Lower Bounds on the Convergence Time of Adaptive MCMC Methods

Dawn Woodard
Operations Research and Information Engineering
Cornell University

Scott Schmidler
Duke University

MCQMC 2010

Questions

Ask questions / clarifications anytime!

Outline

Outline

Adaptive MCMC

Adaptive MCMC methods have shown empirically better performance than MCMC for a number of examples

There is great interest in general results on the amount of improvement afforded by these methods

Bounds on the convergence ("mixing") time are hard to obtain for these non-Markovian, time-inhomogeneous processes

Mixing Time Bounds for Adaptive MCMC

We derive lower bounds on the mixing time of adaptive methods

- ...using the concepts of hitting time and conductance from the theory of Markov chains

Use these to show slow mixing on multimodal examples:

- A mixture of normals
- The mean-field Potts model

Appear to be the first non-asymptotic bounds on convergence for these methods

Mixing Time Bounds for Adaptive MCMC

- These bounds suggest and in some cases prove that these adaptive methods cannot be rapidly mixing when the Markov chains on which they are based are slowly mixing
  - I.e., no qualitative improvement in convergence in high dimensions
  - They may have shorter autocorrelation times
  - Some methods may successfully *optimize* over the set of Markov kernels
- Their time to convergence is controlled by the conductance of the Markov kernel on which they are based

Adaptive MCMC

2 types of adaptive MCMC methods: **“Multichain resampling”** (MR) and **“Invariant adaptive”** (IA)

Related Work

Existing results relate AMCMC to its limiting kernel, and are asymptotic (in n):

- Andreiu & Atchadé (2007) show that IA methods such that $\theta_n \longrightarrow \theta^*$ share many asymptotic properties with the Markov chain with transition kernel T_{θ^*}
- Atchadé (2009:a) shows that the asymptotic variance of a MR method is at least as large as the asymptotic variance of its limiting kernel

Outline

Outline

MR Methods

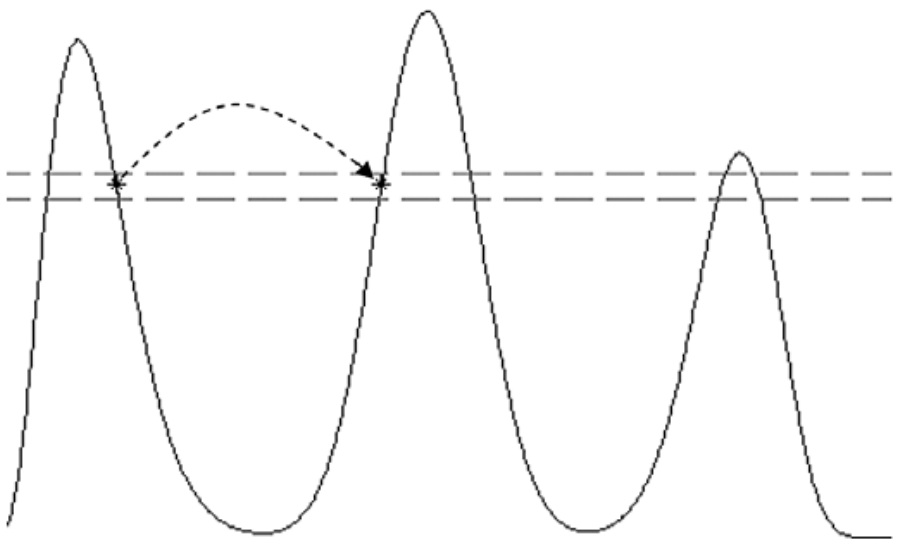
"Multichain resampling" (MR) methods:

- One or more parallel chains
- Supplement local moves with jumps to locations previously visited by one of the chains
- ith chain $X^{(i)}$ has target $\pi^{(i)}$ and transition kernel $K_{i,n} = \alpha T_i + (1-\alpha) R_{i,n}$ where T_i is a $\pi^{(i)}$-reversible Markov kernel
- $R_{i,n}$ are resampling kernels proposing from the set of previous samples $X_{0:n-1}^{(1:I)}$
- $\pi^{(1)} = \pi$

MR Methods

Ex: Equi-energy sampler (Kou, Zhou, & Wong 2006):

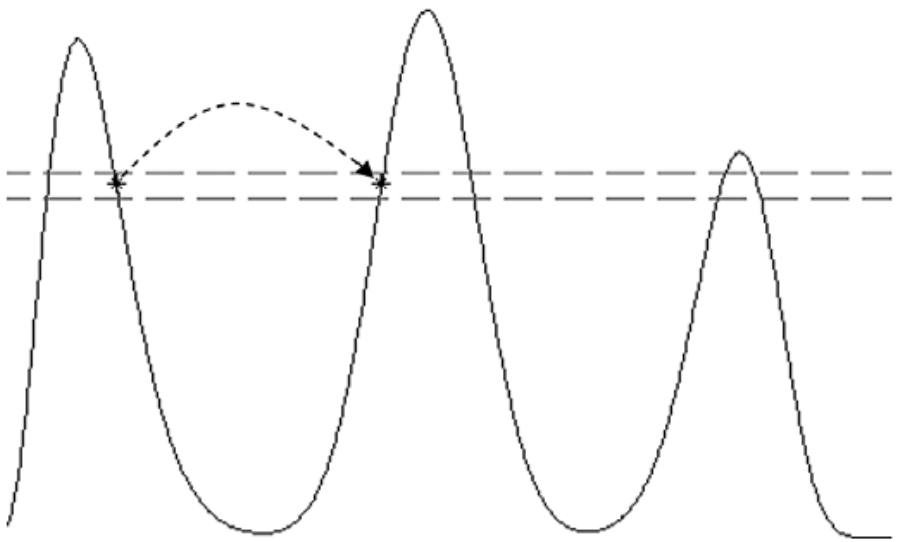
- Introduce moves between points of similar energy (density), allowing jumps between modes:



From Kou et al. 2006

MR Methods

- Multiple chains at different temperatures $\pi^{(i)} \propto \pi^{\beta_i}$
- Occasionally resample from a higher-temperature chain a state having similar energy (density) to the current state



From Kou et al. 2006

Mixing Time

- We are concerned with convergence of $X^{(1)}$ to π:

$$\|\pi_n - \pi\|_{TV} = \sup_{A \subset \mathcal{X}} |\pi_n(A) - \pi(A)|$$

where π_n is the marginal dist'n of $X_n^{(1)}$

- The convergence ("mixing") time is:

$$\tau_\epsilon = \sup_{\pi_0} \left(\min\{n : \|\pi_{n'} - \pi\|_{TV} < \epsilon \quad \forall n' \geq n\} \right) \tag{1}$$

I.e. the time to be within ϵ of the target π for any starting dist'n

- Rapid mixing: τ_ϵ grows polynomially in the dimension

- Slow mixing: τ_ϵ grows exponentially

Mixing Time

- We are concerned with convergence of $X^{(1)}$ to π:

$$\|\pi_n - \pi\|_{TV} = \sup_{A \subset \mathcal{X}} |\pi_n(A) - \pi(A)|$$

  where π_n is the marginal dist'n of $X_n^{(1)}$

- The convergence ("mixing") time is:

$$\tau_\epsilon = \sup_{\pi_0} \left(\min\{n : \|\pi_{n'} - \pi\|_{TV} < \epsilon \quad \forall n' \geq n\} \right) \tag{1}$$

  I.e. the time to be within ϵ of the target π for any starting dist'n

- Rapid mixing: τ_ϵ grows polynomially in the dimension

- Slow mixing: τ_ϵ grows exponentially

Mixing Time

- We are concerned with convergence of $X^{(1)}$ to π:

$$\|\pi_n - \pi\|_{TV} = \sup_{A \subset \mathcal{X}} |\pi_n(A) - \pi(A)|$$

  where π_n is the marginal dist'n of $X_n^{(1)}$

- The convergence ("mixing") time is:

$$\tau_\epsilon = \sup_{\pi_0} \left(\min\{n : \|\pi_{n'} - \pi\|_{TV} < \epsilon \quad \forall n' \geq n\} \right) \tag{1}$$

  I.e. the time to be within ϵ of the target π for any starting dist'n

- Rapid mixing: τ_ϵ grows polynomially in the dimension

- Slow mixing: τ_ϵ grows exponentially

Mixing Time

Will bound the mixing time using:

The hitting time for $A \subset \mathcal{X}$:

$$H_A = \min\{n : \exists i \text{ with } X_n^{(i)} \in A\}$$

The conductance of a π-reversible Markov kernel T:

$$\Phi_T(A) = \frac{\int_A \pi(dv)\, T(v, A^c)}{\pi(A)\pi(A^c)}$$

The prob. of moving between A & A^c

$$\Phi_T = \inf_{\substack{A \subset \mathcal{X}: \\ 0 < \pi(A) < 1}} \Phi_T(A)$$

Quantifies the worst bottle-neck

Mixing Time

Key idea:

$$\Pr(H_A \leq n) \leq \pi(A) - \epsilon \qquad \Rightarrow \qquad \pi_n(A) \leq \pi(A) - \epsilon$$

(to be in A at time n, must have hit A by time n)

$$\Rightarrow \qquad \|\pi_n - \pi\|_{TV} \geq \epsilon \qquad \Rightarrow \qquad \tau_\epsilon > n$$

If the chains are initialized in A^c, then H_A is controlled by $\Phi_T(A)$

so "The time to convergence of an adaptive MCMC method is controlled by the conductance of the Markov kernel on which it is based"

Bound for MR Samplers

Theorem

For any $\epsilon > 0$ and any $A \subset \mathcal{X}$ such that $0 < \pi^{(i)}(A) < 1$ for all i, the mixing time τ_ϵ of the MR process satisfies:

$$\tau_\epsilon \geq (\pi(A) - \epsilon) \left[cl \max_i \gamma(A, i) \Phi_{T_i}(A) \right]^{-1}.$$

where $\gamma(A, i) = \min \left\{ 1, \frac{\pi^{(i)}(A)}{\pi(A)} \right\}$ is the *persistence* (Woodard, Schmidler, & Huber 2009).

Bound for MR Samplers

The **persistence factor** arises since we have **multiple chains** that exchange samples via resampling

Bound for MR Samplers

c arises because adaptive MCMC only asymptotically π-invariant; doesn't monotonically approach π.

We **assume the drift away from π is bounded**:

Assumption

There is a constant $1 \leq c < \infty$ *such that for any* $A \subset \mathcal{X}$ *having* $\pi^{(i)}(A) > 0$, *the sampler* $Y = X|_A$ *with* $Y_0^{(i)} \overset{ind.}{\sim} \pi^{(i)}|_A$ *satisfies the following for all* i *and* n: $\mathcal{L}(Y_n^{(i)})$ *has a density wrt* $\pi^{(i)}|_A$ *that is everywhere* $\leq c$.

(Holds with $c = 1$ for method of Atchadé (2009b) and when $\alpha = 1$).

Bound for MR Samplers

Corollary

For a single chain ($I = 1$), $\tau_\epsilon \geq \frac{1}{4c\Phi_T}$

$\Rightarrow$ Slow mixing of the Markov chain w/ kernel T implies slow mixing of any single-chain MR process based on T
(e.g. method of Atchadé 2009b).

Corollary

For the sampler of Gelfand & Sahu (1994), $\tau_\epsilon \geq \frac{1}{4cI\Phi_T}$

$\Rightarrow$ Slow mixing of the Markov chain w/ kernel T implies slow mixing of any Gelfand-Sahu sampler based on T.

Bound for MR Samplers

Our bound:

$$\tau_\epsilon \geq (\pi(A) - \epsilon) \left[cl \max_i \gamma(A, i) \Phi_{T_i}(A) \right]^{-1}$$

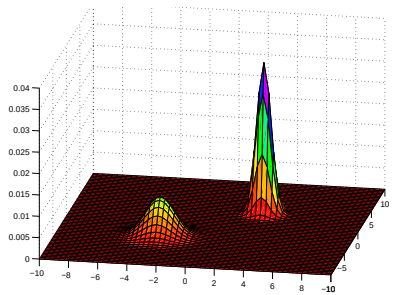
looks like a bound for non-adaptive swapping processes (MC^3) obtained by Woodard, Schmidler & Huber (2009):

$$\tau_\epsilon^* \geq 2^{-8} \ln(2\epsilon)^{-1} \left[\max_i \gamma(A, i) \Phi_{T_i}(A) \right]^{-1/2}.$$

Slow Mixing on Multimodal Examples

Ex: Mixture of normals in $\mathbb{R}^d$

$$\pi(x) = \frac{1}{2} N_d(x; -\mu \mathbf{1}_d, \sigma_1^2 \mathsf{I}_d) + \frac{1}{2} N_d(x; \mu \mathbf{1}_d, \sigma_2^2 \mathsf{I}_d)$$



Theorem (WSH09a): MC^3 is rapidly mixing for $\sigma_1 = \sigma_2$
Theorem (WSH09b): MC^3 is slowly mixing for $\sigma_1 \neq \sigma_2$

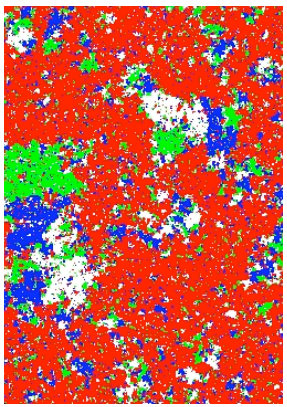
Theorem: MR adaptive methods are slowly mixing for $\sigma_1 \neq \sigma_2$
(based on local M-H proposals, & for $\pi^{(i)}$ constructed by tempering)

Slow Mixing on Multimodal Examples

Ex: Mean-field Potts model (ferromagnetic, $q > 3$)

$$\pi(z) \propto \exp\left\{ \frac{\alpha}{2M} \sum_{i,j} \mathbf{1}_{\{z_i = z_j\}} \right\}$$

$$z \in \{1, \ldots, q\}^M, \qquad \alpha \geq 0$$



Theorem (BR04,WSH09): MC^3 is slowly mixing

Theorem: So are MR adaptive methods
(based on local M-H proposals, & for $\pi^{(i)}$ constructed by tempering)

Outline

IA Methods

1. "Invariant adaptive" (IA) methods:

- Adaptively optimize over a family of transition kernels $\{T_\theta : \theta \in \Theta\}$ that are invariant with respect to the target distribution π.

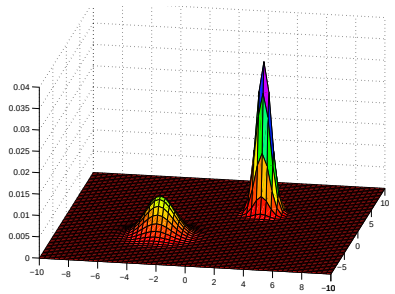
Ex: Adaptive Metropolis (Haario, Saksman, & Tamminen 2001):

- Adapt the covariance θ_n of the MVN proposal using the empirical covariance of the samples

Bound for IA Samplers

Ex: Mixture of normals in $\mathbb{R}^d$

$$\pi(x) = \frac{1}{2} N_d(x; -\mu \mathbf{1}_d, \sigma_1^2 \mathsf{I}_d) + \frac{1}{2} N_d(x; \mu \mathbf{1}_d, \sigma_2^2 \mathsf{I}_d)$$



Theorem

Inter-chain Adaptation (Craiu, Rosenthal, & Yang 2009) and Adaptive Metropolis are slowly mixing for $\sigma_1 \neq \sigma_2$

Outline

Conclusions

- To our knowledge we have given the first non-asymptotic bounds on convergence for adaptive MCMC
- Our results for MR samplers extend results for $(MC)^3$
  - suggest that they may not be able to provide a speedup from slow to rapid mixing
  - these methods may not be worth their greater difficulty (theoretical, practical)
  - formalize intuition that jumps back to previously visited locations cannot speed exploration of new areas
  - imply slow mixing on a mixture of normals and the mean-field Potts model
- Results showing specific IA samplers are slowly mixing on a multimodal example

Conclusions

Our manuscript and these slides can be found on my website:
people.orie.cornell.edu/woodard

References


Andrieu, C. and Atchadé, Y. F. (2007).
On the efficiency of adaptive MCMC algorithms.
Electronic Communications in Probability, 12:336-349.


Atchadé, Y. F. (2009:a).
A cautionary tale on the efficiency of some adaptive MCMC algorithms.
Annals of Applied Probability, accepted.


Atchadé, Y. F. (2009:b).
Resampling from the past to improve on MCMC algorithms.
Far East Journal of Theoretical Probability, 27:81-99.


Gelfand, A. E. and Sahu, S. K. (1994).
On Markov chain Monte Carlo acceleration.
J. of Computational and Graphical Statistics, 3, 261–276.


Kou, S. C., Zhou, Q. and Wong, W. H. (2006).
Equi-energy sampler with applications in statistical inference and statistical mechanics.
Annals of Statistics, 34, 1581-1619.


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

Extra Material

The remaining slides provide extra information on the topics covered in the seminar.

Adaptive MCMC

- Adaptive Markov chain Monte Carlo methods have been introduced for high-dimensional integration problems arising in Bayesian analysis and statistical mechanics
- Markov chain techniques cannot use the information from previous iterations of the chain to alter the transition kernel to speed convergence
- Adaptive MCMC techniques do this, relying on alternative arguments to ensure convergence to the target distribution

Adaptive MCMC

- Adaptive Markov chain Monte Carlo methods have been introduced for high-dimensional integration problems arising in Bayesian analysis and statistical mechanics
- Markov chain techniques cannot use the information from previous iterations of the chain to alter the transition kernel to speed convergence
- Adaptive MCMC techniques do this, relying on alternative arguments to ensure convergence to the target distribution

Adaptive MCMC

- Adaptive Markov chain Monte Carlo methods have been introduced for high-dimensional integration problems arising in Bayesian analysis and statistical mechanics
- Markov chain techniques cannot use the information from previous iterations of the chain to alter the transition kernel to speed convergence
- Adaptive MCMC techniques do this, relying on alternative arguments to ensure convergence to the target distribution

IA Methods

More IA methods:

Inter-chain adaptation (Craiu, Rosenthal, & Yang 2009):

- Parallel Metropolis chains with invariant distribution π and normal proposal with covariance θ_n
- Adapt θ_n using past samples from all chains

MR Methods

More MR methods:

- Importance-resampling MCMC (Atchadé 2009:a)
- Gelfand & Sahu (1994)

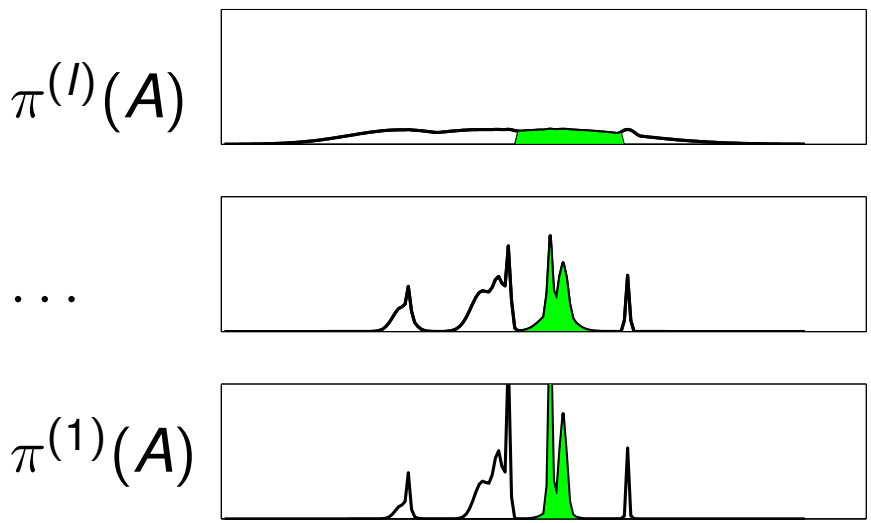
MR Methods

Equi-energy sampler:

- Simulate chains at different temperatures $\pi^{(i)} \propto \pi^{\beta_i}$ for $1 = \beta_1 > \ldots > \beta_I \geq 0$
- Bin the state histories of each process i according to energy
- For $i < I$ the process $X^{(i)}$ occasionally proposes a move from the next-highest-temp. chain $X^{(i+1)}_{0:n-1}$ that is in the same energy bin as the current state $X^{(i)}_{n-1}$
- Accept with probability $\rho(x, y) = \min\left\{1, \frac{\pi^{(i)}(y)\pi^{(i+1)}(x)}{\pi^{(i)}(x)\pi^{(i+1)}(y)}\right\}$

The Persistence: Does a set persist across levels?

Consider the probability of A under $\pi^{(i)}$ as a function of i:



The persistence involves the ratio of $\pi^{(i)}(A)$ to $\pi(A)$