = \max_{i=1,\ldots,N_C} \hat{b}(\mathbf{x}_i)$.

Estimating Minorization Coefficient

We let $R \subset \mathbb{R}^m$ be a grid of points that covers the small set C and where each point in R is in the state space, and let N_R be the number of grid points.

- We take advantage of Equation (3).
- The grid is designed so that all of the hyperrectangles formed by it have the same volume η. (Equal volume is not necessary, but simplifies computation and has little impact on the result).
- Choose N_ε points inside C that could serve as initial states of a Gibbs sampler. Call this collection of points $\hat{C}_\varepsilon$.
- For each point $\mathbf{x} \in R$, compute the transition density $k(\mathbf{x}|\mathbf{x}_t)$ for each $\mathbf{x}_t \in \hat{C}_\varepsilon$.
- Take the minimum of these transition densities over the $\mathbf{x}_t \in \hat{C}_\varepsilon$ and call this minimum $\hat{\varepsilon}(\mathbf{x}_i)$, where $\mathbf{x}_i$ is the i^{th} point in R.
- Do this for each of the points in R and then estimate ε using $\hat{\varepsilon}$ $= \eta \sum_{i=1}^{N_R} \hat{\varepsilon}(\mathbf{x}_i)$.
- Note that if the drift function is analytically intractable, it is difficult to form a grid over C.–Can form grid over $\hat{C}_\varepsilon$ without losing mathematical justification for this procedure.

Simulation Study: The Pareto Distribution

Estimating the parameters of a Pareto distribution with scale parameter c and shape parameter α based on a set of 52 data points from this distribution. A noninformative, improper prior is placed on the pair (α, c), resulting in the posterior density

$$p(\alpha, c|\mathbf{y}) \propto \prod_{i=1}^{52} y_i^{-(\alpha+1)} \alpha^{52} c^{52\alpha} \prod_{i=1}^{52} \mathbb{I}_{[c,\infty)}(y_i),$$

where $\mathbb{I}_A(x)$ denotes the 0-1 indicator function for the set A.
Full conditional distributions

$$\begin{aligned} \alpha|c, \mathbf{y} &\sim \text{Gamma}\left(53, \sum_{i=1}^{52} \log y_i - 52 \log c\right) \\ p(c|\alpha, \mathbf{y}) &\sim c^{52\alpha} \mathbb{I}_{(0,y^*]}(c), \end{aligned}$$

where $y^* = \max_{i=1,\ldots,52} y_i$.

Pareto Example

The drift function

$$V(\alpha, c) = \alpha^{-5.2} c^{-(5.2+0.1\alpha)}$$

was chosen, and the small set was chosen to be

$$C = \{(\alpha, c) \in (0, \infty) \times (0, y^*] : V(\alpha, c) \leq 10\}.$$

- λ Settings: 200 initial states outside of C, 200 one-step chains run from each. $\hat{\lambda} = 0.0003(0.0001)$.
- b settings: 500 points inside C, 200 one-step chains from each. $\hat{b} = 0.008(0.0004)$.
- ε settings: 1000 points inside C.
- Grid formed over the set of selected points from minimum to maximum in each component direction. Increments of 0.05 in each direction. $\hat{\varepsilon} = 0.0480(0.004)$.
- Bound on the mixing time: 940 steps. 51.512 seconds to complete process.

Trace Plots

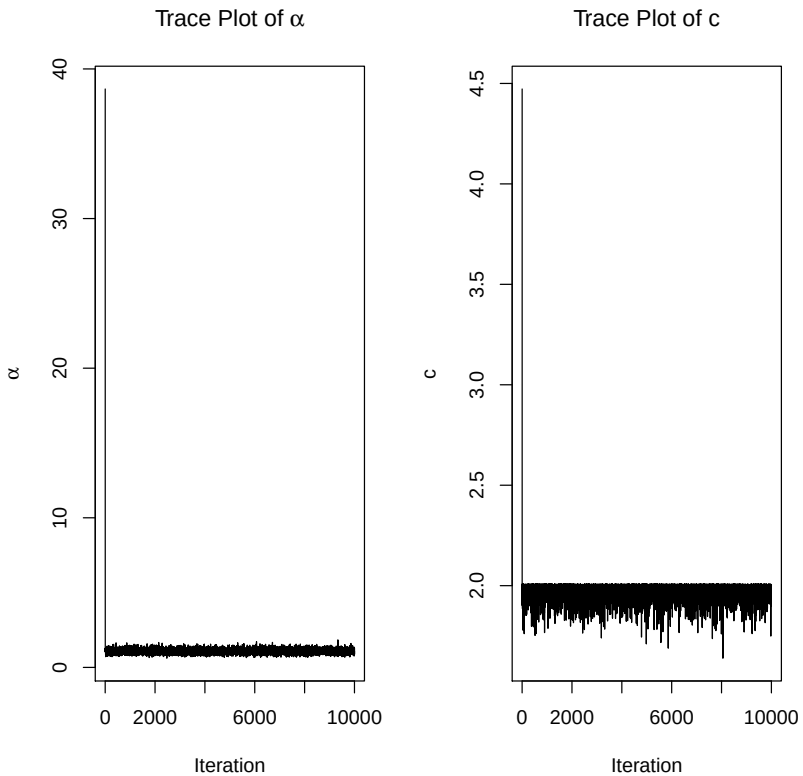

Figure: Trace Plots indicating that 940 steps is sufficient for mixing.

Closing Remarks

The idea here has been to present ways to obtain an approximate bound on mixing time efficiently for Metropolis-Hastings algorithms and Gibbs samplers.

- Advantages:
  - More efficient than the Cowles and Rosenthal (1998) method.
  - Can handle higher dimensions than existing methods.
- Limitations
  - Will still run into the "curse of dimensionality" somewhere–number of grid points needed to estimate ε for the Gibbs sampler gets large in high dimensions.
  - For both algorithms, the number of chains in estimating parameters, as well as the number of initial points needed to obtain stable estimates, gets large in high dimensions.
- Takeaway: While not without some issues, these techniques provide useful intermediaries between output-based diagnostics of convergence and often intractable analytical methods that outperform the previously-existing techniques.