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) 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, 2012; 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) 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 (2010) 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 another class of adaptive

MCMC techniques which tune parameters of a parametric transition kernel. Atchadé (2009) 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. When the multiple processes are based on the same Markov kernel, the bound shows that the mixing time of the adaptive sampler is limited by the conductance of the Markov kernel (Corollary 4.1). 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. We use our results to show that multichain resampling methods (including the equi-energy sampler and IR-MCMC) mix slowly on two examples: mixtures of normal distributions in $\mathbb{R}^d$, and the mean-field ferromagnetic Potts model. We also show that adaptive samplers in the second class described above (*invariant adaptive* methods) are slowly mixing on a mixture of normals.

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 framework for designing algorithms of this type has been provided by Wang and Schmidler (2012*b*).

In Section 2 and 3 we define the class of adaptive methods under consideration and give background on mixing time. Section 4 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 Section 6 gives our results for invariant adaptive methods. 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 $X_{0:n}^{(1:I)}$:

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

where $\sum_{kj} w_{ijkn} = 1$ 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 slow mixing results in Section 5 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 $w_{ijkn} \propto \delta_{E_{n-1}}(X_j^{(k)})\mathbf{1}_{\{k=i+1\}}$.) 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). 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 $w_{ijkn} \propto \frac{\pi^{(i)}(X_j^{(k)})}{\pi^{(k)}(X_j^{(k)})}\mathbf{1}_{\{k=i+1\}}$ chosen to be importance weights.

Adaptive Metropolized independence sampling (AMIS). AMIS methods construct an approximation of the target distribution from the current sample history $X_{0:n}$ to use as the proposal at time n; see e.g. Andrieu and Thoms (2008); Ji and Schmidler (2012). When this approximation takes the form of a non-parametric kernel-density estimator, the resulting transition kernel is of the form (1) with $\alpha = 0$ and Q_n instead a kernel mixture of the X_j for some mixing kernel $\mathcal{K}$. Although this does not strictly satisfy our definition of MRAM samplers, it behaves similarly and we will obtain a similar result (Theorem 5.2). In practice, a mixture distribution is often used in place of a kernel density estimate (Andrieu and Thoms, 2008; Ji and Schmidler, 2012); in this case our results are expected to hold, but technical conditions required for proof become more complicated.

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" d (typically the dimension of $\mathcal{X}$). The process is *slowly* mixing if the mixing time grows exponentially in the problem size. (To avoid trivial conditions on our theorems, we will also call the process slowly mixing if the number of chains $I(d)$ grows exponentially in d, as this also requires exponential computational effort.) 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, 2012). 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 $\pi(A) > 0$, denote the restriction of π to A by $\pi|_A(dy) \propto \pi(dy)\mathbf{1}_{\{y \in A\}}$. The restriction $T|_A$ of a Markov kernel T to A is defined to reject any move that would leave A:

$$T|_A(x, B) = T(x, B) + \mathbf{1}_{\{x \in B\}} T(x, A^c) \qquad x \in A, B \subset A.$$

Similarly, we define the restriction $Y = X|_A$ of an (adaptive) process X to A as a stochastic process that is independent of X, but defined identically to X except that any move leaving A is rejected. So for a MRAM sampler the ith chain of the restricted process Y is initialized according to the same distribution as X (so $X_0 \stackrel{\mathrm{d}}{=} Y_0$) and transitions according to $\alpha T_i|_A + (1 - \alpha)R^Y_{i,n}$, where $R^Y_{i,n}$ is the kernel that resamples from the history of the process Y (according to the rules of the specific algorithm).

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 $\hat{\theta}_n = n^{-1} \sum_{i=1}^n \theta(X_n)$ to $E_\pi(\theta(X))$, 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

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

and let $\Pr(H_A \leq n) \leq \epsilon$ for some $A \subset \mathcal{X}$. Then taking $\theta = 1_A(x)$ gives $\mathrm{Bias}^2(\hat{\theta}_n) \geq (\pi(A) - \epsilon)^2$. In this case the bias term may dominate and cannot be ignored.

4. Bounds for MRAM Processes. We first obtain bounds for MRAM samplers. Although the transition kernels $K_{i,n}$ depend on the history of the chain, the Markov kernels T_i on which they are based do not. This enables us 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 in chain i,

$$\rho(x,y) = \min\left\{1, \frac{\pi^{(i)}(dy)\pi^{(i+1)}(dx)}{\pi^{(i)}(dx)\pi^{(i+1)}(dy)}\right\} \tag{3}$$

leaves $\pi^{(i)}$ invariant only if the current and proposed states are *independent*; 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. In contrast, Markov chain methods monotonically approach their limiting distribution. The mixing parameter α in the MRAM kernel 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 χ^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 or n large this drift should be minimal; in order to analyze adaptive methods in the presence of this drift, we quantify it as follows. Recall the definition of the restricted process $X|_A$ from Section 3.

DEFINITION 4.1. *Let $A \subset \mathcal{X}$ such that $0 < \pi^{(i)}(A) < 1$ for all i, and consider the sampler $Y = X|_{A^c}$ with $Y_0^{(i)} \overset{ind.}{\sim} \pi^{(i)}|_{A^c}$. Let $\nu_n^{(i)}(dy)$ indicate the marginal distribution of $Y_n^{(i)}$, and define the ratio*

$$\psi_A = \max_{i,n} \frac{\int_{A^c} T_i(y,A)\nu_n^{(i)}(dy)}{\int_{A^c} T_i(y,A)\pi^{(i)}|_{A^c}(dy)}.$$

Since T_i is ergodic, the denominator $\int_{A^c} T_i(y,A)\pi^{(i)}|_{A^c}(dy)$ is strictly positive. Additionally, the numerator $\int_{A^c} T_i(y,A)\nu_n^{(i)}(dy)$ is ≤ 1, hence $\psi_A \leq \max_i \left[\int_{A^c} T_i(y,A)\pi^{(i)}|_{A^c}(dy)\right]^{-1} < \infty$. The quantity ψ_A, essentially a ratio of conductances, measures the extent to which the drift of the process away from its target distribution inflates the probability of transitioning to A under Markov kernel T_i.

In certain special cases $\psi_A = 1$: the degenerate case $\alpha = 1$; the single-chain method of Atchadé (2009), and more generally any multi-chain MRAM sampler having $\pi^{(i)} \equiv \pi$ and fixed resampling probabilities w_{ijkn} for all n, so the acceptance rate for resampling moves is one (see Corollary 4.1 below).

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 $0 < \pi^{(i)}(A) < 1$, 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. We have the following result:

THEOREM 4.1. *For any $\epsilon > 0$ and any $A \subset \mathcal{X}$, the mixing time τ_ϵ of the MRAM process satisfies:*

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

PROOF. Let $X_0^{(i)} \overset{\text{ind.}}{\sim} \pi^{(i)}|_{A^c}$, and 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 $\tau_\epsilon > n$.

Let $Y = X|_{A^c}$, and 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. The probability that X first hits A at time n (i.e. $H_A = n$) is equal to the probability that Y first attempts a move (in any of I chains) to A at time n but rejects due to restriction, so

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

where the second inequality comes from the mixture representation of (1), since the resampling proposal for the process Y satisfies $Q^Y_{i,j}(y, A) = 0$ 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

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

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 appearance of ψ_A in Theorem 4.1 is quite informative. Comparing with Theorem A strongly suggests that slow mixing of the original non-adaptive process implies slow mixing of the MRAM process, *unless ψ_A increases exponentially (in d)* for at least some set A. (We say "strongly suggests" because slight differences in the quantities appearing in the upper and lower bounds of Woodard et al. (2009*a*) and Woodard et al. (2009*b*) prevent this conclusion from following immediately.) Moreover since T_i is unchanged in MRAM processes, this means that $\nu_n^{(i)}$ must increase exponentially relative to the invariant distribution. More precisely, for every $A \subset \mathcal{X}$ with exponentially small conductance or persistence ($\max_i \gamma(A, i)\Phi_{T_i}(A)$), $\nu_n^{(i)}$ must increase exponentially relative to $\pi^{(i)}|_{A^c}$, on a set $\partial A \subset A^c$ that is "close" to A in the sense that $\inf_{y\in\partial A} T_i(y, A)$ is at most polynomially decreasing. That is, the drift of $\nu_n^{(i)}$ away from the stationary distribution must be exponentially large and done in precisely the right way so as to (exponentially) improve the conductance of all such A relative to the non-adaptive process T_i, i.e. $\nu_n^{(i)}(\partial A)/\pi^{(i)}(\partial A) \geq c^d$ for some $c > 1$. While some adaptation schemes may have the potential to achieve this, the methods in common use discussed in this paper (MRAM and IAMC), which focus primarily on adapting the *proposal* distribution, do not. This tells us a great deal about the utility of these approaches on hard problems, and provides guidance for future design of adaptive strategies; we return to this point in Section 7.

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, Theorem 4.1 is sufficient to analyze the effect of problem size on mixing time for fixed ϵ (as in Section 5).

4.1. *Common Markov kernel.* Consider a MRAM process with a common Markov kernel $T_i \equiv T$ and common target density $\pi^{(i)} \equiv \pi$, for which the resampling probabilities w_{ijkn} are fixed (do not depend on the history of the process) for each n and the resampling acceptance probability is one. The single-chain sampler of Atchadé (2009) is included in this class.

COROLLARY 4.1. *For any $0 < \epsilon < 1/4$, the mixing time τ_ϵ of a MRAM sampler with $\pi^{(i)} \equiv \pi$, $T_i \equiv T$, and fixed resampling probabilities w_{ijkn} and resampling acceptance probability one, satisfies:*

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

PROOF. For any measurable $A \subset \mathcal{X}$ such that $1/2 \leq \pi(A) < 1$ we have $\psi_A = 1$ by the following induction. Let $Y_0^{(i)} \sim \pi|_{A^c}$. Now assuming $Y_j^{(i)} \sim \pi|_{A^c}$ for $j = 0, \ldots, n-1$ for all i and some n, the distribution of $Y_n^{(i)}$ is a mixture of the distributions obtained by transitioning according to $T_i|_{A^c}$ and $R_{i,n}^Y$, which are $\pi|_{A^c}$ and (a mixture with each component equal to) $\pi|_{A^c}$, respectively. So $Y_n^{(i)} \sim \pi|_{A^c}$ and therefore $\psi_A = 1$ by induction. Theorem 4.1 then gives $\tau_\epsilon \geq (\pi(A) - \epsilon)[I\Phi_T(A)]^{-1}$, and the result follows from $\Phi_T(A) = \Phi_T(A^c)$. □

Compare this result with standard results for Markov chains. For $\mathcal{X}$ finite, assuming $T(x,x) \geq 3/4$ for all $x \in \mathcal{X}$ (which can be achieved by 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

$$\frac{1}{8\Phi_T} \ln(2\epsilon)^{-1} \leq \tau_\epsilon^* \leq \frac{8}{\Phi_T^2} \left[\ln(\max_x \pi(x)^{-1}) + \ln(\epsilon^{-1}) \right]. \tag{4}$$

The lower bounds in (4) 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 conductance of T decreases exponentially in the problem size d (so T is slowly mixing) then any MRAM process based on T of the type described in Corollary 4.1 is also 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 as described in Corollary 4.1.*

Corollary 4.2 proves in particular the conjecture of Atchadé (2009) 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 $\max_x \pi(x)^{-1}$ in Corollary 4.3 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. Moreover, if the original chain is slowly mixing due to initialization arbitrarily far away, the MRAM process would be expected to suffer similarly.

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:

$$\pi(x) = \frac{1}{2} N_d(x; -\mu \mathbf{1}_d, \sigma_1^2 \mathrm{I}_d) + \frac{1}{2} N_d(x; \mu \mathbf{1}_d, \sigma_2^2 \mathrm{I}_d) \tag{5}$$

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 conditions that Frieze et al. (1994) show rapid mixing of Metropolis-Hastings with local proposals on log-concave target densities in $\mathbb{R}^d$. ($\mathcal{X}$ unrestricted 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 $A_0 = \{x \in \mathbb{R}^d : \sum_i x_i \geq 0\}$ and show that if the target distributions $\pi^{(i)}$ are tempered versions of π, then $\max_i \gamma(A_0, i)\Phi_{T_i}(A_0)$ is exponentially decreasing for any choice of I temperatures whenever I is polynomial, and that consequently parallel tempering is slowly mixing. Since $\pi(A_0) \geq 1/2$ for all d large enough, it follows immediately from Theorem 4.1 that

COROLLARY 5.1. *Any MRAM process based on proposal S and tempered densities, with any number of chains $I(d)$, is slowly mixing on distribution (5) for $\sigma_1 > \sigma_2$ unless $\psi_{A_0}(d)$ grows exponentially in d.*

This shows that the class of samplers discussed after Definition 4.1, including the single-chain sampler of Atchadé (2009), is slowly mixing on the target (5). It also leads us to the following slow mixing result for the equi-energy and IR-MCMC samplers:

THEOREM 5.1. *Any equi-energy or IR-MCMC sampler based on proposal S, any number of chains $I(d)$, tempered densities with any inverse temperatures $\{\beta_i(d)\}_{i=1}^{I(d)}$, and any energy bin thresholds, is slowly mixing on distribution (5) for any fixed values of $(\mu, \sigma_1, \sigma_2)$ 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. One could prove Theorem 5.1 by applying Corollary 5.1 and bounding $\psi_{A_0}(d)$; however, we prove it directly (Appendix B) by adapting the proof of Theorem 6.1. A nearly identical proof (not provided here) yields Theorem 5.2.

THEOREM 5.2. *Any AMIS sampler with rotationally symmetric mixing kernel $\mathcal{K}$ is slowly mixing on distribution (5) for any fixed values of $(\mu, \sigma_1, \sigma_2)$ such that $\sigma_1 > \sigma_2$, $\mu > 2\sigma_1$ and $\sigma_1/\sigma_2 < \sqrt{e}$.*

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 d sites has distribution on $z \in \mathbb{Z}_q^d$ given by:

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

and we will be concerned with the "ferromagnetic" case $\lambda \geq 0$. 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 λ_c of the interaction parameter. For the mean-field ferromagnetic Potts ($q \geq 3$) model with $\lambda \geq \lambda_c$, 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). Define the subset $A_{\text{P}} = \left\{z : \sum_i \mathbf{1}(z_i = 1) > \frac{d}{2}\right\}$ of the Potts model state space. From Theorem 4.1, we have the following:

COROLLARY 5.2. *Any MRAM process based on single-site dynamics and using tempered densities is slowly mixing on the mean-field Potts model with $\lambda > \lambda_c$ unless $\psi_{A_P}(d)$ grows exponentially in d.*

This shows that the class of samplers discussed after Definition 4.1, including the single-chain sampler of Atchadé (2009), is slowly mixing on the mean field Potts model. We suspect that Corollary 5.2 holds for $\lambda = \lambda_c$, but this cannot be proven using Theorem 4.1 due to the term $(\pi(A) - \epsilon)$.

PROOF. Woodard et al. (2009*b*) show that for $\lambda \geq \lambda_c$ and any choice of a polynomial number I of temperatures, the quantity $\max_i \{\gamma(A_{\text{P}}, i)\Phi_{T_i}(A_{\text{P}})\}$ decreases exponentially as a function of d. Appendix C shows that, for $\lambda > \lambda_c$ and d sufficiently large, $\pi(A_{\text{P}}) > b$ for some positive constant b. Then from Theorem 4.1 the mixing time increases exponentially in d for any $\epsilon \in (0, b)$, i.e. the process 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 (5). 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.

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 (5) 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 this 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:

$$A = \left\{ x \in \mathbb{R}^d : \frac{N_d(x; \mu \mathbf{1}_d, \sigma_2^2 \mathrm{I}_d)}{N_d(x; -\mu \mathbf{1}_d, \sigma_1^2 \mathrm{I}_d)} > \delta^d \right\} \tag{6}$$

$$B_1 = \{x \in \mathbb{R}^d : \|x + \mu \mathbf{1}_d\| \leq \sigma_1 \sqrt{2d}\}$$

$$B_2 = \{x \in \mathbb{R}^d : \|x - \mu \mathbf{1}_d\| \leq 2\sigma_2 \sqrt{d}\}.$$

B_1 and B_2 are hyperspheres centered at the mean vectors of the components of π respectively, with $B_1 \subset A^c$ and $B_2 \subset A$ (Proposition A.2 in Appendix A). Standard concentration inequalities for sub-Gaussian random variables (Ledoux, 2001) yield $\Pr(\|Z - \mu \mathbf{1}_d\| > 2\sigma_2\sqrt{d}) \leq 2\mathrm{e}^{-\frac{d}{\kappa}}$ for $Z \sim N_d(\mu \mathbf{1}_d, \sigma_2^2 \mathrm{I}_d)$ and $\kappa = \frac{2}{1-\log(2)}$; hence for all d large enough ($d > 7$), we have $N_d(B_2; \mu \mathbf{1}_d, \sigma_2^2 \mathrm{I}_d) > 2/3$ and thus $\pi(A) \geq \pi(B_2) \geq 1/3$. Moreover, π concentrates in B_1 and B_2 as $d \to \infty$, and we will see that the Adaptive Metropolis and Inter-chain Adaptation algorithms have increasing difficulty moving between B_1 and B_2, causing slow mixing.

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 is the hitting time of A. As in the proof of Theorem 4.1, define $n_\epsilon(A) = \min\{n : \Pr(H_A \leq n) > \pi(A) - \epsilon\}$ and recall that $\tau_\epsilon \geq n_\epsilon(A)$. Lemma A.1 (Appendix A) shows $\exists \xi < 1$ such that $\Pr(H_A \leq n) \leq nI\xi^d$ by a coupling construction. Since $\pi(A) \geq 1/3$, for any $\epsilon < 1/6$ we therefore have $\tau_\epsilon \geq 1/(6I\xi^d)$. Unless I grows exponentially in d (which yields slow mixing by definition), we have that τ_ϵ grows exponentially in d. □

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.

7. Conclusions. These results appear to be the first non-asymptotic bounds on convergence for adaptive MCMC samplers. Our results for some commonly used adaptive samplers show that they perform no better 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 attainable 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. Our results also emphasize the importance of adapting the target distribution, not just the proposal kernel, to achieve improved mixing. This suggests consideration of other classes of adaptation schemes, such as those which encourage movements 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 (Liang and Wong, 1999; Liu et al., 2001; Wang and Schmidler, 2012*a*). A framework for designing sampling algorithms that achieve both (i) and (ii) above has been provided by Wang and Schmidler (2012*b*).

Acknowledgments. We thank Jeff Rosenthal for pointing out an error in a previous statement of Theorem 6.1. This work was partially supported by National Science Foundation grants CMMI-0926814 and DMS-1209103.

APPENDIX A: RESULTS FOR PROOF OF THEOREM 6.1

We first establish some simple properties of the sets A, B_1, and B_2. Let $c = \mu \left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right)^{-1} \left(\frac{1}{\sigma_2^2} + \frac{1}{\sigma_1^2}\right)$ and

$$r(d) = \sqrt{2d \log \frac{\sigma_1}{\delta \sigma_2} \left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right)^{-1} + d(c^2 - \mu^2)}.$$

PROPOSITION A.1. *A is an open ball of radius $r(d)$ centered at $c\mathbf{1}_d$.*

PROOF. For $x \in A$, rearranging the definition gives

$$\left(\frac{\sigma_1}{\delta \sigma_2}\right)^d \exp\left\{-\frac{d\mu^2}{2}\left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right)\right\} > \exp\left\{\frac{1}{2}\sum_{j=1}^{d}\left[x_j^2\left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right) - 2\mu x_j\left(\frac{1}{\sigma_2^2} + \frac{1}{\sigma_1^2}\right)\right]\right\}.$$

Taking the logarithm and completing the square then gives

$$2d \log \frac{\sigma_1}{\delta \sigma_2} - d\mu^2\left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right) + \sum_{j=1}^{d} \mu^2 \left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right)^{-1}\left(\frac{1}{\sigma_2^2} + \frac{1}{\sigma_1^2}\right)^2 > \left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right)\sum_{j=1}^{d}\left(x_j - \mu\left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right)^{-1}\left(\frac{1}{\sigma_2^2} + \frac{1}{\sigma_1^2}\right)\right)^2$$

which is the desired ball. □

PROPOSITION A.2. *$B_1 \subset A^c$ and $B_2 \subset A$.*

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

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

Using $\left(\frac{\sigma_1}{\sigma_2}\right)^d \leq e^{\frac{d}{2}}$, we then have for $x \in B_1$

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

Similarly, for $x \in B_2$ we have $\|x + \mu \mathbf{1}_d\| \geq \mu\sqrt{d}$ and

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

so $N_d(x; \mu\mathbf{1}_d, \sigma_2^2 \mathrm{I}_d) > N_d(x; -\mu\mathbf{1}_d, \sigma_1^2 \mathrm{I}_d)$. □

We now use these properties to establish the following lemma used in the proof of Theorem 6.1:

LEMMA A.1. *There is some $\xi < 1$ such that $Pr(H_A \leq n) \leq nI\xi^d$ for all d large enough and all n.*

To prove Lemma A.1 we will need the following results.

PROPOSITION A.3. *Let $W = W_0, W_1, \ldots$ be an Adaptive Metropolis chain with target distribution $\nu(x) = N_d(x; -\mu\mathbf{1}_d, \sigma_1^2 I_d)$ and $W_0 \sim \nu$. Let $\nu_n(\cdot)$ denote the marginal distribution of W_n, and $P_n(\cdot)$ the marginal distribution of the proposed state W_n^* at iteration n. There exists $\eta < 1$ such that for all d sufficiently large and all n we have $\max\{\nu_n(A), P_n(A)\} \leq \eta^d$.*

PROOF. Recall that B_1 is the ball of radius $\sigma_1\sqrt{2d}$ centered at $-\mu\mathbf{1}_d$, that A is the ball of radius $r(d)$ centered at $c\mathbf{1}_d$ (Prop. A.1), and that B_1 and A are disjoint (Prop. A.2). So $r(d) + \sigma_1\sqrt{2d} \leq \| - \mu\mathbf{1}_d - c\mathbf{1}_d\| = (\mu + c)\sqrt{d}$, yielding $r(d) \leq (\mu + c - \sqrt{2}\sigma_1)\sqrt{d}$. Let b be the angle of a right triangle with hypotenuse length $(\mu + c)\sqrt{d}$ and opposite side length $(\mu + c - \sqrt{2}\sigma_1)\sqrt{d}$; b does not depend on d. A simple geometric argument shows that there is an infinite (circular) cone with apex $-\mu\mathbf{1}_d$ and fixed aperture angle $2b < \pi$ that entirely contains the set A. The distribution ν_n is symmetric with respect to rotations about the point $-\mu\mathbf{1}_d$ (since the sampler's initial distribution, target distribution, and construction are symmetric with respect to such rotations). Hence $\nu_n(A)$ is bounded above by $\frac{1}{2} I_{\sin^2 b}(\frac{d-1}{2}, \frac{1}{2})$ where $I_x(a, b)$ is the regularized incomplete beta function, which is the proportion of the surface area of a hypersphere centered at $-\mu\mathbf{1}_d$ that lies inside of the cone of angle $2b$ (Li, 2011). Since $I_{\sin^2 b}(\frac{d-1}{2}, \frac{1}{2})$ decreases exponentially in d, $\nu_n(A)$ is bounded above by a quantity that does not depend on n and decreases exponentially in d. The same holds for $P_n(A)$.

To see that I decreases exponentially in d, note that $I_{\sin^2 b}(\frac{d-1}{2}, \frac{1}{2})$ is the cumulative distribution function of $X \sim \mathrm{Beta}(\frac{d-1}{2}, \frac{1}{2})$, evaluated at $\sin^2 b <$

1. So for $d > 2$,

$$I_{\sin^2 b}\left(\frac{d-1}{2}, \frac{1}{2}\right) \leq \frac{\int_0^{\sin^2 b} x^{\frac{d-3}{2}}(1-x)^{-\frac{1}{2}}dx}{\int_{\frac{d-1}{d}}^1 x^{\frac{d-3}{2}}(1-x)^{-\frac{1}{2}}dx} \leq \frac{(\sin^2 b)^{\frac{d-1}{2}}(1-\sin^2 b)^{-\frac{1}{2}}}{\left(\frac{d-1}{d}\right)^{\frac{d-3}{2}}\left(\frac{1}{d}\right)^{\frac{1}{2}}}$$

which decreases exponentially in d. □

A.1. Proof of Lemma A.1. Here we give the proof for the Adaptive Metropolis algorithm; the case of the Inter-chain Adaptation algorithm is nearly identical. The proof technique is inspired by that of Roberts and Rosenthal (2007), Theorem 1.

Let $T_\theta^X(x, \cdot)$ be the Metropolis kernel with normal random walk proposal $N_d(x, \theta)$ and target density $\pi(\cdot)$ defined in (5), 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 covariance adaptation function for Adaptive Metropolis.

Let $\delta, \eta < 1$ be defined as in (6) and Prop. A.3, respectively. We claim ("Claim A") that for all d, we can construct stochastic processes $W_0, W_1, \ldots$ and $X_0, X_1, \ldots$ such that $W_0 = X_0 \sim N_d(-\mu\mathbf{1}_d, \sigma_1^2 \mathrm{I}_d)$ and such that, for all $n \leq (8\eta^d + 2\delta^d)^{-1}$,

1. $X_n | X_{0:n-1} \sim T_{\theta_n^X}^X(X_{n-1}, \cdot)$ and $W_n | W_{0:n-1} \sim T_{\theta_n^W}^W(W_{n-1}, \cdot)$
2. $\Pr(W_{0:n} = X_{0:n}) \geq 1 - n(4\eta^d + \delta^d)$.

This says that $X_{0:n}$ is the Adaptive Metropolis process of interest, having target distribution $\pi(\cdot)$; that $W_{0:n}$ is the Adaptive Metropolis process described in Prop. A.3 having target distribution $N_d(-\mu\mathbf{1}_d, \sigma_1^2 \mathrm{I}_d)$; and that the two processes, when initialized in the same state, remain equal for time exponential in d with high probability.

Claim A is trivially true for $n = 0$. Suppose that it is true for some value $n-1$, where $0 < n \leq (8\eta^d + 2\delta^d)^{-1}$. Then, conditional on $W_{0:n-1} = X_{0:n-1}$ we have $\theta_n^W = \theta_n^X$, so the proposal distributions in the nth iteration for the two chains X and W are identical. Let X_n^* and W_n^* be the respective proposed states; we can define them jointly to satisfy $W_n^* = X_n^*$ with probability one. If $W_{0:n-1} \neq X_{0:n-1}$, generate X_n^* and W_n^* independently. Let $\rho_X(X_{n-1}, X_n^*)$ and $\rho_W(W_{n-1}, W_n^*)$ be the acceptance probabilities of the two proposals. Draw $U \sim \text{Uniform}(0, 1)$; if $U \leq \rho_X(X_{n-1}, X_n^*)$ then let $X_n = X_n^*$ and otherwise $X_n = X_{n-1}$. If $U \leq \rho_W(W_{n-1}, W_n^*)$ then let $W_n = W_n^*$ and otherwise $W_n = W_{n-1}$. This construction yields $X_n | X_{0:n-1} \sim T_{\theta_n^X}^X(X_{n-1}, \cdot)$ and $W_n | W_{0:n-1} \sim T_{\theta_n^W}^W(W_{n-1}, \cdot)$.

Now applying Prop. A.3) gives

(7)
$$\Pr(W_{n-1} \in A \text{ or } W_n^* \in A | X_{0:n-1} = W_{0:n-1}) \leq \frac{2\eta^d}{1-(n-1)(4\eta^d+\delta^d)} \leq 4\eta^d.$$

Conditional on $X_{0:n-1} = W_{0:n-1}$, on $W_{n-1} \in A^c$, and on $X_n^* = W_n^* \in A^c$, $\rho_X(X_{n-1}, X_n^*)$ and $\rho_W(W_{n-1}, W_n^*)$ are within a factor of $1+\delta^d$ of each other, since for $x, y \in A^c$

$$\frac{\pi(y)}{\pi(x)} \in \left[\left(\frac{1}{1+\delta^d}\right)\frac{N_d(y; -\mu\mathbf{1}_d, \sigma_1^2\mathrm{I}_d)}{N_d(x; -\mu\mathbf{1}_d, \sigma_1^2\mathrm{I}_d)},\ (1+\delta^d)\frac{N_d(y; -\mu\mathbf{1}_d, \sigma_1^2\mathrm{I}_d)}{N_d(x; -\mu\mathbf{1}_d, \sigma_1^2\mathrm{I}_d)}\right]. \tag{8}$$

So for any $X_{0:n-1} = W_{0:n-1}$ for which $W_{n-1} \in A^c$, and any $W_n^* \in A^c$, $\Pr(X_n \neq W_n | X_{0:n-1}, W_{0:n-1}, W_n^*) \leq \delta^d$. This yields

$$\begin{aligned}
&\Pr(X_n \neq W_n | X_{0:n-1} = W_{0:n-1}) \\
&\leq \Pr(W_{n-1} \in A \text{ or } W_n^* \in A | X_{0:n-1} = W_{0:n-1}) \\
&\qquad + \Pr(X_n \neq W_n | X_{0:n-1} = W_{0:n-1}, W_{n-1} \in A^c,\ W_n^* \in A^c) \\
&\leq (4\eta^d + \delta^d).
\end{aligned}$$

Hence Claim A is proven by induction.

Therefore for $n \leq (8\eta^d + 2\delta^d)^{-1}$ (using Prop. A.3)

$$\begin{aligned}
\Pr(H_A \leq n) &\leq \Pr(\exists j \leq n : W_j \in A) + \Pr(\exists j \leq n : X_j \neq W_j) \\
&\leq n(5\eta^d + \delta^d) \leq n(8\eta^d + 2\delta^d).
\end{aligned}$$

Notice that the same inequality, $\Pr(H_A \leq n) \leq n(8\eta^d + 2\delta^d)$, holds trivially for any other value of $n > (8\eta^d + 2\delta^d)^{-1}$. So taking $\xi = \max\{\eta, \delta\}^{\frac{1}{2}} < 1$ we have that for all d large enough $(8\eta^d + 2\delta^d) \leq \xi^d$, so $\Pr(H_A \leq n) \leq n\xi^d$ for all n. Recalling that $I = 1$ for Adaptive Metropolis, this is the desired result. □

APPENDIX B: PROOF OF THEOREM 5.1

We phrase the proof in terms of EES but the proof for IR-MCMC is nearly identical. We will use the definitions and results from the proof of Theorem 6.1. Recall that $\beta_i \in (0, 1]$ is the inverse temperature for chain i in the equi-energy sampler. Define $\nu(x) = N_d(x; -\mu\mathbf{1}_d, \sigma_1^2\mathrm{I}_d)$, and let ν^{β_i} be shorthand for the density proportional to $\nu(x)^{\beta_i}$ (which is $N_d(x; -\mu\mathbf{1}_d, \beta_i^{-1}\sigma_1^2\mathrm{I}_d)$). Initialize the chains as $X_0^{(i)} \sim \nu^{\beta_i}$. As in the proof of Theorem 6.1, we will show that there is some $\xi < 1$ such that $\Pr(H_A \leq n) \leq nI\xi^d$ for all d large enough and all n. We first require a result analogous to Prop. A.3:

PROPOSITION B.1. *Let $W = W_0, W_1, \ldots$ be an equi-energy sampler based on Metropolis proposal kernel S, with I chains having target densities $\nu^{(i)} = \nu^{\beta_i}$ and $W_0^{(i)} \overset{ind.}{\sim} \nu^{\beta_i}$. Let $\nu_n^{(i)}(\cdot)$ denote the marginal distribution of $W_n^{(i)}$. There exists $\eta < 1$ (independent of I, $\{\beta_i\}_{i=1}^I$, and the energy bin thresholds) such that for d sufficiently large, $\nu_n^{(i)}(A) \leq \eta^d$ for all i and n.*

PROOF. The initial, target, and proposal distributions are symmetric about $-\mu\mathbf{1}_d$, so by definition of the equi-energy sampler $\nu_n^{(i)}$ is also symmetric about $-\mu\mathbf{1}_d$ for all n. The result follows from the proof of Prop. A.3. □

Next we prove that $\Pr(H_A \leq n) \leq nI\xi^d$. Let $K_{i,n}^X = \alpha T_i^X + (1-\alpha)R_{i,n}^X$ for $i < I$ be the equi-energy transition kernel with target density proportional to $\pi(x)^{\beta_i}$ for $\pi(\cdot)$ in (5), where T_i^X is a Metropolis kernel with proposal S, and $R_{i,n}^X$ resamples the history $X_{0:n-1}$. $K_{I,n}^X = T_I^X$ since there is no resampling in chain I. Let $K_{i,n}^W = \alpha T_i^W + (1-\alpha)R_{i,n}^W$ be analogously defined for target densities ν^{β_i}, using the same energy bins and proposal S as X. We claim ("Claim B") that for all d we can construct stochastic processes $W_0, W_1, \ldots$ and $X_0, X_1, \ldots$ such that $W_0^{(i)} = X_0^{(i)} \overset{\text{ind.}}{\sim} \nu^{\beta_i}$ for each i and such that, for all $n \leq I^{-1}(8\eta^d + 6\delta^d)^{-1}$,

1. $X_n^{(i)}|X_{0:n-1} \sim K_{i,n}^X(X_{n-1}^{(i)}, \cdot)$ and $W_n^{(i)}|W_{0:n-1} \sim K_{i,n}^W(W_{n-1}^{(i)}, \cdot)$ independently across $i = 1, \ldots, I$
2. $\Pr(W_{0:n} = X_{0:n}) \geq 1 - nI(4\eta^d + 3\delta^d)$.

Claim B is trivially true for $n = 0$. Suppose that it is true for some value $n-1$, where $0 < n \leq I^{-1}(8\eta^d + 6\delta^d)^{-1}$. If $X^{(i)}$ transitions according to T_i^X, it proposes $X^{(i)*}$ according to S and accepts with probability $\rho_{X,T}(X_{n-1}^{(i)}, X^{(i)*}) = \min\{1, \pi(X^{(i)*})^{\beta_i}/\pi(X_{n-1}^{(i)})^{\beta_i}\}$. If it instead transitions according to $R_{i,n}^X$ then the acceptance rate is

$$\rho_{X,R}(X_{n-1}^{(i)}, X^{(i)*}) = \min\left\{1, \frac{\pi(X^{(i)*})^{\beta_i}\,\pi(X_{n-1}^{(i)})^{\beta_{i+1}}}{\pi(X_{n-1}^{(i)})^{\beta_i}\,\pi(X^{(i)*})^{\beta_{i+1}}}\right\}.$$

Similarly, chain $W^{(i)}$ will accept proposal $W^{(i)*}$ with probability $\rho_{W,T}(W_{n-1}^{(i)}, W^{(i)*}) = \min\{1, \nu(W^{(i)*})^{\beta_i}/\nu(W_{n-1}^{(i)})^{\beta_i}\}$ or

$$\rho_{W,R}(W_{n-1}^{(i)}, W^{(i)*}) = \min\left\{1, \frac{\nu(W^{(i)*})^{\beta_i}\,\nu(W_{n-1}^{(i)})^{\beta_{i+1}}}{\nu(W_{n-1}^{(i)})^{\beta_i}\,\nu(W^{(i)*})^{\beta_{i+1}}}\right\}.$$

Conditional on $W_{0:n-1} = X_{0:n-1}$ the proposal distributions for $X^{(i)*}$ and $W^{(i)*}$ are identical; only the acceptance probabilities differ. We can define

$X^{(i)*}$ and $W^{(i)*}$ jointly to satisfy $X^{(i)*} = W^{(i)*}$ and ensure that the same move type (Metropolis or equi-energy) is proposed, for each i. Doing this (and applying Prop. B.1) gives

$$\Pr(W_{n-1}^{(i)} \in A \text{ or } W^{(i)*} \in A | X_{0:n-1} = W_{0:n-1}) \leq \frac{2\eta^d}{1-(n-1)I(4\eta^d + 3\delta^d)} \leq 4\eta^d. \tag{9}$$

Conditional on $X_{0:n-1} = W_{0:n-1}$, on $W_{n-1}^{(i)} \in A^c$, and on $X^{(i)*} = W^{(i)*} \in A^c$, the acceptance probabilities of the proposals $X^{(i)*}$ and $W^{(i)*}$ are within a factor of $1 + 3\delta^d$ of each other, shown as follows. Using (8), for $x, y \in A^c$ and any i'

$$\left(\frac{\pi(y)}{\pi(x)}\right)^{\beta_{i'}} \in \left[\left(\frac{1}{1+\delta^d}\right)^{\beta_{i'}} \frac{\nu(y)^{\beta_{i'}}}{\nu(x)^{\beta_{i'}}},\ (1+\delta^d)^{\beta_{i'}} \frac{\nu(y)^{\beta_{i'}}}{\nu(x)^{\beta_{i'}}}\right] \subset \left[\left(\frac{1}{1+\delta^d}\right) \frac{\nu(y)^{\beta_{i'}}}{\nu(x)^{\beta_{i'}}},\ (1+\delta^d) \frac{\nu(y)^{\beta_{i'}}}{\nu(x)^{\beta_{i'}}}\right].$$

So conditional on $X_{0:n-1} = W_{0:n-1}$, on $W_{n-1}^{(i)} \in A^c$, and on $X^{(i)*} = W^{(i)*} \in A^c$, the acceptance rates $\rho_{X,T}(X_{n-1}^{(i)}, X^{(i)*})$ and $\rho_{W,T}(W_{n-1}^{(i)}, W^{(i)*})$ are within a factor of $(1+\delta^d)^2 \leq 1 + 3\delta^d$ of each other, as are the acceptance rates $\rho_{X,R}(X_{n-1}^{(i)}, X^{(i)*})$ and $\rho_{W,R}(W_{n-1}^{(i)}, W^{(i)*})$.

Claim B, and therefore $\Pr(H_A \leq n) \leq nI\xi^d$ for all n, then follows by an identical argument to that of Lemma A.1 using Prop. B.1 in place of Prop. A.3. □

APPENDIX C: PROOF THAT $\pi(A_{\mathrm{P}})$ IS BOUNDED FOR 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

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

and the marginal distribution of σ is given by

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

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

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

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

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

Note f_λ does not depend on d, and for $\lambda > \lambda_c$ 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 $A_i = \left\{ z : \sigma_i(z) > \frac{d}{2} \right\}$, 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_λ, in the sense that for any $\epsilon > 0$, $\Pr\{\min_{s \in S_3} \|a(z) - s\bar{a}\|_2 < \epsilon\} \to 1$ as $d \to \infty$ (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 $d \to \infty$, and there is some d^* such that $\pi(A_1) \geq \frac{1}{4}$ for $d > d^*$. For $q > 3$, the same argument yields some d^{**} such that $\pi(A_{\text{P}}) \geq \frac{1}{q+1}$ for $d > d^{**}$.

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.

Andrieu, C., and Thoms, J. (2008), "A Tutorial on Adaptive MCMC," *Statistics and Computing*, 18, 343–373.

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

Atchadé, Y. F. (2010), "A cautionary tale on the efficiency of some adaptive Monte Carlo schemes," *Annals of Applied Probability*, 20, 841–868.

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*, 488, 1454–1466.

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.

Ji, C., and Schmidler, S. C. (2012), "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.

Ledoux, M. (2001), *The Concentration of Measure Phenomenon*, Vol. 89 of *Mathematical Surveys and Monographs* American Mathematical Society.

Li, S. (2011), "Concise Formulas for the Area and Volume of a Hyperspherical Cap," *Asian Journal of Mathematics and Statistics*, 4(1), 66–70.

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. (2012), 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.

Wang, J., and Schmidler, S. C. (2012*a*), Adaptive Energy Partitioning for Generalized Wang-Landau Sampling,. (submitted).

Wang, J., and Schmidler, S. C. (2012*b*), An Exploration/Exploitation Approach to Adaptive MCMC,. (in preparation).

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