Submitted to the Annals of Statistics

LOWER BOUNDS ON THE CONVERGENCE RATES OF ADAPTIVE MCMC METHODS

By Scott C. Schmidler
and Dawn B. Woodard

Duke University and Cornell University

We consider the convergence properties of recently proposed adaptive Markov chain Monte Carlo (MCMC) algorithms for approximation of high-dimensional integrals arising in Bayesian analysis and statistical mechanics. Despite their name, in the general case these algorithms produce non-Markovian, time-inhomogeneous, irreversible stochastic processes. Nevertheless, we show that lower bounds on the mixing times of these processes can be obtained using familiar ideas of hitting times and conductance from the theory of reversible Markov chains. While loose in some cases, the bounds obtained are sufficient to demonstrate slow mixing of several recently proposed algorithms including the adaptive Metropolis algorithm of Haario et al. (2001), the equi-energy sampler (Kou et al., 2006), and the importance-resampling MCMC algorithm (Atchadé, 2009*b*) on some multimodal target distributions including mixtures of normal distributions and the mean-field Potts model. These results appear to be the first non-trivial bounds on the mixing times of adaptive MCMC samplers, and suggest that the adaptive methods considered may not provide qualitative improvements in mixing over the simpler Markov chain algorithms on which they are based. Our bounds also indicate properties which adaptive MCMC algorithms must have to achieve exponential speed-ups, suggesting directions for further research in these methods.

1. Introduction. Markov chain Monte Carlo (MCMC) sampling techniques are currently the most widely used approach to approximating the high-dimensional integrals arising in Bayesian statistics, as well as in related areas such as statistical mechanics. As such, derivation of new MCMC methods, and formal analysis of their properties, has become an important area of Bayesian statistics research (Andrieu and Roberts, 2009; Douc et al., 2007; Ji and Schmidler, 2009; Jones and Hobert, 2001; Kou et al., 2006; Mengersen and Tweedie, 1996; Mira, 2001; Neal, 2003; Roberts and Rosenthal, 2001; Tierney, 1994).

A common construction for MCMC utilizes a (Metropolis-Hastings) random walk that explores the state space via local moves; however, for some

Keywords and phrases: adaptive Monte Carlo, Markov chain convergence, equi-energy sampler, rapid mixing, tempering

target distributions this random walk takes an impractically long time to explore the target distribution. For example, when the target distribution is multimodal, a local random walk may rarely move between modes. Many algorithms have been introduced to address the challenge of efficient sampling from high-dimensional and multimodal distributions. Parallel tempering (Geyer, 1991) supplements a basic Metropolis-Hastings chain with a set of auxiliary chains, the states of which are occasionally swapped, "seeding" the primary chain with samples from other chains. These auxiliary chains are typically constructed via a temperature parameter, which flattens the target distribution in order to enable crossing of energy barriers (regions of low density), and can allow rapid movement between multiple modes. The related technique of simulated tempering (Geyer and Thompson, 1995; Marinari and Parisi, 1992) uses a single chain with alternating transition kernels.

An alternative is to adapt the transition kernel of the chain, using information obtained from previous iterations to speed convergence - such methods are termed *adaptive* MCMC. The recently proposed equi-energy sampler (Kou et al., 2006), like parallel tempering, constructs auxiliary sampling chains typically constructed by temperature. However, rather than swapping, the equi-energy sampler seeds the primary chain with proposed jumps to locations visited previously by the other chains, specifically those locations having approximately equal energy (density) to the current state. Such jumps potentially enable movement between distinct modes of the target distribution. Two other adaptive algorithms, the importance-resampling MCMC (IR-MCMC) algorithm (Atchadé, 2009*b*) and a method proposed by Gelfand and Sahu (1994), also utilize multiple (non-Markovian) processes which can supplement local moves with jumps to locations previously visited by another process. Again, these methods aim to improve upon the efficiency of a single Markov chain.

Such adaptive algorithms have been shown empirically to have more rapid convergence and more rapid decay of autocorrelation than their non-adaptive counterparts on several examples (Kou et al., 2006; Minary and Levitt, 2006). However, Atchadé (2009*a*) gives an example for which the empirical performance of the equi-energy sampler and IR-MCMC is comparable to that of random-walk Metropolis, and argues that the equi-energy and IR-MCMC samplers are not themselves asymptotically as efficient as their (very efficient) limiting kernels.

Few rigorous bounds on the convergence rates of adaptive MCMC techniques are available. Andrieu and Moulines (2006) and Andrieu and Atchadé (2007) obtain asymptotic efficiency results for a different class of adaptive

MCMC techniques which tune parameters of a parametric transition kernel. Atchadé (2009*b*) considers an adaptive process that at some fixed set of times jumps back to a previously visited location, and shows that if the underlying process converges geometrically then the adaptive process converges at least polynomially (in the number of steps n, not in problem size).

Here we consider the non-asymptotic behavior of such adaptive algorithms, specifically whether they yield convergence ("mixing") times that improve significantly on their non-adaptive counterparts. A major obstacle to obtaining non-asymptotic bounds is the non-Markovian, time-inhomogeneous, irreversible nature of the algorithms, preventing direct application of spectral analysis and other common methods used for Markov chains. Our main result (Theorem 4.1) extends a bound by Woodard et al. (2009*b*) for parallel and simulated tempering to these adaptive methods. For a single process, the bound shows that the mixing time of the adaptive sampler, like that of the underlying Markov chain on which it is based, is limited by the conductance of the Markov kernel (Corollary 4.1). This result holds irrespective of how the non-local jumps are taken and the values of any adaptation tuning parameters. Therefore this type of adaptivity, which we call *multichain resampling*, cannot provide a qualitative speedup from slow to rapid mixing (defined formally in Section 3). This result is not immediately obvious since it might seem advantageous, if the current route of exploration proves unfruitful, to jump back to a more promising location and restart exploration from that point. The same result holds for the multi-process sampler of Gelfand and Sahu (1994), regardless of the number of processes. Combined with results obtained by Woodard et al. (2009*b*), our bounds immediately imply that multichain resampling methods (including the equi-energy sampler, IR-MCMC, and Gelfand-Sahu sampler) mix slowly on two examples considered there: mixtures of normal distributions in $\mathbb{R}^d$, and the mean-field ferromagnetic Potts model.

Our results formalize the intuitive notion that jumping back to locations already visited cannot speed exploration of new, as yet unseen, regions of the target distribution. However, such adaptation may indeed yield improvements in autocorrelation times (and hence *asymptotic* efficiency) relative to their non-adaptive counterparts. Indeed, this is suggested by the empirical results demonstrated in these papers. However, our lower bounds indicate that qualitative improvements in convergence to equilibrium may not be obtainable under the type of adaptivity utilized in these algorithms. Instead, algorithms that encourage exploration of new regions, in addition to speeding mixing among previously visited regions, must be explored. A preliminary step in this direction is given by Heaton and Schmidler (2009).

In Section 2 we define the class of adaptive methods under consideration. Section 3 obtains bounds on the mixing time of these techniques and relates them to existing results on Markov chains. Section 5 shows slow mixing on the two examples, and we conclude with some discussion in Section 7.

2. Adaptive MCMC Techniques. We divide adaptive MCMC techniques considered here into two classes, which capture the majority of methods proposed to date. The first class simulates one or more parallel chains, and for each chain i attempts to adaptively optimize over a family of transition kernels $\{T_\theta : \theta \in \Theta^{(i)}\}$ that are invariant with respect to the target distribution of that chain. We call these methods *invariant adaptive* Markov chain (IAMC) methods. The second class also simulates one or more parallel chains, but sometimes resamples from the history of the chains in order to share information among the chains, or to speed mixing among previously visited regions. The transition kernels of such methods generally are only invariant with respect to the target distribution in a limiting sense. We call these methods *multichain resampling* adaptive Markov chain (MRAM) methods.

To fix notation, let π denote the target distribution of interest on state space $\mathcal{X}$. Let $X^{(1)}, \ldots, X^{(I)}$ be a set of discrete time stochastic processes $X^{(i)} = X_0^{(i)}, X_1^{(i)}, \ldots$ on $\mathcal{X}$, targeted at distributions $\pi^{(i)}$. At least one $X^{(i)}$ is assumed to have $\pi^{(i)} = \pi$; call it $X^{(1)}$.

2.1. *IAMC Methods.* The most familiar approach to adapting MCMC samplers is to optimize the proposal kernel of a Metropolis-Hastings chain. More generally, let $\{T_\theta\}_{\theta \in \Theta^{(i)}}$ be a set of ergodic, $\pi^{(i)}$-reversible Markov transition kernels on $\mathcal{X}$, and denote by $X_{0:n-1}^{(i)}$ the history of the i^{th} process at time n. We consider adaptive sampling algorithms for which the $X^{(i)}$ are generated by respective time-inhomogeneous *but* $\pi^{(i)}$*-invariant* transition kernels $T_{i,n} = T_{\theta_{i,n}}$ where $\theta_{i,n} = g_i(X_{0:n-1}^{(1:I)}) \in \Theta^{(i)}$. We call such algorithms IAMC methods. Here g_i are functions defining the adaptation; IAMC methods are typically constructed to ensure $\theta_{i,n} \stackrel{n\to\infty}{\longrightarrow} \theta^*$ for some optimal value θ^*, but our results do not depend on this property. For concreteness we restrict to the case $\pi^{(i)} \equiv \pi$ and $\Theta^{(i)} \equiv \Theta$ for a common set Θ, which captures all such algorithms proposed to date.

Adaptive Metropolis. The adaptive Metropolis scheme of Haario et al. (2001) was the first of this type to provide formal proof of convergence under continuous adaptation, and helped spark a resurgence of interest in adaptive MCMC methods. The Haario et al. (2001) scheme uses a single chain with π-invariant Metropolis-Hastings kernels T_θ on $\mathcal{X} = \mathbb{R}^d$ constructed from a

multivariate normal random-walk proposal. The adaptive parameter θ is the covariance matrix of the random-walk proposal.

Parallel Chains. Craiu et al. (2009) propose simulating parallel Metropolis-Hastings chains with common invariant distribution π and a common proposal kernel P_θ, adapting the parameters of that kernel using the past samples from all of the chains ("Inter-chain Adaptation").

2.2. *MRAM Methods.* We distinguish a second type of adaptivity proposed for MCMC algorithms, which we refer to as multichain resampling (or MRAM), as follows. We define the MRAM class to include those adaptive sampling algorithms for which the $X^{(i)}$ are generated by respective time-inhomogeneous transition kernels $K_{i,n}$ given by:

$$K_{i,n} = \alpha\, T_i + (1-\alpha) R_{i,n}. \tag{1}$$

for $\alpha \in (0,1]$, where each T_i is an ergodic time-homogeneous $\pi^{(i)}$-reversible Markov transition kernel on $\mathcal{X}$, and $R_{i,n}$ is a sequence of *resampling* Metropolis-type kernels which propose from the set of previously drawn samples $X_{0:n}^{(1:I)}$:

$$Q_{i,n}(x, dy) = \sum_{k=1}^{I} \sum_{j=0}^{n} w_{ijk} \delta(y - X_j^{(k)})$$

where $\sum_{kj} w_{ijk} = 1$ and δ is Dirac's delta, and accept with probability calculated to ensure limiting distribution $\pi^{(i)}$. The resulting sequence of random vectors $X = X_0, X_1, \ldots$ where $X_n = (X_n^{(1)}, \ldots, X_n^{(I)})$ forms a non-Markovian, irreversible, time-inhomogeneous joint stochastic process with limiting marginal distributions $\pi^{(i)}$. Commonly T_i may be a Metropolis-Hastings (MH) kernel using a local random walk proposal; then $R_{i,n}$ supplements these local moves with jumps to potentially distant regions of the state space.

Equi-energy sampler. Of the MRAM methods published to date, the *equi-energy* sampler (EES) of Kou et al. (2006) has perhaps received the most attention. The EES aims to enable moves between points of similar energy (equivalently, density) throughout the state space, potentially allowing the sampler to cross between modes.

Similar to parallel tempering, the EES constructs processes $X^{(i)}$ with tempered target densities $\pi^{(i)} \propto \pi^{\beta_i}$ for a sequence of "inverse temperatures" $1 = \beta_1 > \ldots > \beta_I \geq 0$. (Kou et al. (2006) also truncate the densities $i > 1$ by $\pi^{(i)} \propto \pi^{\beta_i} \wedge c_i$ for some constant $c_i > 0$; this truncation does not alter our results and is omitted here for simplicity.) Each process $X^{(i)}$ is constructed

by specifying T_i to be a $\pi^{(i)}$-reversible MH kernel for some common (across i) proposal P. Adaptivity is obtained by binning the state histories of each process i according to energy; then for $i < I$ the process $X^{(i)}$ occasionally proposes to move to one of the states previously visited by the $i+1$ process $(X_{0:n}^{(i+1)})$ that lie in the same energy bin of π as the current state $X_n^{(i)}$, and accepts with probability calculated to ensure that $\pi^{(i)}$ is the limiting distribution of $X^{(i)}$. (Hence the EES takes $w_{ijk} \propto \delta_{E_{n-1}}(X_j^{(k)})\mathbf{1}_{\{k=i+1\}}$.) Such "equi-energy" moves can be non-local in the state space, potentially involving moves between distinct modes of π.

Importance-resampling MCMC. Two other MRAM methods are proposed by Atchadé (2009*b*). The first is a simplification of EES, using a single process X with non-local moves sampled uniformly from the entire history $X_{0:n}$. (That is, the proposed moves are not restricted to an energy bin corresponding to the current state X_n as done in EES, so Q_n is simply the empirical process $X_{0:n}$.) The second method, referred to as importance-resampling MCMC (IR-MCMC), uses auxiliary chains as in EES, but samples from $X_{0:n}^{(i+1)}$ using weights $w_{ijk} \propto \frac{\pi^{(i)}(X_j^{(k)})}{\pi^{(k)}(X_j^{(k)})}\mathbf{1}_{\{k=i+1\}}$ chosen to be importance weights.

Gelfand-Sahu sampler. Gelfand and Sahu (1994) introduced a MRAM sampler that constructs multiple parallel processes each with limiting distribution π. Each $X^{(i)}$ transitions according to a common Markov kernel T, and occasionally the set of current states are resampled using weights obtained from a kernel density estimate of π based on the entire history of the sampler. This resampling is intended to bring the distribution of the current states closer to π.

3. Mixing Times. The algorithms described in Section 2 construct multiple (non-Markovian) dependent stochastic processes $X^{(1)}, \ldots, X^{(I)}$ on $\mathcal{X}$ often having distinct limiting distributions; denote by X the joint process $X_0, X_1, \ldots$ where $X_n = (X_n^{(1)}, \ldots, X_n^{(I)})$. However, it is convergence of the (marginal) process $X^{(1)}$ with limiting distribution π which is of interest. For $\pi_n = \mathcal{L}_{\pi_0}(X_n^{(1)})$ the marginal distribution of $X_n^{(1)}$ under the joint initial distribution π_0, the total variation norm

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

measures the distance to π, where the supremum is over measurable subsets. Define the *mixing time* τ_ϵ as the number of iterations required to be within

distance ϵ of the target π for any initial distribution π_0:

$$\tau_\epsilon = \sup_{\pi_0} \min\{n : \|\pi_{n'} - \pi\|_{\mathrm{TV}} < \epsilon \quad \forall n' \geq n\}. \tag{2}$$

By analogy to Markov chains (Aldous, 1982; Sinclair, 1992), we say X is *rapidly* mixing if for every fixed ϵ the mixing time grows at most polynomially in the "problem size" (typically the dimension of $\mathcal{X}$). The process is *slowly* mixing if the mixing time grows exponentially in the problem size. The rapid/slow distinction provides a categorization of computational feasibility: while polynomial factors are presumed to eventually be overwhelmed by increases in computing power, exponential factors are presumed to cause persistent computational difficulties. Rapidly mixing processes lead to efficient approximation algorithms for combinatorial counting (Sinclair, 1992) and expectations of bounded variance functions under the target distribution (Schmidler and Woodard, 2009). In unbounded state spaces the inf over π_0 may lead to $\tau_\epsilon = \infty$; in such cases it is desirable to assume $\sup \frac{\pi_0(x)}{\pi(x)}$ is bounded, e.g. by restriction to some compact set.

Many of the standard techniques for bounding mixing times of Markov chains are not immediately applicable to the adaptive processes of Section 2, which under the general construction produce non-Markovian, time-inhomogeneous, irreversible stochastic processes. However, we will obtain lower bounds on mixing times via the *hitting time* for subsets $A \subset \mathcal{X}$:

$$H_A = \min_i H_A^{(i)} \qquad\qquad H_A^{(i)} = \min\{n : X_n^{(i)} \in A\}$$

and involving the familiar *conductance* of a π-reversible Markov kernel T:

$$\Phi_T = \inf_{\substack{A \subset \mathcal{X}: \\ 0 < \pi(A) < 1}} \Phi_T(A) \qquad\qquad \Phi_T(A) = \frac{\int_A \pi(dv) T(v, A^c)}{\pi(A)\pi(A^c)}$$

where $\Phi_T(A)$ captures the probability of moving between A and A^c under T, and Φ_T quantifies the worst "bottleneck" in the transition kernel.

For any $A \subset \mathcal{X}$ with $\pi^{(i)}(A) > 0$, denote the restriction of π to A by $\pi|_A(dy) \propto \pi(dy)\mathbf{1}(y \in A)$. Then we define the restriction $Y = X|_A$ of the process X to A by taking $Y_0 = X_0$ (if $X_0 \in A^I$) and defining Y identically to X, except that any move leaving A is rejected.

Convergence of estimators. Some authors have questioned the relevance of L_1 convergence to MCMC (Mira and Geyer, 2000) , where interest lies in convergence of ergodic averages $\hat{\theta}_n = n^{-1} \sum_{i=1}^n \theta(X_n)$, arguing for restricting attention to asymptotic variance (Flegal, 2008; Mira, 2001). When negative

eigenvalues are present the former can be slow even when the latter is small. However, for finite-length MCMC runs the relevant quantity is the expected mean-squared error:

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Bias}^2(\hat{\theta}) + \mathrm{Var}(\hat{\theta})$$

and focusing on the integrated autocorrelation considers only the second term. Convergence of the Markov chain to its stationary distribution appears in the bias term; this is the formal justification for the standard practice of discarding an initial transient ("burn-in") period.

Although our bounds are stated in terms of L_1 convergence, our proofs use hitting times and thus immediately imply bounds on the MSE convergence of the ergodic averages as well. To see this, define

$$\|\hat{\theta}_n - \theta\|_{\mathrm{MSE}} = \sup_{\theta \in L_2(\pi);\, \|\theta\| \leq 1} \mathrm{MSE}\ \hat{\theta}_n$$

and let $\Pr(\tau_A \leq n) \leq \epsilon$ for some $A \subset \mathcal{X}$. Then taking $\theta = 1_A(x)$ gives $\mathrm{Bias}^2(\hat{\theta}_n) \geq (\pi(A) - \epsilon)^2$. Thus negative autocorrelation *can* hurt convergence of estimators based on ergodic averages, if it arises from multimodality of the target distribution. In this case the bias term may dominate and cannot be ignored.

4. Bounds for MRAM Processes. We first obtain bounds for MRAM samplers. Since the Markov kernels T_i on which they are based do not depend on the history of the chain, we are able to obtain very general results.

Adaptive processes are not in general invariant with respect to their target distributions. For example, in the EES algorithm it is easily seen that the acceptance ratio for resampling moves

$$\rho(x, y) = \min\left\{1, \frac{\pi^{(i-1)}(dy)\pi^{(i)}(dx)}{\pi^{(i-1)}(dx)\pi^{(i)}(dy)}\right\}$$

leaves $\pi^{(i-1)}$ invariant only if the current and proposed states are drawn *independently*; but the resampling process makes the chains dependent (e.g. inflating $\Pr(X_n^{(i)} = X_n^{(i+1)})$). Thus even when initialized according to the target distribution π, the EES process wanders away from π before returning in the limit.

This drift is not desirable; by contrast, Markov chain methods monotonically approach their limiting distribution. The parameter α controls the amount of drift; as $\alpha \to 1$ the number of T_i moves increases relative to the number of $R_{i,n}$ moves. T_i moves reduce the (L_2) distance to $\pi^{(i)}$ by at least

a factor equal to the spectral gap of T_i, while $R_{i,n}$ moves can inflate this distance. For α relatively large the drift should be minimal; in order to analyze adaptive methods in the presence of this drift, we assume that it is bounded as follows.

ASSUMPTION 4.1. *There exists a constant $1 \leq c < \infty$ (independent of problem size) such that for any $A \subset \mathcal{X}$ having $\pi^{(i)}(A) > 0$ for all i, the sampler $Y = X|_A$ with $Y_0^{(i)} \overset{ind.}{\sim} \pi^{(i)}|_A$ satisfies the following for all i and n: the marginal distribution $\mathcal{L}(Y_n^{(i)})$ has a density with respect to $\pi^{(i)}|_A$ that is everywhere $\leq c$.*

This holds with $c = 1$ for the single-chain method of Atchadé (2009*b*), and for the degenerate case $\alpha = 1$.

The bound we obtain for MRAM algorithms is a generalization of a mixing time bound for parallel tempering given by Woodard et al. (2009*b*). Define the *persistence* for any $A \subset \mathcal{X}$ and any $i \in \{1, \ldots, I\}$ as:

$$\gamma(A, i) = \min\left\{1, \frac{\pi^{(i)}(A)}{\pi(A)}\right\}.$$

Then the following bound for parallel tempering follows directly from the spectral gap bounds obtained in Woodard et al. (2009*b*):

THEOREM A. *(Woodard et al., 2009*b*)*
For $\mathcal{X}$ finite, $\epsilon > 0$, and any $A \subset \mathcal{X}$ with $0 < \pi^{(i)}(A) < 1$ for all i, the mixing time τ_ϵ^* of parallel tempering satisfies

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

Now let X be a MRAM process on general $\mathcal{X}$ as defined in Section 2, satisfying Assumption 4.1. We have the following result:

THEOREM 4.1. *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 MRAM process satisfies:*

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

PROOF. Let $X_0^{(i)} \overset{\text{ind.}}{\sim} \pi^{(i)}|_{A^c}$ and let $Y = X|_{A^c}$, so that $\psi_{i,n} = \mathcal{L}(Y_n^{(i)})$ has a density with respect to $\pi^{(i)}|_{A^c}$ that is everywhere $\leq c$. Define the sequences

$Z^{(i)}$ of Boolean random variables, where $Z_n^{(i)}$ is true if a move of $Y^{(i)}$ at time n is rejected because it would leave A^c, and false otherwise.

Consider the hitting time H_A for X. Since $H_A^{(1)} \geq H_A$, for any n such that $\Pr(H_A \leq n) \leq \pi(A) - \epsilon$ we have $\|\pi_n - \pi\|_{\text{TV}} \geq \epsilon$ and so $\tau_\epsilon > n$. The probability that $H_A = n$ is equal to the probability that Y first attempts a move to A at time n but rejects due to restriction, so

$$\begin{aligned}
\Pr(H_A \leq n) \leq \sum_{i=1}^{I}\sum_{j=1}^{n} \Pr(Z_j^{(i)}) &\leq \sum_{i=1}^{I}\sum_{j=1}^{n} \int_{A^c} T_i(y, A)\psi_{i,j-1}(dy) \\
&\leq c\sum_{i=1}^{I}\sum_{j=1}^{n} \int_{A^c} T_i(y, A)\pi^{(i)}|_{A^c}(dy) \\
&= cn\sum_{i=1}^{I} \pi^{(i)}(A)\Phi_{T_i}(A)
\end{aligned}$$

where the second inequality comes from the mixture representation of (1), since for the sampler Y we have $Q_{i,j}(y, A) = 0$ for all y, i, and j. The last equality uses reversibility of T_i. Now define $n_\epsilon(A) = \min\{n : \Pr(H_A \leq n) > \pi(A) - \epsilon\}$, so that

$$\begin{aligned}
\tau_\epsilon \geq n_\epsilon(A) &\geq (\pi(A) - \epsilon)\left[c\sum_i \pi^{(i)}(A)\Phi_{T_i}(A)\right]^{-1} \\
&\geq (\pi(A) - \epsilon)\left[cI\max_i \pi^{(i)}(A)\Phi_{T_i}(A)\right]^{-1}.
\end{aligned}$$

Then $\pi^{(i)}(A) \leq \gamma(A, i)$ gives the desired result. □

The factor of I appearing in Theorem 4.1 but not Theorem A comes from the slightly different definitions of mixing in the two cases: the parallel tempering mixing time is for convergence of the joint chain process to its limiting product distribution, required for the spectral analysis of Woodard et al. (2009*b*).

The difference in the dependence on ϵ between the two theorems comes from our use of hitting times to bound variation distance directly for the time-inhomogeneous MRAM processes, compared to standard time-change arguments for time-homogeneous processes. We suspect this can be improved; the bound in Theorem 4.1 is certainly loose as a function of ϵ for some MRAM processes, since parallel tempering is trivially in this set. However, we can use Theorem 4.1 to analyze the effect of problem size on mixing time for fixed ϵ (as in Section 5).

4.1. *Single Chain Resampling.* Consider a MRAM process with a single chain ($I = 1$), i.e. an adaptive process constructed from a Markov kernel T mixed with occasional jumps back to previously visited locations. These jumps can be made in any manner, subject to Assumption 4.1. In this case, Theorem 4.1 simplifies:

COROLLARY 4.1. *For any $0 < \epsilon < 1/4$, the mixing time τ_ϵ of a MRAM sampler based on T, with $I = 1$, satisfies:*

$$\tau_\epsilon \geq \frac{1}{4c\Phi_T}.$$

PROOF. For measurable $A \subset \mathcal{X}$ such that $1/2 \leq \pi(A) < 1$, Theorem 4.1 gives $\tau_\epsilon \geq (\pi(A) - \epsilon)[c\Phi_T(A)]^{-1}$, and the result follows from $\epsilon < 1/4$ and $\Phi_T(A) = \Phi_T(A^c)$. □

We can compare this result with standard results for Markov chains. For the case of $\mathcal{X}$ finite, if we assume that $T(x, x) \geq 3/4$ for all $x \in \mathcal{X}$ (which can be achieved by simply adding a holding probability of $3/4$), results in Sinclair (1992) give the following bounds on the mixing time τ_ϵ^* of the Markov chain T

$$\frac{1}{8\Phi_T} \ln(2\epsilon)^{-1} \leq \tau_\epsilon^* \leq \frac{8}{\Phi_T^2} \left[\ln(\max_x \pi(x)^{-1}) + \ln(\epsilon^{-1}) \right]. \tag{3}$$

The lower bounds in (3) and Corollary 4.1 on the mixing times τ_ϵ^* of the Markov chain and τ_ϵ of the adaptive sampler are of the same order as a function of Φ_T. Combining with results in Lawler and Sokal (1988) we have:

COROLLARY 4.2. *For general $\mathcal{X}$ and T geometrically ergodic, if the spectral gap of T decreases exponentially in the problem size then any MRAM process based on T with $I = 1$ is slowly mixing.*

COROLLARY 4.3. *For finite $\mathcal{X}$, if $\ln(\max_x \pi(x)^{-1})$ grows polynomially as a function of the problem size, then slow mixing of the Markov chain with transition kernel T implies slow mixing of any MRAM process based on T that has $I = 1$.*

In particular, Corollary 4.2 proves for the first time the hypothesis of Atchadé (2009*b*) that the single-chain sampler defined in that paper is never qualitatively more efficient than the Markov chain on which it is based.

The condition on $\max_x \pi(x)^{-1}$ means that the smallest probability $\pi(x)$ can decrease exponentially in the problem size, but not, e.g., doubly-

exponentially, and comes from the consideration of worst-case (over initial distributions) mixing time. This condition is satisfied by the mean-field Potts model example of Section 5. When it does not hold for a particular example, it is often possible to remove the low-probability states from the state space without significantly altering either the mixing time of the sampler or the Monte Carlo estimates.

4.2. *Gelfand-Sahu Sampler.* The sampler proposed by Gelfand and Sahu (1994) and described in Section 2 is constructed from multiple processes with common Markov kernel T and common limiting density $\pi^{(i)} \equiv \pi$. Then Theorem 4.1 simplifies:

COROLLARY 4.4. *For any $0 < \epsilon < 1/4$, the mixing time τ_ϵ of the Gelfand-Sahu sampler based on T satisfies:*

$$\tau_\epsilon \geq \frac{1}{4cI\Phi_T}.$$

Combining this result with (3) we find that:

COROLLARY 4.5. *For general $\mathcal{X}$ and T geometrically ergodic, if the spectral gap of T decreases exponentially in the problem size then any Gelfand-Sahu sampler based on T is slowly mixing.*

COROLLARY 4.6. *For finite $\mathcal{X}$, if $\ln(\max_x \pi(x)^{-1})$ grows polynomially as a function of the problem size then slow mixing of the Markov chain with transition kernel T implies slow mixing of any Gelfand-Sahu sampler based on T.*

Note that we discount the possibility of obtaining rapid mixing when the number of processes I grows exponentially in the problem size, since this case automatically requires exponential computational effort.

Therefore the mixing time of the Gelfand-Sahu sampler is limited by the conductance of the Markov kernel, and it cannot be rapidly mixing unless the underlying Markov chain is already rapidly mixing.

5. Examples of Slow Mixing.

5.1. *MRAM Samplers on a Mixture of Normals.* Consider sampling from a target distribution given by a mixture of two multivariate normal distributions in $\mathbb{R}^d$, with density:

$$\pi(x) = \frac{1}{2} N_d(x; -\mu \mathbf{1}_d, \sigma_1^2 \mathrm{I}_d) + \frac{1}{2} N_d(x; \mu \mathbf{1}_d, \sigma_2^2 \mathrm{I}_d) \tag{4}$$

where $N_d(x;\nu,\Sigma)$ denotes the multivariate normal density for $x \in \mathbb{R}^d$ with mean vector ν and $d\times d$ covariance matrix Σ, and $\mathbf{1}_d$ and I_d denote the vector of d ones and the $d \times d$ identity matrix, respectively. This can be expected to reasonably approximate many multimodal posterior distributions arising in Bayesian statistics.

Restrict to any convex $K \subset \mathbb{R}^d$ such that $\pi(K) \stackrel{d\to\infty}{\longrightarrow} 1$ and such that $\ln(\sup_{x\in K} \pi(x)^{-1})$ increases polynomially in d; it is under such restricted conditions that Frieze et al. (1994) show rapid mixing of Metropolis-Hastings with local proposals on log-concave target densities in $\mathbb{R}^d$. (The unrestricted case leads to $\tau_\epsilon = \infty$ due to the presence of starting states arbitrarily far from the modes.)

Let S be the proposal kernel that is uniform on the ball of radius d^{-1} centered at the current state. When $\sigma_1 = \sigma_2$, Woodard et al. (2009*a*) have given an explicit construction of parallel and simulated tempering chains that is rapidly mixing. However, when $\sigma_1 > \sigma_2$, Woodard et al. (2009*b*) set $A = \{x \in \mathbb{R}^d : \sum_i x_i \geq 0\}$ and show that if the target distributions $\pi^{(i)}$ are tempered versions of π, then $\max_i \gamma(A,i)\Phi_{T_i}(A)$ is exponentially decreasing for any choice of I temperatures whenever I is polynomial, and that consequently parallel tempering is slowly mixing. Since $\pi(A) \geq 1/2$ for all d large enough, it follows immediately from Theorem 4.1 that

COROLLARY 5.1. *Any MRAM process satisfying Assumption 4.1 that is based on the proposal S and uses tempered densities is slowly mixing on the normal mixture (4) for $\sigma_1 \neq \sigma_2$.*

5.2. *MRAM Samplers on the Mean-Field Potts Model.* Potts models are Gibbs random fields defined on graphs, which arise in statistical physics (Binder and Heermann, 2002), image processing (Geman and Geman, 1984), and spatial statistics (Green and Richardson, 2002). The mean-field Potts model is the special case of a complete interaction graph, which admits simpler analysis but nonetheless retains the important characteristics of general Potts models, namely a first-order phase transition at a critical temperature (for $q \geq 3$). A mean-field Potts model with M sites has distribution on $z \in \mathbb{Z}_q^M$ given by:

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

and we will be concerned with the "ferromagnetic" case $\alpha \geq 0$. We consider the standard single-site (Glauber) dynamics as the base Metropolis kernel,

which proposes changing the color of a single site chosen uniformly at random at each time. The convergence rate of single-site dynamics on Potts models exhibits a phase transition, slowing down dramatically at a critical value α_c of the interaction parameter. For the mean-field ferromagnetic Potts ($q \geq 3$) model with $\alpha \geq \alpha_c$, the Metropolis chain is slowly mixing, as is the Swendsen-Wang algorithm (Gore and Jerrum, 1999) and parallel and simulated tempering (Bhatnagar and Randall, 2004). From Theorem 4.1, we have the following:

COROLLARY 5.2. *Any MRAM process satisfying Assumption 4.1 based on single-site dynamics and using tempered densities is slowly mixing in M for the mean-field Potts model with $\alpha > \alpha_c$.*

We suspect Lemma 5.2 to hold for $\alpha = \alpha_c$, but this cannot be proven using Theorem 4.1 due to the term $(\pi(A) - \epsilon)$.

PROOF. Define the subset $A = \left\{ z : \sum_i \mathbf{1}(z_i = 1) > \frac{M}{2} \right\}$ of the Potts model state space. Woodard et al. (2009*b*) show that for $\alpha \geq \alpha_c$ and any choice of a polynomial number I of temperatures, the quantity $\max_i \{\gamma(A, i)\Phi_{T_i}(A)\}$ decreases exponentially as a function of M. We show that for $\alpha > \alpha_c$ and all M large enough, $\pi(A) > b$ for some positive constant b (see Appendix B). It then follows immediately from Theorem 4.1 that the mixing time of the MRAM sampler increases exponentially in M for any fixed $\epsilon \in (0, b)$, i.e. the sampler is slowly mixing. □

6. Bounds for IAMC Processes. While in MRAM methods the transition kernel is a mixture of a fixed transition kernel T_i and a resampling kernel, in IAMC samplers the parameters θ of the transition kernel T_θ depend on the entire history of the sampler. This makes it harder to obtain general bounds on the mixing time of IAMC algorithms. Instead we show how to obtain bounds for two IAMC methods on the example (4). Our proof technique bounds the hitting time of a set A that has low conductance $\Phi_{T_\theta}(A)$ for "most" θ. We expect that this approach can be used to obtain lower bounds on mixing time for other examples and other IAMC techniques.

Subject to Assumption 4.1, we have:

THEOREM 6.1. *The Adaptive Metropolis method of Haario et al. (2001) and the Inter-chain Adaptation method of Craiu et al. (2009) are slowly mixing in d for the mixture of normals (4) with any fixed values of μ, σ_1, and σ_2 such that $\sigma_1 > \sigma_2$, $\mu > 2\sigma_1$ and $\sigma_1/\sigma_2 < \sqrt{e}$.*

We expect that the result in fact holds for any fixed values of μ, σ_1, and σ_2.

PROOF. Take any $\delta \in (\exp\{-1/4\}, 1)$ and define the sets:

$$B_1 = \{x \in \mathbb{R}^d : |x + \mu \mathbf{1}_d| \leq \sigma_1 \sqrt{2d}\}$$
$$B_2 = \{x \in \mathbb{R}^d : |x - \mu \mathbf{1}_d| \leq 2\sigma_2 \sqrt{d}\}$$
$$A = \left\{ x \in \mathbb{R}^d : \frac{N_d(x; \mu \mathbf{1}_d, \sigma_2^2 \mathrm{I}_d)}{N_d(x; -\mu \mathbf{1}_d, \sigma_1^2 \mathrm{I}_d)} \leq \delta^d \right\}. \tag{5}$$

B_1 and B_2 are hyperspheres, each centered at one of the modes of π. As we will see, π concentrates in B_1 and B_2 as $d \to \infty$, and the Adaptive Metropolis and Inter-chain Adaptation algorithms have increasing difficulty moving between B_1 and B_2, causing slow mixing. We have $B_1 \subset A$ and $B_2 \subset A^c$ (Proposition A.1 in Appendix A).

Initialize $X_0^{(i)} \overset{\text{ind.}}{\sim} N_d(-\mu \mathbf{1}_d, \sigma_1^2 \mathrm{I}_d)$, and recall that H_{A^c} is the hitting time of A^c. As in the proof of Theorem 4.1, define $n_\epsilon(A^c) = \min\{n : \Pr(H_{A^c} \leq n) > \pi(A^c) - \epsilon\}$ and observe that $\tau_\epsilon \geq n_\epsilon(A^c)$.

Prop. A.3 in Appendix A constructs a coupling to show $\exists \beta < 1$ such that $\Pr(H_{A^c} \leq n) \leq n\beta^d$. Since $\pi(A^c) \geq 1/3$ (Prop. A.2 in Appendix A), for any $\epsilon < 1/6$ we have that $\tau_\epsilon \geq 1/(6\beta^d)$, which grows exponentially in d; this proves Theorem 6.1.

□

Theorem 6.1 says that these IAMC samplers do not qualitatively improve the convergence rate over their simpler, non-adaptive counterparts. Instead, in multimodal target distributions the chain adapts to the *local* shape of the distribution, and may actually *prevent* it from exploring more globally, decreasing the rate of convergence. (See Heaton and Schmidler (2009) for an empirical demonstration of this behavior.)

7. Conclusions. These results appear to be the first non-asymptotic bounds on convergence for adaptive MCMC samplers. Our results for IAMC samplers show that commonly used adaptive schemes can perform no better, and may perform worse, than their non-adaptive counterparts on multimodal target distributions. We then use this to show that current methods can converge exponentially slowly on simple multimodal target distributions, suggesting that some caution is needed in applying these methods.

Our results for the MRAM class formalize the intuitive notion that jumping back to locations already visited cannot speed exploration of unseen regions of the target distribution (convergence rate), although it may improve mixing among previously visited regions (autocorrelation). Thus for

the multimodal problems where sophisticated MCMC methods are most needed, the adaptive MRAM methods are slowly mixing when the underlying non-adaptive chain is, and so do not provide a qualitative improvement over simpler methods. Our lower bounds indicate that qualitative improvements in convergence to equilibrium may not be obtainable under the type of adaptivity utilized in MRAM algorithms, emphasizing the need to develop algorithms that encourage exploration of new regions in addition to speeding mixing among previously visited regions.Thus an adaptive sampling algorithm must achieve *both* of two criteria: it must (i) adapt to mix efficiently among previously visited regions, and (ii) adapt to encourage exploration of unseen regions. Trading off these desiderata will require further exploration, and may be thought of as a standard bandit (exploration/exploitation) type problem. As one approach, we suggest that a guiding principle for designing adaptive algorithms may be to use a mixture kernel of the form:

$$K_{\text{adapt}} = \alpha K_{\text{AMIS}} + (1 - \alpha) K_{\text{explore}}$$

where one component adapts to the previously seen samples and the other uses methods to encourage moving away from previous samples. Examples of the latter have received significant interest in recent years especially in statistical physics (Wang and Landau, 2001), and have recently been introduced in statistics (Liu et al., 2001); other examples include Heaton and Schmidler (2009); Liang and Wong (1999). A preliminary step in this direction is given by Heaton and Schmidler (2009), but this seems a fruitful area for further research.

Acknowledgments. We thank Jeff Rosenthal for pointing out an error in a previous statement of Theorem 6.1.

APPENDIX A: RESULTS FOR PROOF OF THEOREM 6.1

PROPOSITION A.1. *$B_1 \subset A$ and $B_2 \subset A^c$.*

PROOF. For $x \in B_1$ we have by the triangle inequality:

$$|x - \mu \mathbf{1}_d| \geq |\mu \mathbf{1}_d + \mu \mathbf{1}_d| - |x + \mu \mathbf{1}_d| \geq 2\mu\sqrt{d} - \sigma_1\sqrt{2d} > \mu\sqrt{d}$$

and analogously $|x + \mu \mathbf{1}_d| \geq \mu\sqrt{d}$ for any $x \in B_2$.

Therefore for $x \in B_1$,

$$\begin{aligned}
\frac{N_d(x; \mu\mathbf{1}_d, \sigma_2^2 \mathrm{I}_d)}{N_d(x; -\mu\mathbf{1}_d, \sigma_1^2 \mathrm{I}_d)} &= \left(\frac{\sigma_1}{\sigma_2}\right)^d \exp\left\{-\frac{1}{2\sigma_2^2}|x - \mu\mathbf{1}_d|^2 + \frac{1}{2\sigma_1^2}|x + \mu\mathbf{1}_d|^2\right\} \\
&\leq \exp\left\{-\frac{1}{2\sigma_2^2}|x - \mu\mathbf{1}_d|^2 + \frac{1}{2\sigma_1^2}|x + \mu\mathbf{1}_d|^2 + \frac{d}{2}\right\} \\
&\leq \exp\left\{-\frac{1}{2\sigma_2^2}|x - \mu\mathbf{1}_d|^2 + \frac{3d}{2}\right\} \\
&\leq \exp\left\{-\frac{\mu^2 d}{2\sigma_2^2} + \frac{3d}{2}\right\} \leq \exp\left\{-2d + \frac{3d}{2}\right\} \\
&= \exp\left\{-\frac{d}{2}\right\} < \delta^d.
\end{aligned}$$

For $x \in B_2$,

$$\begin{aligned}
\frac{1}{\sigma_2^d} \exp\left\{-|x - \mu\mathbf{1}_d|^2/(2\sigma_2^2)\right\} &\geq \frac{1}{\sigma_2^d} \exp\{-2d\} \\
&\geq \frac{1}{\sigma_1^d} \exp\{-2d\} > \frac{1}{\sigma_1^d} \exp\{-\mu^2 d/(2\sigma_1^2)\} \\
&> \frac{1}{\sigma_1^d} \exp\{-|x + \mu\mathbf{1}_d|^2/(2\sigma_1^2)\}
\end{aligned}$$

since $|x + \mu\mathbf{1}_d| \geq \mu\sqrt{d}$. Therefore $N_d(x; \mu\mathbf{1}_d, \sigma_2^2 \mathrm{I}_d) > N_d(x; -\mu\mathbf{1}_d, \sigma_1^2 \mathrm{I}_d)$. □

PROPOSITION A.2. *$\pi(A^c) \geq 1/3$.*

PROOF. For $Z \sim N_d(\mu\mathbf{1}_d, \sigma_2^2 \mathrm{I}_d)$, we have $|Z - \mu\mathbf{1}_d|^2/\sigma_2^2 \sim \chi_d^2$. Using a normal approximation we find that $\Pr(|Z - \mu\mathbf{1}_d|^2/\sigma_2^2 > 4d)$ decreases exponentially in d. So for all d large enough, $N_d(\mu\mathbf{1}_d, \sigma_2^2 \mathrm{I}_d)(B_2) > 2/3$, and thus $\pi(A^c) \geq \pi(B_2) \geq 1/3$. □

PROPOSITION A.3. *There is some $\beta < 1$ such that $Pr(H_{A^c} \leq n) \leq n\beta^d$ for all d large enough and all n.*

To prove Proposition A.3 we will need several intermediate results.

A.1. Auxiliary results for proof of Proposition A.3.

PROPOSITION A.4. *Define*

$$\rho_W(x,y) = 1 \wedge \frac{N_d(y; -\mu\mathbf{1}_d, \sigma_1^2 I_d)}{N_d(x; -\mu\mathbf{1}_d, \sigma_1^2 I_d)} \qquad \rho_X(x,y) = 1 \wedge \frac{\pi(y)}{\pi(x)}.$$

Then there is some $\alpha < 1$ for which

$$\rho_W(x,y), \rho_X(x,y) \leq 2\alpha^d \qquad \forall x \in B_1, y \in A^c \backslash B_2.$$

PROOF. Define $C = \{x : |x + \mu\mathbf{1}_d| \leq \sigma_1\sqrt{5d/2}\}$; the proof of Prop. A.1 also gives $C \subset A$. So $\exists\tilde{\alpha} < 1$ such that for $x \in B_1$ and $y \in A^c$,

$$N_d(y; -\mu\mathbf{1}_d, \sigma_1^2 \mathrm{I}_d)/N_d(x; -\mu\mathbf{1}_d, \sigma_1^2 \mathrm{I}_d) \leq \tilde{\alpha}^d,$$

since $|x + \mu\mathbf{1}_d| \leq \sigma_1\sqrt{2d}$ and $|y + \mu\mathbf{1}_d| > \sigma_1\sqrt{5d/2}$. Similarly, $\exists\hat{\alpha} < 1$ such that for $x \in B_1$ and $y \in A^c\backslash B_2$,

$$N_d(y; \mu\mathbf{1}_d, \sigma_2^2 \mathrm{I}_d)/N_d(x; -\mu\mathbf{1}_d, \sigma_1^2 \mathrm{I}_d) \leq \hat{\alpha}^d.$$

For $x \in B_1$, $\pi(x)$ is within a factor of two of $N_d(x; -\mu\mathbf{1}_d, \sigma_1^2 \mathrm{I}_d)$; the result follows where $\alpha = \max\{\tilde{\alpha}, \hat{\alpha}\}$. □

PROPOSITION A.5. *Consider an Adaptive Metropolis chain $W = W_0, W_1, \ldots$ with $W_0 \sim N_d(-\mu\mathbf{1}_d, \sigma_1^2 I_d)$ and $W_n \sim T_n^W(W_{n-1}, \cdot)$ where T_n^W is the Metropolis transition kernel with target $N_d(-\mu\mathbf{1}_d, \sigma_1^2 I_d)$ and normal proposal having covariance $\theta_n^W = g(W_{0:n-1})$, and g is the adaptation function for Adaptive Metropolis. Then there is some $\rho < 1$ such that for all d large enough and all n, $Pr(W_n \in B_1^c) \leq \rho^d$.*

PROOF. For a random variable $Z \sim N_d(-\mu\mathbf{1}_d, \sigma_1^2 \mathrm{I}_d)$, we have $|Z + \mu\mathbf{1}_d|^2/\sigma_1^2 \sim \chi_d^2$. Using a normal approximation we find that

$$\Pr(|Z + \mu\mathbf{1}_d|^2/\sigma_1^2 > 2d) = N_d(-\mu\mathbf{1}_d, \sigma_1^2 \mathrm{I}_d)(B_1^c)$$

decreases exponentially in d. By Assumption 4.1, $\mathcal{L}(W_n)$ has a density with respect to $N_d(-\mu\mathbf{1}_d, \sigma_1^2 \mathrm{I}_d)$ that is everywhere $\leq c$, giving the result. □

PROPOSITION A.6. *Consider the chain W from Prop. A.5. There is some $\gamma < 1$ such that the marginal probability that W proposes a move to B_2 at time n is $\leq \gamma^d$ for all d large enough and all n. I.e., letting $P_n^W(W_{n-1}, \cdot)$ be the proposal kernel,*

$$\int P_n^W(W_{n-1}, B_2)\mathcal{L}(W_{0:n-1}) \leq \gamma^d.$$

PROOF. The distribution of $W_{0:n-1}$ is symmetric with respect to rotations around the point $-\mu\mathbf{1}_d$, as is the distribution of proposed state W^*. B_2 is a hypersphere and $-\mu\mathbf{1}_d \notin B_2$; consider the infinite (circular) cone with apex $-\mu\mathbf{1}_d$ that contains B_2 and has minimal aperature angle b. A simple geometric argument shows $b < \pi$ radians and b does not depend on d. The number of non-overlapping cones with aperature angle b and apex $-\mu\mathbf{1}_d$ increases exponentially in d, and they have equal proposal probabilities by symmetry. □

A.2. Proof of Proposition A.3. Here we give the proof for the Adaptive Metropolis algorithm; the case of the Inter-chain Adaptation algorithm is nearly identical but notationally more cumbersome. Our proof technique is inspired by that used in Roberts and Rosenthal (2007), Theorem 1.

Let $T_\theta^X(x,\cdot)$ be the Metropolis kernel with proposal $N_d(x,\theta)$ and target π defined in (4), and let $T_\theta^W(x,\cdot)$ be the Metropolis kernel with proposal $N_d(x,\theta)$ and target $N_d(-\mu\mathbf{1}_d, \sigma_1^2\mathrm{I}_d)$. For a chain W let $\theta_n^W = g(W_{0:n-1})$, where g is the adaptation function for Adaptive Metropolis.

Let $\delta, \alpha, \rho, \gamma < 1$ be as defined as in (5) and Propositions A.4, A.5, and A.6, respectively. We claim ("Claim A") that for all d large enough, we can construct a stochastic process $W_0, W_1, \ldots$ such that $W_0 = X_0$ and, for all n,

1. $X_n \sim T_n^X(X_{n-1},\cdot)$, where $T_n^X = T_{\theta_n^X}^X$
2. $W_n \sim T_n^W(W_{n-1},\cdot)$, where $T_n^W = T_{\theta_n^W}^W$
3. $\Pr(W_j = X_j \text{ for } 0 \le j \le n) \ge 1 - n(2\rho^d + 4\gamma^d + \delta^d + 2\alpha^d)$.

Claim A is trivially true for $n = 0$. Suppose that it is true for some value $n-1$. In this case Prop. A.7 (below) shows that, conditional on the event (E) that $W_j = X_j$ for $0 \le j \le n-1$ and $W_{n-1} \in B_1$,

$$\|T_n^W(W_{n-1},\cdot)|_{B_2^c} - T_n^X(X_{n-1},\cdot)|_{B_2^c}\|_{TV} \le \delta^d + 2\alpha^d.$$

So on E the conditional distributions of W_n and X_n restricted to B_2^c are within $\delta^d + 2\alpha^d$ of each other. We show below that the probability on E of proposing a move to B_2 is $\le 4\gamma^d$. Therefore by Roberts and Rosenthal (2004, Proposition 3(g)), on E we can ensure that $W_n = X_n$ with probability $\ge 1 - (4\gamma^d + \delta^d + 2\alpha^d)$.

Recalling that $W_0 = X_0 \sim N_d(-\mu\mathbf{1}_d, \sigma_1^2\mathrm{I}_d)$, by Prop. A.5 the marginal probability $\Pr(W_{n-1} \in B_1^c)$ is $\le \rho^d$. Therefore for all d large enough,

$$\Pr(W_{n-1} \in B_1^c | X_j = W_j \text{ for } 0 \le j \le n-1) \le 2\rho^d.$$

By Prop. A.6, the marginal probability that the W chain proposes a move to B_2 at time n is $\le \gamma^d$. Then, letting W^* be the proposed state, for all d

large enough

$$\Pr(W^* \in B_2 | W_{n-1} \in B_1 \text{ and } X_j = W_j \text{ for } 0 \le j \le n-1) \le 2\gamma^d.$$

Notice that the distribution of the proposal X^* in chain X is the same as that of W^*, conditional on $W_{n-1} \in B_1$ and $X_j = W_j$ for $0 \le j \le n-1$. So

$$\Pr(X^* \in B_2 | W_{n-1} \in B_1 \text{ and } X_j = W_j \text{ for } 0 \le j \le n-1) \le 2\gamma^d.$$

Now we have:

$$\begin{aligned}
&\Pr[W_j = X_j \text{ for } 0 \le j \le n] \\
&\ge \Pr[W_j = X_j \text{ for } 0 \le j \le n-1] \left(1 - 2\rho^d\right) \left(1 - (4\gamma^d + \delta^d + 2\alpha^d)\right) \\
&\ge \left(1 - (n-1)(2\rho^d + 4\gamma^d + \delta^d + 2\alpha^d)\right) \left(1 - (2\rho^d + 4\gamma^d + \delta^d + 2\alpha^d)\right) \\
&\ge 1 - n(2\rho^d + 4\gamma^d + \delta^d + 2\alpha^d).
\end{aligned}$$

Hence Claim A is proven by induction. Therefore (using Prop. A.5)

$$\begin{aligned}
\Pr(H_{A^c} \le n) &\le \Pr(\exists j \le n : W_j \in A^c) + \Pr(\exists j \le n : X_j \ne W_j) \\
&\le n(3\rho^d + 4\gamma^d + \delta^d + 2\alpha^d).
\end{aligned}$$

□

PROPOSITION A.7. *Using the definitions in the proof of Prop. A.3, conditional on the event that $W_j = X_j$ for $0 \le j \le n-1$ and $W_{n-1} \in B_1$,*

$$\|T_n^W(W_{n-1}, \cdot)|_{B_2^c} - T_n^X(X_{n-1}, \cdot)|_{B_2^c}\|_{TV} \le \delta^d + 2\alpha^d.$$

PROOF. If $W_j = X_j$ for all $j \le n-1$ then $\theta_n^W = \theta_n^X$.
For $x \in B_1$ and $y \in A$,

$$\frac{\pi(y)}{\pi(x)} \in \left[\frac{1}{1+\delta^d}, 1+\delta^d\right] \frac{N_d(y; -\mu\mathbf{1}_d, \sigma_1^2 \mathrm{I}_d)}{N_d(x; -\mu\mathbf{1}_d, \sigma_1^2 \mathrm{I}_d)}.$$

Let $\rho_W(x,y)$ and $\rho_X(x,y)$ be the acceptance probability functions for W and X, respectively (notice that they do not depend on the proposal covariance). Then, if $X_{n-1} \in B_1$ and the proposal X^* is in A, $\rho_W(X_{n-1}, X^*)$ is within a factor of $1+\delta^d$ of $\rho_X(X_{n-1}, X^*)$.

If $X_{n-1} \in B_1$ and $X^* \in A^c \backslash B_2$, by Prop. A.4 we have

$$\rho_W(X_{n-1}, X^*), \rho_X(X_{n-1}, X^*) \le 2\alpha^d$$

giving the result. □

APPENDIX B: PROOF THAT $\pi(A) > B$ FOR THE POTTS MODEL

Letting $\sigma(z) = (\sigma_1(z), \ldots, \sigma_q(z))$ denote the sufficient statistic vector $\sigma_k(z) = \sum_i \mathbf{1}(z_i = k)$, we have

$$\pi(z) \propto \exp\left\{\frac{\alpha}{2M}\sum_{k=1}^{q}\sigma_k(z)^2\right\}$$

and the marginal distribution of σ is given by

$$\rho(\sigma) \propto \binom{M}{\sigma_1, \ldots, \sigma_q} \exp\left\{\frac{\alpha}{2M}\sum_{k=1}^{q}\sigma_k^2\right\}.$$

For $q \geq 3$ the critical value of the interaction parameter is $\alpha_c = \frac{2(q-1)\ln(q-1)}{q-2}$. Using Stirling's formula, Gore and Jerrum (1999) write $\binom{M}{\sigma_1,\ldots,\sigma_q}$ in terms of $a = (a_1, \ldots, a_q) = \sigma/M$ (the proportion of sites in each color):

$$\binom{M}{\sigma_1, \ldots, \sigma_q} = \exp\left\{ - M\sum_{k=1}^{q} a_k \ln a_k + \Delta(a)\right\}$$

where $\Delta(a)$ satisfies $\sup_a |\Delta(a)| = O(\ln M)$, and apply this to obtain:

$$\rho(\sigma) \propto \exp\{f_\alpha(a)M + \Delta(a)\} \quad \text{where} \quad f_\alpha(a) = \sum_{k=1}^{q}\left[\frac{\alpha}{2}a_k^2 - a_k \ln a_k\right]$$

Note f_α does not depend on M, and for $\alpha > \alpha_c$ has global maxima at permutations of $\bar{a} = \left(x, \frac{1-x}{q-1}, \ldots, \frac{1-x}{q-1}\right)$ for some $x \in \left[\frac{q-1}{q}, 1\right)$ (Gore and Jerrum, 1999; Woodard et al., 2009*b*).

Consider subsets $A_i = \left\{z : \sigma_i(z) > \frac{M}{2}\right\}$, and observe that when $q = 3$ we have $\pi(A_1) = \pi(A_2) = \pi(A_3)$ by symmetry. The distribution π concentrates near the global maxima of f_α, in the sense that for any $\epsilon > 0$, $\Pr\{\min_{s\in S_3} \|a(z) - s\bar{a}\|_2 < \epsilon\} \to 1$ as $M \to \infty$ (Gore and Jerrum, 1999), where S_3 is the symmetric group of 3 elements. If $\min_{s\in S_3} \|a(z) - s\bar{a}\|_2 < 1/6$ then $z \in A_1 \cup A_2 \cup A_3$, so $\pi(A_1 \cup A_2 \cup A_3) \to 1$ as $M \to \infty$, and there is some M^* such that $\pi(A_1) \geq \frac{1}{4}$ for $M > M^*$. For $q > 3$, the same argument yields some M^{**} such that $\pi(A) \geq \frac{1}{q+1}$ for $M > M^{**}$.

REFERENCES

Aldous, D. (1982), "Some inequalities for reversible Markov chains," *Journal of the London Mathematical Society*, 25, 564–576.

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

Andrieu, C., and Moulines, E. (2006), "On the ergodicity properties of some adaptive MCMC algorithms," *Annals of Applied Probability*, 16, 1462–1505.

Andrieu, C., and Roberts, G. O. (2009), "The Pseudo-Marginal Approach for Efficient Monte Carlo Computations," *Annals of Statistics*, 37(2), 697–725.

Atchadé, Y. F. (2009*a*), "A cautionary tale on the efficiency of some adaptive Monte Carlo schemes," *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.

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*, pp. 478–487.

Binder, K., and Heermann, D. W. (2002), *Monte Carlo Simulation in Statistical Physics*, 4^{th} edn Springer.

Craiu, R. V., Rosenthal, J., and Yang, C. (2009), "Learn from thy neighbor: Parallel-chain and regional adaptive MCMC," *Journal of the American Statistical Association*, . In press.

Douc, R., Guillin, A., Marin, J., and Robert, C. P. (2007), "Convergence of Adaptive Mixtures of Importance Sampling Schemes," *Annals of Statistics*, 35(1), 420–448.

Flegal, J. M. (2008), "Markov Chain Monte Carlo: Can We Trust the Third Significant Figure?," *Statistical Science*, 23(2), 250–260.

Frieze, A., Kannan, R., and Polson, N. (1994), "Sampling from log-concave distributions," *Annals of Applied Probability*, 4, 812–837.

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

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.

Geyer, C. J. (1991), Markov chain Monte Carlo maximum likelihood,, in *Computing Science and Statistics, Volume 23: Proceedings of the 23rd Symposium on the Interface*, ed. E. Keramidas, Interface Foundation of North America, Fairfax Station, VA, pp. 156–163.

Geyer, C. J., and Thompson, E. A. (1995), "Annealing Markov chain Monte Carlo with applications to ancestral inference," *Journal of the American Statistical Association*, 90, 909–920.

Gore, V. K., and Jerrum, M. R. (1999), "The Swendsen-Wang process does not always mix rapidly," *J. of Statist. Physics*, 97, 67–85.

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

Haario, H., Saksman, E., and Tamminen, J. (2001), "An adaptive Metropolis algorithm," *Bernoulli*, 7, 223–242.

Heaton, M., and Schmidler, S. C. (2009), A Multiscale Adaptive MCMC Algorithm,. (submitted).

Ji, C., and Schmidler, S. C. (2009), "Adaptive Markov chain Monte Carlo for Bayesian Variable Selection," *Journal of Computational and Graphical Statistics*, (to appear).

Jones, G. L., and Hobert, J. P. (2001), "Honest Exploration of Intractable Probability Distributions via Markov Chain Monte Carlo," *Statistical Science*, 16(4), 312–334.

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.

Lawler, G. F., and Sokal, A. D. (1988), "Bounds on the L^2 spectrum for Markov chains and Markov processes: a generalization of Cheeger's inequality," *Transactions of the American Mathematical Society*, 309, 557–580.

Liang, F., and Wong, W. H. (1999), "Dynamics Weighting in Simulatinos of Spin Systems," *PhysLettA*, 252, 257–262.

Liu, J. S., Liang, F., and Wong, W. H. (2001), "A Theory for Dynamics Weighting in Monte Carlo Computation," *Journal of the American Statistical Association*, 96(454), 561–573.

Madras, N., ed (2000), *Fields Institute Communications Volume 26: Monte Carlo Methods*, Providence, RI: American Mathematical Society.

Marinari, E., and Parisi, G. (1992), "Simulated tempering: a new Monte Carlo scheme," *Europhysics Letters*, 19, 451–458.

Mengersen, K. L., and Tweedie, R. L. (1996), "Rates of Convergence of Hastings and Metropolis Algorithms," *Annals of Statistics*, 24(1), 101–121.

Minary, P., and Levitt, M. (2006), "Discussion of "Equi-energy sampler" by Kou, Zhou, and Wong," *Annals of Statistics*, 34, 1636–1641.

Mira, A. (2001), "Ordering and Improving the Performance of Monte Carlo Markov Chains," *Statistical Science*, 16(4), 340–350.

Mira, A., and Geyer, C. J. (2000), "On Non-Reversible Markov Chains,". In Madras (2000).

Neal, R. (2003), "Slice sampling (with discussion)," *Annals of Statistics*, 31, 705–767.

Roberts, G. O., and Rosenthal, J. S. (2001), "Optimal Scaling for Various Metropolis-Hastings Algorithms," *Statistical Science*, 16(4), 351–367.

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 Rosenthal, J. S. (2007), "Coupling and Ergodicity of Adaptive MCMC," *Journal of Applied Probability*, 44, 458–475.

Schmidler, S. C., and Woodard, D. B. (2009), Computational complexity and Bayesian analysis,. In preparation.

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

Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions," *Annals of Statistics*, 22(4), 1701–1728.

Wang, F. G., and Landau, D. P. (2001), "Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States," *Physical Review Letters*, 86(10), 2050–2053.

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

E-mail: schmidler@stat.duke.edu

Department of Statistical Science
Box 90251
Duke University
Durham, NC 27708-0251
E-mail: schmidler@stat.duke.edu
URL: http://www.stat.duke.edu/∼scs

E-mail: dbw59@cornell.edu

School of Operations Research and
Information Engineering
206 Rhodes Hall
Ithaca, NY
E-mail: dbw59@cornell.edu
URL: http://people.orie.cornell.edu/∼woodard