CONDITIONS FOR RAPID AND TORPID MIXING OF PARALLEL AND SIMULATED TEMPERING ON MULTIMODAL DISTRIBUTIONS

by

Dawn B. Woodard

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Scott C. Schmidler, Supervisor

_____
Robert L. Wolpert

_____
Michael L. Lavine

_____
Mark L. Huber

_____

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Statistical Science
in the Graduate School of
Duke University

2007

ABSTRACT

CONDITIONS FOR RAPID AND TORPID MIXING OF PARALLEL AND SIMULATED TEMPERING ON MULTIMODAL DISTRIBUTIONS

by

Dawn B. Woodard

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Scott C. Schmidler, Supervisor

_____
Robert L. Wolpert

_____
Michael L. Lavine

_____
Mark L. Huber

_____

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University

2007

Abstract

Stochastic sampling methods are ubiquitous in statistical mechanics, Bayesian statistics, and theoretical computer science. However, when the distribution that is being sampled is multimodal, many of these techniques converge slowly, so that a great deal of computing time is necessary to obtain reliable answers. Parallel and simulated tempering are sampling methods that are designed to converge quickly even for multimodal distributions. In this thesis, we assess the extent to which this goal is acheived.

We give conditions under which a Markov chain constructed via parallel or simulated tempering is guaranteed to be rapidly mixing, meaning that it converges quickly. These conditions are applicable to a wide range of multimodal distributions arising in Bayesian statistical inference and statistical mechanics. We provide lower bounds on the spectral gaps of parallel and simulated tempering. These bounds imply a single set of sufficient conditions for rapid mixing of both techniques. A direct consequence of our results is rapid mixing of parallel and simulated tempering for several normal mixture models in $\mathbb{R}^M$ as M increases, and for the mean-field Ising model.

We also obtain upper bounds on the convergence rates of parallel and simulated tempering, yielding a single set of sufficient conditions for torpid mixing of both techniques. These conditions imply torpid mixing of parallel and simulated tempering

on a normal mixture model with unequal covariances in $\mathbb{R}^M$ as M increases and on the mean-field Potts model with $q \geq 3$, regardless of the number and choice of temperatures, as well as on the mean-field Ising model if an insufficient (fixed) set of temperatures is used. The latter result is in contrast to the rapid mixing of parallel and simulated tempering on the mean-field Ising model with a linearly increasing set of temperatures.

Contents

List of Figures

Chapter 1

Introduction

Stochastic sampling methods have become ubiquitous in statistics, computer science, and statistical physics. In physics, samples according to the energy function are used to estimate the probability that a physical system in equilibrium is in a particular configuration, the frequency with which the system changes between configurations, and other physical quantities of interest. In Bayesian statistical model-fitting, samples from the posterior distribution of model parameters are used for simultaneous estimation of the parameters along with the associated uncertainty, which can then be used for prediction or for estimation of additional quantities of interest. Model comparison (hypothesis testing) is another area of Bayesian statistics in which it is often necessary to obtain samples from a distribution of interest. Model comparison may be done, for instance, by drawing samples from the joint posterior distribution of the model and the parameters.

Another area where sampling methods are used in Bayesian statistics is in estimation of missing data. One such technique is to draw samples from the joint posterior distribution of the parameters and the missing data. This approach has an advantage over traditional methods such as Expectation-Maximization in that it measures the uncertainty in the missing data and incorporates that uncertainty into parameter estimation. Missing data problems include censored data, classification, and latent variable models such as stochastic volatility models.

For all of these applications, samples from a distribution of interest are used via Monte Carlo integration to solve the described problem. In special cases it is possible to draw independent samples exactly from the target distribution, for instance in conjugate Bayesian models; however, in general this is a difficult problem. In Bayesian inference the target is a posterior distribution, the form of which can be complex and which typically is only known up to a normalizing constant. In cases such as this one, a widely-applicable approach is to construct a Markov chain having the target distribution as its limiting distribution. Although the resulting samples are not independent, they satisfy laws of large numbers and often central limit theorems, and thus can still be used for Monte Carlo integration (Tierney 1994; Robert and Casella 1999). Such Markov chain Monte Carlo (MCMC) methods have revolutionized computation in Bayesian statistics (Gilks et al. 1996), provided significant breakthroughs in theoretical computer science (Jerrum and Sinclair 1996), and become a staple of physical simulations (Binder and Heermann 2002).

Due to the variety of applications of MCMC, a number of MCMC procedures have been developed. In Bayesian model-fitting, if the conditional posterior distribution is available for each parameter then Gibbs sampling can be used (Geman and Geman 1984; Gelfand and Smith 1990). If this is not the case, then sampling can be performed via Metropolis-Hastings (Metropolis et al. 1953), although this requires specification of a proposal kernel that is potentially difficult to tune. In addition, adaptive rejection Metropolis sampling can be embedded into a Gibbs or Metropolis-Hastings scheme (Gilks et al. 1995). The use of these methods together in hybrid samplers is common, where the sampling method for each parameter or block of parameters can be chosen using application-specific knowledge if that is available (Tierney 1994).

For the problem of model selection, reversible jump provides an MCMC method for sampling from the joint posterior distribution of the model and parameters (Green 1995). For estimation of missing data, data augmentation can be used to sample from the joint posterior distribution of the parameters and missing data (Tanner and Wong 1987).

A common difficulty arising in the application of MCMC methods is that many target distributions arising in statistics and statistical physics are strongly multimodal; in such cases a Markov chain consisting of only local moves can take an impractically long time to reach stationarity. Even after reaching stationarity, the chain can have very long-range dependence (slow "mixing"), which decreases the

accuracy of the Monte Carlo estimates. Multimodality occurs in physical systems that have multiple low-energy configurations separated by energy barriers, such as the ferromagnetic Potts model (Swendsen and Wang 1987). In Bayesian statistics, if there are multiple likely parts of the parameter space separated by regions of low likelihood, then the posterior distribution can be multimodal. This can occur in nonparametric models, mixture models, model selection and change point problems (Neal 1996; Liang and Wong 2001; Lauritzen 1996; Liang and Wong 2000). For the example of model selection in linear regression, the posterior probability of the possible models is a criterion for model selection and model averaging; Liang and Wong (2000) show that multimodality of this posterior distribution over models can lead to slow mixing of Metropolis-Hastings with local moves. The marginal distribution of the C_p statistic for one of their examples is shown in Figure 1.1, illustrating this multimodality.

Complicating matters further, standard techniques for detecting slow convergence and mixing of MCMC cannot detect the presence of an undiscovered mode (Geweke 1992; Cowles and Carlin 1996). If the MCMC is trapped in a local mode, but has mixed well within that mode, then it can pass such diagnostics. This means that important modes can be missed by using MCMC. This is illustrated in Figure 1.2 for a mixture of two bivariate normals. Independent samples from the mixture are shown, as are dependent samples from the same mixture obtained using a Metropolis-Hastings chain with local steps. Although only one of the modes has been sampled,

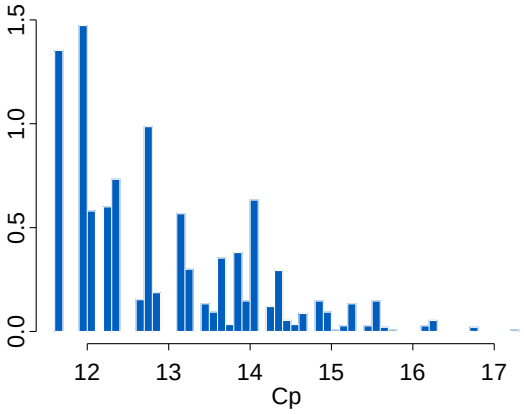

Figure 1.1: The marginal distribution of the C_p statistic over the possible models, for a data set simulated as described in Liang and Wong (2000).

the autocorrelation and time series plots for this chain do not show the lack of convergence.

In order to rigorously analyze the convergence and mixing of a Markov chain, it is necessary to estimate or bound the true convergence and mixing rates. Since the most commonly used MCMC algorithms construct reversible Markov chains, or can be made reversible without significant alteration, the convergence and mixing rates can both be bounded using the spectral gap of the transition operator (kernel) (Madras and Slade 1993). A variety of techniques have been developed to obtain bounds on the spectral gap of reversible Markov chains (Lawler and Sokal 1988; Diaconis and Stroock 1991; Sinclair 1992).

Application of such techniques to Metropolis-Hastings with local moves shows impractically slow mixing for a number of multimodal examples (Madras and Zheng

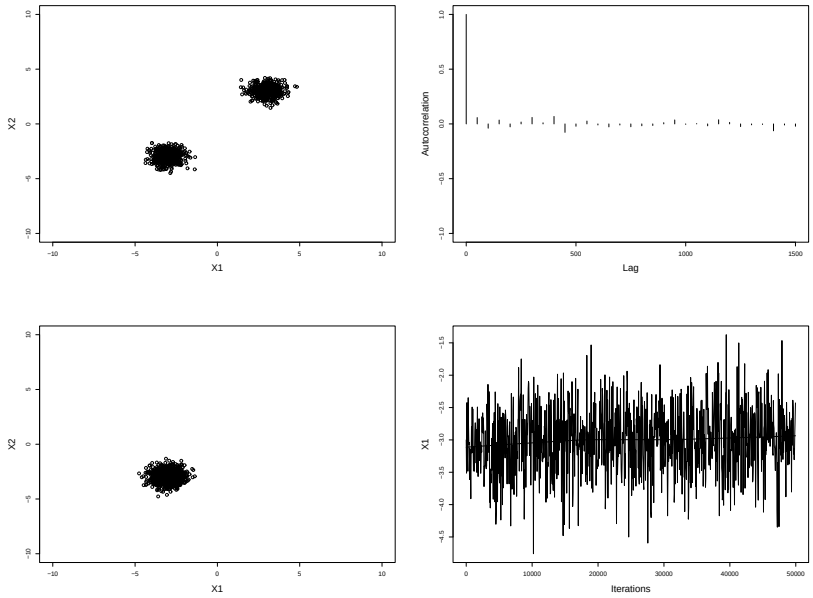

Figure 1.2: Top left: Independent samples from a mixture of two bivariate normal distributions. Bottom left: Samples from the same mixture obtained by a Metropolis-Hastings chain. Top and bottom right: The autocorrelation and time series plots for this chain.

2003), reinforcing the empirical evidence of slow mixing on multimodal distributions. A number of alternative MCMC algorithms have therefore been proposed for sampling from multimodal distributions. The evolutionary Monte Carlo algorithm (Liang and Wong 2000) maintains a population of samples rather than a single sample, and occasionally proposes hybridization of samples with the goal of discovering unexplored modes. The slice sampler (Neal 2003) attempts to jump between the modes by introducing a move that samples uniformly from the subset of the state space with density at least as high as that of the current state.

Two of the most popular and empirically successful MCMC algorithms for multimodal problems are Metropolis-coupled MCMC or *parallel tempering* (Geyer 1991), and *simulated tempering* (Geyer and Thompson 1995; Marinari and Parisi 1992). These algorithms flatten a distribution via "heating"; this technique arises naturally in physical systems where the temperature is a parameter of the system. Due to their common use, adequate theoretical characterization of the convergence and mixing of chains constructed via simulated or parallel tempering is of significant interest. Zheng (2003) bounds the spectral gap of simulated tempering below by a multiple of the spectral gap of parallel tempering, where the multiplier depends on a measure of the overlap between distributions at adjacent temperatures. Madras and Piccioni (1999) analyze a variant of simulated tempering as a mixture of the component chains at each temperature.

Madras and Randall (2002) develop decomposition theorems for bounding the

spectral gap of a Markov chain, then use those theorems to bound the mixing of simulated tempering in terms of the slowest mixing of the tempered chains. If Metropolis-Hastings mixes slowly on the original (untempered) distribution, their bound cannot be used to show fast mixing of simulated tempering.

However, fast mixing of simulated tempering has been shown for several multimodal distributions for which local Metropolis-Hastings mixes slowly. Madras and Zheng (2003) bound the spectral gap of parallel and simulated tempering on two examples, the "exponential valley" density and the mean-field Ising model. They use the decomposition theorems of Madras and Randall (2002). However, unlike Madras and Randall they decompose the state spaces of these examples into subsets on which the target distribution is unimodal, one subset for each local mode, and bound the mixing of parallel and simulated tempering in terms of the mixing within each subset. Since for these examples Metropolis-Hastings is rapidly mixing on the unimodal subsets, their bounds are able to show rapid mixing of parallel and simulated tempering. This is in contrast to the standard (untempered) Metropolis-Hastings chain, which is torpidly mixing. Here torpid mixing means that the spectral gap decreases exponentially in the problem size, while rapid mixing means that it decreases polynomially. The rapid / torpid mixing distinction is a measure of the computational tractability of the algorithm.

The results of Madras and Zheng (2003) are extended by Bhatnagar and Randall (2004) to show the rapid mixing of parallel and simulated tempering on an asymmet-

ric exponential valley density and the rapid mixing of a variant of parallel tempering on the mean-field Ising model with external field. These authors use the same decomposition as Madras and Zheng (2003). Bhatnagar and Randall (2004) also show that parallel and simulated tempering are torpidly mixing on the mean-field Potts model with $q = 3$, regardless of the number and choice of temperatures.

We generalize the decomposition approach of Madras and Zheng (2003) and Bhatnagar and Randall (2004) to obtain lower bounds on the spectral gap of any parallel or simulated tempering chain for any target distribution, defined on any state space (Theorem 3.1.1 and Corollary 3.1.1). As these authors do, we partition the state space into subsets on which the target density is unimodal. Then we bound the spectral gap of parallel and simulated tempering in terms of the mixing within each unimodal subset and the mixing among the subsets. Since Metropolis-Hastings for a unimodal distribution is often rapidly mixing, these bounds can be tighter than the simulated tempering bound of Madras and Randall (2002).

We then use our bounds to obtain a set of conditions under which parallel and simulated tempering chains are guaranteed to be rapidly mixing. We require that Metropolis-Hastings is rapidly mixing when restricted to any one of the unimodal subsets. The challenge is then to ensure that the tempering chain is able to cross between the modes efficiently. In order to guarantee rapid mixing of the tempering chain, we require that the highest-temperature chain mixes rapidly among the unimodal subsets. We also require that the the overlap between distributions at ad-

jacent temperatures decreases no more than polynomially in the problem size, which is necessary in order to mix rapidly among the temperatures.

In the case where the modes are symmetric, we show that these conditions are sufficient to guarantee rapid mixing, and give two examples where they hold: a mixture of normal distributions with equal covariance matrices in $\mathbb{R}^M$ as M increases (Section 4.2), and the mean-field Ising model (Section 4.1). Mixtures of normal distributions are of interest due to their ubiquity in statistics. We also obtain an additional condition that guarantees rapid mixing in the general (asymmetric) case, and use this to show rapid mixing for a weighted mixture of normal distributions in $\mathbb{R}^M$ as M increases. The additional condition states that a quantity related to the persistence of the unimodal subsets is polynomially decreasing, where persistence is as defined in Section 5.1.

We also show how the failure of this last condition can imply torpid mixing of parallel and simulated tempering. We state that if there is a set with exponentially low conductance at low temperatures (e.g. a unimodal subset of a highly multimodal distribution) and exponentially low persistence, then parallel and simulated tempering are torpidly mixing. This is the case for a normal mixture model with unequal covariances in $\mathbb{R}^M$ and for the mean-field Potts model with $q \geq 3$, as we show. These results are regardless of the number and choice of temperatures in parallel or simulated tempering.

Additionally, we show how the failure of the condition on the overlap can imply

torpid mixing. If the distribution has multiple high modes, so that local Metropolis-Hastings is torpidly mixing at low temperatures, and if two adjacent temperatures are widely spaced, so that the overlap between those levels is exponentially decreasing, then parallel and simulated tempering are torpidly mixing. This is the case if fixed temperatures are used for the mean-field Ising model, as we show.

We obtain a set of upper bounds on the spectral gaps of parallel and simulated tempering (Theorems 5.1.1 and 5.1.2 and Corollary 5.1.1), which hold for any target distribution, on any state space. These bounds imply the above conditions for torpid mixing, which we use to show the torpid mixing of parallel and simulated tempering on the normal mixture with unequal covariances in Section 6.1, on the mean-field Potts model with $q \geq 3$ in Section 6.2, and on the mean field Ising model with fixed temperatures in Section 6.3. The second example builds on the results of Bhatnagar (2007), who shows the torpid mixing of parallel and simulated tempering on the mean-field Potts model with $q = 3$. Potts-type models are used in statistical physics as well as in Bayesian image analysis and for modeling spatial random effects (Banerjee et al. 2004; Geman and Geman 1984; Green and Richardson 2002).

Chapter 2

Background

2.1 Sampling Using Markov Chains

Take any σ-finite measure space $(\mathcal{X}, \mathcal{F}, \lambda)$ with $\mathcal{F}$ countably generated; for example, $\mathcal{X}$ can be $\mathbb{R}^d$ and λ Lebesgue measure, or $\mathcal{X}$ can be countable and λ counting measure. We will consider "arbitrary" subsets $A \subset \mathcal{X}$, by which we are referring to any $A \in \mathcal{F}$. In order to draw samples from a distribution μ on $(\mathcal{X}, \mathcal{F})$, one may simulate a Markov chain that has μ as its limiting distribution, as we now describe. Let P be a transition kernel on $\mathcal{X}$, defined as in Tierney (1994), which operates on a distribution μ on the left:

$$(\mu P)(A) = \int \mu(dx) P(x, A) \qquad \forall A \subset \mathcal{X}.$$

If $\mu P = \mu$ then call μ a stationary distribution of P. One way of finding a transition kernel with stationary distribution μ is by constructing it to be reversible with respect to μ, as we now describe.

P operates on complex-valued functions f on the right:

$$(Pf)(x) = \int f(y)P(x, dy) \qquad \forall x \in \mathcal{X}.$$

Define the inner product $(f, g)_\mu = \int \overline{f(x)} g(x) \mu(dx)$ and denote by $L_2(\mu)$ the set of functions f such that $(f, f)_\mu$ is finite. P is called *reversible* with respect to μ if $(f, Pg)_\mu = (Pf, g)_\mu$ for all $f, g \in L_2(\mu)$ and *nonnegative definite* if $(Pf, f)_\mu \geq 0$ for all $f \in L_2(\mu)$. If P is reversible with respect to μ then μ is easily seen to be a stationary distribution of P.

We will primarily be interested in the case where μ has a density π with respect to λ. Define $\pi[A] = \mu(A)$ and define $(f, g)_\pi$, $L_2(\pi)$, and π-reversibility to be equal to the corresponding quantities for μ.

If P is ϕ-irreducible and aperiodic (defined as in Roberts and Rosenthal (2004)), nonnegative definite, and μ-reversible, then the Markov chain with transition kernel P converges in distribution to μ at a rate related to the *spectral gap*:

$$\mathbf{Gap}(P) = \inf_{\substack{f \in L_2(\mu) \\ \mathrm{Var}_\mu(f) > 0}} \left(\frac{\mathcal{E}(f, f)}{\mathrm{Var}_\mu(f)} \right). \tag{2.1}$$

Here $\mathcal{E}(f, f)$ is the *Dirichlet form* $(f, (I - P)f)_\mu$, and $\mathrm{Var}_\mu(f)$ is the variance $(f, f)_\mu - (f, 1)_\mu^2$. It can easily be shown that $\mathbf{Gap}(P) \in [0, 1]$ (for P not nonnegative definite, $\mathbf{Gap}(P) \in [0, 2]$).

For any distribution μ_0 having a density π_0 with respect to μ, define the L_2-norm $\|\mu_0\|_2 = (\pi_0, \pi_0)_\mu^{1/2}$. For the Markov chain with transition kernel P define the rate of convergence to stationarity as:

$$r = \inf_{\mu_0} \lim_{n\to\infty} \frac{-\ln(\|\mu_0 P^n - \mu\|_2)}{n} \tag{2.2}$$

where the infimum is taken over distributions μ_0 that have a density π_0 with respect to μ such that $\pi_0 \in L_2(\mu)$. The rate r is equal to $-\ln(1 - \mathbf{Gap}(P))$; for every μ_0 that has a density $\pi_0 \in L_2(\mu)$,

$$\|\mu_0 P^n - \mu\|_2 \leq \|\mu_0 - \mu\|_2 e^{-rn} \qquad \forall n \in \mathbb{N}.$$

In addition, r is the largest quantity for which this holds for all such μ_0. These are facts from functional analysis (see e.g. Yuen (2001); Madras and Slade (1993); Roberts and Tweedie (2001)). Therefore for a particular such starting distribution μ_0, the number of iterations n until the L_2-distance to stationarity is less than some $\epsilon > 0$ is on the order $O(r^{-1})$. The constant here depends on $\|\mu_0 - \mu\|_2$ and ϵ. Analogous results hold when the chain is started deterministically at x for μ-a.e. $x \in \mathcal{X}$, rather than drawn randomly from a starting distribution μ_0 (Roberts and Tweedie 2001).

Madras and Slade (1993) also show that the autocorrelation of the chain decays at a rate r. Their proof is stated for finite state spaces but applies to general state spaces as well. Therefore, informally speaking, the number of iterations of the chain required to obtain some number N_0 of approximately independent samples from μ is $O(N_0 r^{-1})$.

Note that the quantity r is monotonically increasing with $\mathbf{Gap}(P)$. Therefore lower (upper) bounds on $\mathbf{Gap}(P)$ correspond to lower (upper) bounds on the rate of convergence to stationarity. In addition, $-\ln(1-\mathbf{Gap}(P))/\mathbf{Gap}(P)$ approaches 1 as $\mathbf{Gap}(P) \to 0$. Therefore the order at which $\mathbf{Gap}(P) \to 0$ as a function of the problem size is equal to the order at which the rate of convergence to stationarity approaches zero. When $\mathbf{Gap}(P)$ is exponentially decreasing as a function of the problem size, we call P *torpidly mixing*. When $\mathbf{Gap}(P)$ is polynomially decreasing as a function of the problem size, we call P *rapidly mixing*. The rapid / torpid mixing distinction is a measure of the computational tractability of an algorithm.

Next we describe a common way of constructing a transition kernel that is reversible with respect to a particular density of interest π.

2.2 Metropolis-Hastings

Consider a transition kernel $P(w, dz)$ (the "proposal" kernel) that has a density $p(w, \cdot)$ with respect to λ for every w, and define the Metropolis-Hastings transition kernel for P with respect to π as follows (Metropolis et al. 1953). Propose a move z according to $P(w, \cdot)$, where w is the current state, accept the move with probability

$$\rho(w, z) = \min\left\{1, \frac{\pi(z)p(z, w)}{\pi(w)p(w, z)}\right\}$$

and otherwise reject. The resulting transition kernel is easily seen to be reversible with respect to π. If $p(w, z) = p(z, w)$ for all $z \neq w$ then P is called symmetric.

2.3 Parallel and Simulated Tempering

If the Metropolis-Hastings proposal kernel moves only locally in the space, and if π has more than one mode, then the Metropolis-Hastings chain may move between the modes of π infrequently. Tempering is a modification of Metropolis-Hastings wherein the density of interest π is "flattened" in order to allow movement among the modes of π, meaning the following. For any *inverse temperature* $\beta \in [0, 1]$ such that $\int \pi(z)^\beta \lambda(dz) < \infty$, define

$$\pi_\beta(z) = \frac{\pi(z)^\beta}{\int \pi(w)^\beta \lambda(dw)} \qquad \forall z \in \mathcal{X}.$$

Note that for any $z, w \in \mathcal{X}$ such that $\pi(z), \pi(w) \neq 0$, the ratio $\pi_\beta(z)/\pi_\beta(w)$ monotonically approaches one as β decreases, flattening the resulting density. For any β, define H_β to be the Metropolis-Hastings chain with respect to π_β; more generally, assume that we have some way to specify a π_β-reversible transition kernel for each β, and call this kernel H_β.

Let $\mathcal{B} = \left\{\beta \in [0, 1] : \int \pi(z)^\beta \lambda(dz) < \infty\right\}$. The parallel tempering algorithm (Geyer 1991) simulates parallel Markov chains with transition kernels H_{β_k} where $\beta_0 < \ldots < \beta_N = 1$ and $\beta_0 \in \mathcal{B}$. The inverse temperatures are commonly specified in a geometric progression, and Predescu et al. (2004) show an asymptotic optimality result for this choice.

Updates of individual chains are alternated with proposed swaps between temperatures, so that the process forms a single Markov chain with state $x = (x_{[0]}, \ldots, x_{[N]})$

on the space $\mathcal{X}_{pt} = \mathcal{X}^{N+1}$ and stationary density

$$\pi_{pt}(x) = \prod_{k=0}^{N} \pi_{\beta_k}(x_{[k]}) \qquad x \in \mathcal{X}_{pt}.$$

with product measure $\lambda_{pt}(dx) = \prod_{k=0}^{N} \lambda(dx_{[k]})$. The marginal density of $x_{[N]}$ under stationarity is π, the density of interest.

A holding probability of 1/2 is added to each update or swap move to guarantee nonnegative definiteness. The update move T chooses k uniformly from $\{0, \ldots, N\}$ and updates $x_{[k]}$ according to H_{β_k}:

$$T(x, dy) = \frac{1}{2(N+1)} \sum_{k=0}^{N} H_{\beta_k}(x_{[k]}, dy_{[k]}) \delta(x_{[-i]} - y_{[-i]}) \qquad x, y \in \mathcal{X}_{pt}$$

where $x_{[-i]} = (x_{[0]}, \ldots, x_{[i-1]}, x_{[i+1]}, \ldots, x_{[N]})$ and δ is Dirac's delta function.

The swap move Q attempts to exchange two of the levels via one of the following schemes:

SC1. sample k, l uniformly from $\{0, \ldots, N\}$ and propose exchanging the value of $x_{[k]}$ with that of $x_{[l]}$. The proposed state, denoted $(k, l)x$, is accepted according to the Metropolis criteria preserving π_{pt}:

$$\rho(x, (k, l)x) = \min \left\{ 1, \frac{\pi_{\beta_k}(x_{[l]}) \pi_{\beta_l}(x_{[k]})}{\pi_{\beta_k}(x_{[k]}) \pi_{\beta_l}(x_{[l]})} \right\}$$

SC2. sample k uniformly from $\{0, \ldots, N-1\}$ and propose exchanging $x_{[k]}$ and $x_{[k+1]}$, accepting with probability $\rho(x, (k, k+1)x)$.

Regardless of the choice of swapping scheme, both T and Q are reversible with respect to π_{pt} by construction, and nonnegative definite due to their 1/2 holding probability.

We define the parallel tempering chain $P_{pt} = QTQ$, which is also nonnegative definite and reversible with respect to π_{pt}, so the rate of convergence of the parallel tempering chain to π_{pt} may be bounded using the spectral gap of P_{pt}.

Observe that the definitions of Q and T can be generalized to use arbitrary densities $\phi_k \neq \pi_{\beta_k}$ by replacing H_{β_k} with any ϕ_k-reversible kernel T_k; we may specify the densities ϕ_k in any convenient way subject to $\phi_N = \pi$. The resulting chain is called a *swapping chain*, and its state space, measure, transition kernel, and associated stationary density will be denoted by $\mathcal{X}_{sc}$, λ_{sc}, P_{sc} and π_{sc}, respectively.

Rather than simulating parallel chains, a single chain can be augmented with a level index k, so that the state is $(z, k) \in \mathcal{X}_{st} = \mathcal{X} \otimes \{0, \ldots, N\}$ and the stationary density is

$$\pi_{st}(z, k) = \frac{1}{N+1}\phi_k(z) \qquad (z, k) \in \mathcal{X}_{st}.$$

The resulting *simulated tempering* chain (Marinari and Parisi 1992; Geyer and Thompson 1995) has two move types: T' samples $z \in \mathcal{X}$ according to T_k, conditional on k, while Q' attempts to change k via one of the following schemes:

ST1. propose a new level l uniformly from $\{0, \ldots, N\}$ and accept with probability $\min\left\{1, \frac{\phi_l(z)}{\phi_k(z)}\right\}$.

ST2. propose a move to $l = k - 1$ or $l = k + 1$ with equal probability and accept with the probability from scheme **ST1**, unless $l \in \{-1, N + 1\}$ in which case the move is rejected.

ST3. draw the level k from its conditional distribution given z.

Once again, a holding probability of 1/2 is added to both T' and Q'. The transition kernel of simulated tempering is then specified as $P_{st} = Q'T'Q'$. Here we do not require that ϕ_k are tempered versions of π, although this is the usual choice.

2.4 Example Target Distributions

2.4.1 Mean-Field Potts Model

The Potts model is from statistical physics. The Potts model and related models are used in Bayesian image analysis and for modeling spatial random effects (Banerjee et al. 2004; Geman and Geman 1984; Green and Richardson 2002).

We will consider the ferromagnetic mean-field Potts model with $q \in \{2, 3, \ldots\}$ colors and M sites, defined as follows for $z \in \{1, \ldots, q\}^M$:

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

with $\alpha \geq 0$. The marginal distribution of the number of sites in each color is shown in Figure 2.1. As shown in that figure, the distribution can be multimodal.

As we will see, the mean-field Potts model exhibits a phenomenon where a small change in the value of the parameter α near a critical value α_c causes a dramatic change in the asymptotic behavior of π in M. This phenomenon occurs in more general Potts models, so it is worthwhile to consider the mean-field Potts case. Consider

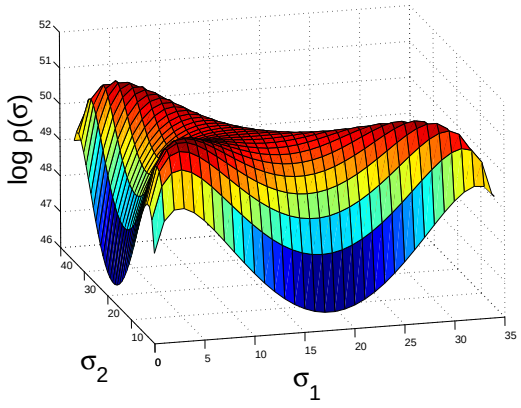

Figure 2.1: The marginal distribution of the number of sites with color 1 (σ_1) and color 2 (σ_2) for the mean-field Potts model with $q = 3$, $M = 35$, and $\alpha = 4\ln(2)$.

the proposal kernel S that proposes changing the color of a single site, where the site and color are drawn uniformly at random. We will analyze the convergence rate of parallel and simulated tempering in terms of the problem size M.

The mean-field Ising model is the mean-field Potts model with $q = 2$. It is commonly rewritten as follows for $w \in \mathcal{X} = \{-1, +1\}^M$:

$$\pi(z) = \frac{1}{Z} \exp\left\{ \frac{\alpha'}{2M} \left(\sum_{i=1}^{M} w_i \right)^2 \right\} \tag{2.3}$$

where $Z = \sum_w \exp\left\{\alpha'(\sum_i w_i)^2/(2M)\right\}$ and $\alpha' = \alpha/2$.

2.4.2 Mixtures of Normal Distributions

Many distributions in statistics are well-approximated by mixtures of normal distributions. In fact, any density function in $\mathbb{R}^M$ can be approximated arbitrarily well by

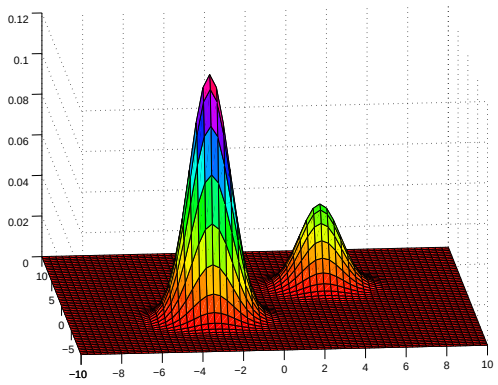

Figure 2.2: $\pi(z)$ as a function of z for a mixture of two normals in $\mathbb{R}^2$

a mixture of normal densities.

We will analyze the mixing of parallel and simulated tempering on various two-component mixtures of normal distributions in $\mathbb{R}^M$ in terms of M. For any length-M vector ν, $M \times M$ covariance matrix Σ, and $z \in \mathbb{R}^M$, let $N_M(z; \nu, \Sigma)$ be the density of a multivariate normal distribution in $\mathbb{R}^M$ with mean ν and covariance Σ, evaluated at z. We will consider mixtures of the form

$$\pi(z) = a_M N_M(z; \nu_M, \Sigma_M) + (1 - a_M) N_M(z; \nu'_M, \Sigma'_M)$$

where $a_M \in [0, 1]$. The density of one such mixture where $M = 2$ is shown in Figure 2.2. The Metropolis-Hastings proposal kernel S that we will use draws uniformly from the ball of radius M^{-1} centered at the current state.

Chapter 3

Lower Bounds on the Convergence Rates of Parallel and Simulated Tempering and Conditions for Rapid Mixing

As shown in the previous chapter, a lower bound on the spectral gap of a Markov chain implies a lower bound on the rate of convergence to stationarity. In this chapter we obtain lower bounds on the spectral gaps of parallel and simulated tempering chains, and show conditions for rapid mixing.

We are particularly interested in the case where Metropolis-Hastings with local proposals is torpidly mixing due to the multimodality of π. We partition the space into subsets on which the target is unimodal, one subset for each local mode. To guarantee rapid mixing of parallel tempering we require that the number of modes J is fixed (or at least bounded) in the problem size. We also require that each chain T_k is rapidly mixing when restricted to any one of the unimodal subsets, and that the

highest-temperature chain mixes rapidly among the subsets. Additionally, we require that the overlap of adjacent temperature levels is at most polynomially decreasing in the problem size, which is needed in order to mix rapidly among the levels. In the case where the modes are symmetric, these conditions imply that the set of parallel tempering chains is rapidly mixing. The same conditions imply the rapid mixing of simulated tempering in the symmetric case. In the general (asymmetric) case, these conditions are insufficient (see Chapters 5 and 6), so we give one additional condition which implies rapid mixing of both techniques. We require that a quantity related to the persistence (Section 5.1) of each unimodal subset decreases at most polynomially in the problem size.

3.1 Lower Bounds on the Spectral Gaps of Swapping and Simulated Tempering Chains and Conditions for Rapid Mixing

We will obtain lower bounds on the spectral gaps of swapping and simulated tempering chains. These will be used to prove a common set of conditions for rapid mixing of swapping and simulated tempering chains. The bounds are in terms of several quantities. Informally, the first quantity measures how well each chain T_k mixes when restricted to each unimodal subset. The second is how well the highest-temperature chain T_0 mixes among the subsets. The third is the overlap of the distributions of adjacent levels, and the fourth concerns the relative probability of a unimodal subset

at distinct (inverse) temperatures, for each subset.

In order to bound the spectral gap of a swapping or simulated tempering chain in terms of the mixing of the chain within each subset and the mixing of the chain among the subsets, we will use a state space decomposition result due to Caracciolo et al. (1992) and first published in Madras and Randall (2002). As in Madras and Randall (2002), we use the following definitions.

For any transition kernel P reversible with respect to a distribution μ and any subset A of the state space of P, define the restriction of P to A as

$$P|_A(x, B) = P(x, B) + \mathbf{1}_B(x)P(x, A^c) \qquad \text{for } x \in A,\ B \subset A. \tag{3.1}$$

where $\mathbf{1}_B$ is the indicator function for the set B. Note that $P|_A$ is reversible with respect to $\mu|_A$, the restriction of μ to A. Now take any partition $\mathcal{A} = \{A_j : j = 1, \ldots, J\}$ of the state space of P such that $\mu(A_j) > 0$ for all j, and define the projection matrix of P with respect to $\mathcal{A}$ to be

$$\bar{P}(i, j) = \frac{1}{\mu(A_i)} \int_{A_i} \int_{A_j} P(x, dy)\mu(dx) \qquad i, j \in \{1, \ldots, J\}. \tag{3.2}$$

Note that $\bar{P}$ is reversible with respect to the distribution on $j \in \{1, \ldots, J\}$ taking value $\mu(A_j)$.

Now consider a swapping chain defined as in Section 2.3 for some density of interest π on a state space $\mathcal{X}$, with ϕ_k-reversible transition kernels T_k, using the swapping scheme **SC2**. Let $\mathcal{A}$ be any partition of $\mathcal{X}$ such that $\phi_k[A_j] > 0$ for all k, j. The first quantity in our bound is the minimum over k and j of $\mathbf{Gap}(T_k|_{A_j})$, which

measures how well each chain T_k mixes within each partition element. The partition would typically be chosen so that this quantity is large; in our examples we choose $\mathcal{A}$ so that $\pi|_{A_j}$ is unimodal (has a single local mode) for each j. However, the results apply to any partition.

Next we consider how well the chain T_0 mixes among the partition elements. Let $\bar{T}_0$ to be the projection matrix of T_0 with respect to $\mathcal{A}$; the second quantity in our bound is $\mathbf{Gap}(\bar{T}_0)$.

The third quantity is the *overlap* of $\{\phi_k : k = 0, \ldots, N\}$ with respect to $\mathcal{A}$, defined as

$$\delta(\mathcal{A}) = \min_{\substack{|k-l|=1 \\ j \in \{1,\ldots,J\}}} \left[\int_{A_j} \min\left\{\phi_k(z), \phi_l(z)\right\} \lambda(dz) \right] / \phi_k[A_j] \tag{3.3}$$

The quantity $\delta(\mathcal{A})$ controls the rate of temperature changes in simulated tempering (scheme **ST2**). For the swapping chain, note that for any $i, j \in \{1, \ldots, J\}$ and any $k \in \{0, \ldots, N-1\}$, the marginal probability at stationarity of accepting a proposed swap between $x_{[k]} \in \mathcal{A}_i$ and $x_{[k+1]} \in \mathcal{A}_j$ is

$$\frac{\int\limits_{z \in A_i} \int\limits_{w \in A_j} \min\left\{\phi_k(z)\phi_{k+1}(w), \phi_k(w)\phi_{k+1}(z)\right\} \lambda(dw)\lambda(dz)}{\phi_k[A_i]\phi_{k+1}[A_j]} \geq \delta(\mathcal{A})^2. \tag{3.4}$$

We will show that our overlap quantity $\delta(\mathcal{A})$ is bounded below by the overlap used in Madras and Randall (2002) and Zheng (2003), and that our definition is equal to theirs in the case of π symmetric (as defined later in this section).

The fourth and final quantity concerns the probability of a single partition element

under ϕ_k, as a function of k, for each partition element:

$$\gamma(\mathcal{A}) = \min_{j\in\{1,\ldots,J\}} \prod_{k=1}^{N} \min\left\{1, \frac{\phi_{k-1}[A_j]}{\phi_k[A_j]}\right\}. \tag{3.5}$$

Note that for any $j \in \{1,\ldots,J\}$ and any $k,l \in \{0,\ldots,N\}$ such that $k<l$, $\phi_k[A_j] \geq \gamma(\mathcal{A})\phi_l[A_j]$. If $\phi_k[A_j]$ is monotonic as a function of k for each j, then $\gamma(\mathcal{A})$ simplifies to

$$\min_{j\in\{1,\ldots,J\}} \frac{\phi_0[A_j]}{\phi_N[A_j]}.$$

In this case, $\gamma(\mathcal{A})$ is equal to the minimum persistence of the partition elements, where persistence is defined in Section 5.1. We will also show that for π symmetric, $\gamma(\mathcal{A}) = 1$. With these definitions, the following theorem bounds the spectral gap of a swapping chain.

Theorem 3.1.1. *Take any swapping chain P_{sc} that uses the swapping scheme* **SC2** *(Section 2.3). Given any partition $\mathcal{A} = \{A_j : j = 1,\ldots,J\}$ of $\mathcal{X}$ such that $\phi_k[A_j] > 0$ for all k, j, and given $\delta(\mathcal{A})$ as in (3.3) and $\gamma(\mathcal{A})$ as in (3.5),*

$$\mathbf{Gap}(P_{sc}) \geq \left(\frac{\gamma(\mathcal{A})^{J+3}\delta(\mathcal{A})^2}{2^{12}(N+1)^4J^3}\right)\mathbf{Gap}(\bar{T}_0)\min_{k,j}\mathbf{Gap}(T_k|_{A_j}).$$

In particular, the bound holds for parallel tempering. Theorem 3.1.1 will be proven in Section 3.3. Note that for any I and set of constants $\{a_i : i = 1,\ldots,I\}$

and $\{b_i : i = 1, \ldots, I\}$, we have that $\min_i\{a_i/b_i\} \leq \sum_i a_i / \sum_i b_i$. Therefore

$$\begin{aligned}
\delta(\mathcal{A}) &= \min_{\substack{k\in\{0,\ldots,N-1\}\\ j\in\{1,\ldots,J\}}} \min\left\{ \frac{\int_{A_j} \min\{\phi_k(z), \phi_{k+1}(z)\}\lambda(dz)}{\phi_k[A_j]}, \frac{\int_{A_j} \min\{\phi_k(z), \phi_{k+1}(z)\}\lambda(dz)}{\phi_{k+1}[A_j]} \right\} \\
&\leq \min_{k\in\{0,\ldots,N-1\}} \frac{\sum_j \int_{A_j} \min\{\phi_k(z), \phi_{k+1}(z)\}\lambda(dz)}{\max\left\{\sum_j \phi_k[A_j], \sum_j \phi_{k+1}[A_j]\right\}} \qquad (3.6) \\
&= \min_{k\in\{0,\ldots,N-1\}} \int \min\{\phi_k(z), \phi_{k+1}(z)\}\lambda(dz).
\end{aligned}$$

The final expression for $\delta(\mathcal{A})$ is the definition of overlap that is used in Madras and Randall (2002) and Zheng (2003). Therefore Theorem 3 of Zheng (2003), along with our Theorem 3.1.1, implies the following bound for simulated tempering:

Corollary 3.1.1. *Take any simulated tempering chain P_{st} that changes levels using scheme* **ST3** *(Section 2.3). Given any partition $\mathcal{A} = \{A_j : j = 1, \ldots, J\}$ of $\mathcal{X}$ such that $\phi_k[A_j] > 0$ for all k, j,*

$$\mathbf{Gap}(P_{st}) \geq \left(\frac{\gamma(\mathcal{A})^{J+3}\delta(\mathcal{A})^3}{2^{14}(N+1)^5 J^3}\right) \mathbf{Gap}(\bar{T}_0) \min_{k,j} \mathbf{Gap}(T_k|_{A_j}).$$

Theorem 3.1.1 and Corollary 3.1.1 then imply a common condition for rapid mixing of swapping and simulated tempering chains. Recall from Section 2.1 that by rapid mixing we mean that the spectral gap decreases at most polynomially in the problem size. Then rapid mixing is implied by Theorem 3.1.1 and Corollary 3.1.1 if the following condition holds.

Condition 3.1.1. *J is fixed or bounded above in M. $\bar{T}_0$ is rapidly mixing. $\delta(\mathcal{A})$, $\gamma(\mathcal{A})$ and $\min_{k,j} \mathbf{Gap}(T_k|_{A_j})$ are polynomially decreasing in the problem size. N is polynomially bounded-above.*

Define π to be *symmetric* if there exists a partition $\{A_j : j = 1, \ldots, J\}$ of $\mathcal{X}$ such that for every pair of partition elements A_i, A_j there is some λ-measure-preserving bijection f_{ij} from A_i to A_j that preserves π. Note that when π is symmetric with respect to the partition $\mathcal{A}$, the inequality in (3.6) is an equality.

If π is symmetric with respect to $\mathcal{A}$, then $\pi_\beta[A_j] = \frac{1}{J}$ for any inverse temperature β. Therefore if the densities ϕ_k for the swapping or simulated tempering chain are chosen as tempered versions of π then $\gamma(\mathcal{A}) = 1$. In this case the restriction on $\gamma(\mathcal{A})$ in Condition 3.1.1 is automatically satisfied. Therefore for a symmetric distribution, in order to have rapid mixing of parallel and simulated tempering we require only that H_β (e.g. Metropolis-Hastings) mixes rapidly when restricted to each partition element, that H_β mixes rapidly among the partition elements for the smallest inverse temperature $\beta = \beta_0$, and that the overlap is polynomially decreasing using a polynomial number of inverse temperatures. These conditions holds for the mean-field Ising model and for a symmetric normal mixture model, as we will show in the next chapter.

3.2 Tools for Bounding Spectral Gaps

In Sections 3.2.1 and 3.2.2 we give some results from the literature and slight extensions thereon. These results will be used in Section 3.3 for the proof of Theorem 3.1.1.

3.2.1 A Bound for Finite State Space Markov Chains

We first consider a method for finite state space Markov chains. Let P and Q be Markov chain transition matrices on state space $\mathcal{X}$ with $|\mathcal{X}| < \infty$, reversible with respect to densities π_P and π_Q, respectively. Denote by $\mathcal{E}_P$ and $\mathcal{E}_Q$ the Dirichlet forms of P and Q, and let $E_P = \{(x, y) : \pi_P(x)P(x, y) > 0\}$ and $E_Q = \{(x, y) : \pi_Q(x)Q(x, y) > 0\}$ be the edge sets of P and Q, respectively.

For each pair $x \neq y$ such that $(x, y) \in E_Q$, fix a path $\gamma_{xy} = (x = x_0, x_1, x_2, \ldots, x_k = y)$ of length $|\gamma_{xy}| = k$ such that $(x_i, x_{i+1}) \in E_P$ for $i \in \{0, \ldots, k-1\}$. Define

$$c = \max_{(z,w)\in E_P} \left\{ \frac{1}{\pi_P(z)P(z, w)} \sum_{\gamma_{xy} \ni (z,w)} |\gamma_{xy}|\pi_Q(x)Q(x, y) \right\}.$$

Then we have the following result.

Theorem 3.2.1. *(Diaconis and Saloff-Coste 1993)*

$$\mathcal{E}_Q \leq c\mathcal{E}_P.$$

3.2.2 Bounds for General State Space Markov Chains

The following results hold for general state space transition kernels P and Q, reversible with respect to distributions μ_P and μ_Q on a space $\mathcal{X}$ with countably generated σ-algebra $\mathcal{F}$.

Theorem 3.2.2. *Let $\{A_j : j = 1, \ldots, J\}$ be any partition of $\mathcal{X}$ such that $\mu_P(A_j) > 0$ for all j. Define $P|_{A_j}$ as in (3.1) and $\bar{P}$ as in (3.2). For P nonnegative definite,*

$$\frac{1}{2}\mathbf{Gap}(\bar{P}) \min_{j=1,\ldots,J} \mathbf{Gap}(P|_{A_j}) \leq \mathbf{Gap}(P) \leq \mathbf{Gap}(\bar{P}).$$

The lower bound in Theorem 3.2.2 is a direct consequence of a result by Caracciolo et al. (1992) that was first published in Madras and Randall (2002), as described in Section 3.2.3. The proof of the upper bound, which is based on the proof of the result in Madras and Randall (2002), is also given in Section 3.2.3.

Theorem 3.2.3. *(Diaconis and Saloff-Coste 1996) Take any $N \in \mathbb{N}$ and let P_k, $k = 0, \ldots, N$, be μ_k-reversible transition kernels on state spaces $\mathcal{X}_k$. Let P be the transition kernel on $\mathcal{X} = \prod_k \mathcal{X}_k$ given by*

$$P(x, dy) = \left(\frac{1}{N+1}\right) \sum_{k=0}^{N} P_k(x_{[k]}, dy_{[k]})\, \delta(x_{[-k]} - y_{[-k]}) \qquad x, y \in \mathcal{X}$$

where δ is Dirac's delta function. P is called a product chain. It is reversible with respect to $\mu(dx) = \prod_k \mu_k(dx_{[k]})$, and

$$\mathbf{Gap}(P) = \frac{1}{N+1} \min_{k=0,\ldots,N} \mathbf{Gap}(P_k).$$

Lemma 3.2 of Diaconis and Saloff-Coste (1996) states this result for finite state spaces; however, the proof of that Lemma holds in the general case.

Lemma 3.2.1. *Let $\mu_P = \mu_Q$. If $Q(x, A\backslash\{x\}) \leq P(x, A\backslash\{x\})$ for every $x \in \mathcal{X}$ and every $A \subset \mathcal{X}$, then* $\mathbf{Gap}(Q) \leq \mathbf{Gap}(P)$.

Proof. As in Madras and Randall (2002), write $\mathbf{Gap}(P)$ in the form

$$\mathbf{Gap}(P) = \inf_{\substack{f \in L_2(\mu_P) \\ \mathrm{Var}_{\mu_P}(f) > 0}} \left(\frac{\int\int |f(x) - f(y)|^2 \mu_P(dx) P(x, dy)}{\int\int |f(x) - f(y)|^2 \mu_P(dx) \mu_P(dy)} \right).$$

and write $\mathbf{Gap}(Q)$ analogously. The result then follows immediately. □

3.2.3 Proof of the Decomposition Bound

We will prove Theorem 3.2.2 using the following results; consider the context of Section 3.2.2.

Theorem 3.2.4. *(Caracciolo, Pelissetto and Sokal 1992) Let $\mu_P = \mu_Q$. Assume that P is nonnegative definite and let $P^{\frac{1}{2}}$ be its nonnegative square root. Then*

$$\mathbf{Gap}(P^{\frac{1}{2}} Q P^{\frac{1}{2}}) \geq \mathbf{Gap}(\bar{P}) \min_{j=1,\ldots,J} \mathbf{Gap}(Q|_{A_j})$$

This result is due to Caracciolo et al. (1992) but was published in Madras and Randall (2002).

Lemma 3.2.2. *(Madras and Zheng 2003)*

$$\mathbf{Gap}(P) \geq \frac{1}{n}\mathbf{Gap}(P^n) \qquad \forall n \in \mathbb{N}$$

Note that although Madras and Zheng (2003) state this result for finite state spaces, their proof extends easily to general spaces.

Lemma 3.2.3. *(Madras and Zheng 2003)*

Assume that $\mu_P = \mu_Q$ and that P is nonnegative definite. Then

$$\mathbf{Gap}(QPQ) \geq \mathbf{Gap}(P)$$

Now consider the context of Theorem 3.2.2. The lower bound in Theorem 3.2.2 follows directly from Theorem 3.2.4 and Lemma 3.2.2:

$$\mathbf{Gap}(P) \geq \frac{1}{2}\mathbf{Gap}(P^2) = \frac{1}{2}\mathbf{Gap}(P^{\frac{1}{2}}PP^{\frac{1}{2}}) \geq \frac{1}{2}\mathbf{Gap}(\bar{P}) \min_j \mathbf{Gap}(P|_{A_j})$$

We now give the proof of the upper bound in Theorem 3.2.2, which is based on the proof of Theorem 3.2.4 given in Madras and Randall (2002). They define the operator $\bar{\Pi}$ on $f \in L_2(\mu_P)$ by

$$(\bar{\Pi}f)(x) = \int_{A_j} f(y)\mu_P(dy) \qquad \text{if } x \in A_j$$

and they define V_A to be the space of all $f \in L_2(\mu_P)$ such that f is constant within each set A_j. They note that the operator $\bar{\Pi}P\bar{\Pi}$ restricted to the M-dimensional

vector space V_A is equal to $\bar{P}$, and that as a consequence $\mathbf{Gap}(\bar{P}) = \mathbf{Gap}(\bar{\Pi}P\bar{\Pi})$. Using this observation, we can apply Lemma 3.2.3 to obtain $\mathbf{Gap}(\bar{P}) \geq \mathbf{Gap}(P)$.

3.3 Proof of the Lower Bound on the Spectral Gap of a Swapping Chain

3.3.1 Overview of the Proof

As in Madras and Zheng (2003), consider the space $\Sigma = \mathbb{Z}_J^{N+1}$ of possible assignments of levels to partition elements, and for $x = (x_{[0]}, \ldots, x_{[N]}) \in \mathcal{X}_{sc}$ let the signature $s(x)$ be the vector $(\sigma_0, \ldots, \sigma_N) \in \Sigma$ with

$$\sigma_k = j \text{ if } x_{[k]} \in A_j \qquad (0 \leq k \leq N).$$

For $\sigma \in \Sigma$, define

$$\mathcal{X}_\sigma = \{x \in \mathcal{X}_{sc} : s(x) = \sigma\}$$

so s induces a partition of $\mathcal{X}_{sc}$. Define $P_\sigma = P_{sc}|_{\mathcal{X}_\sigma}$, and let $\bar{P}_{sc}$ be the projection matrix of P_{sc} with respect to the partition $\{\mathcal{X}_\sigma\}$. Since P_{sc} is nonnegative definite, Theorem 3.2.2 gives

$$\mathbf{Gap}(P_{sc}) \geq \frac{1}{2}\mathbf{Gap}(\bar{P}_{sc}) \min_{\sigma \in \Sigma} \mathbf{Gap}(P_\sigma). \tag{3.7}$$

Theorem 3.1.1 then follows by deriving bounds on $\mathbf{Gap}(\bar{P}_{sc})$ and $\mathbf{Gap}(P_\sigma)$.

3.3.2 Bounding the Spectral Gap of P_σ

For $\sigma \in \Sigma$ consider the mixing of P_{sc} when restricted to the set $\mathcal{X}_\sigma$. If each of the chains T_k mixes well when restricted to the set A_{σ_k}, then the product chain T and thus P_{sc} will mix well when restricted to $\mathcal{X}_\sigma$. Let $T_\sigma = T|_{\mathcal{X}_\sigma}$ and note that for any $x, y \in \mathcal{X}_\sigma$ with $x \neq y$,

$$T_\sigma(x, dy) = \frac{1}{2(N+1)} \sum_{k=0}^{N} T_k|_{A_{\sigma_k}}(x_{[k]}, dy_{[k]})\ \delta(x_{[-k]} - y_{[-k]}).$$

Therefore T_σ is also a product chain, and Theorem 3.2.3 provides its spectral gap:

$$\begin{aligned}\mathbf{Gap}(T_\sigma) &= \frac{1}{2(N+1)} \min_{k \in \{0,\ldots,N\}} \mathbf{Gap}(T_k|_{A_{\sigma_k}}) \\ &\geq \frac{1}{2(N+1)} \min_{\substack{k \in \{0,\ldots,N\} \\ j \in \{1,\ldots,J\}}} \mathbf{Gap}(T_k|_{A_j}).\end{aligned}$$

Note that since $P_{sc} = QTQ$, and since Q has a $1/2$ holding probability,

$$P_\sigma(x, dy) \geq \frac{1}{4} T_\sigma(x, dy) \qquad \forall x, y \in \mathcal{X}_\sigma.$$

Using Lemma 3.2.1 we have $\mathbf{Gap}(P_\sigma) \geq \mathbf{Gap}(T_\sigma)/4$. Therefore

$$\mathbf{Gap}(P_\sigma) \geq \frac{1}{8(N+1)} \min_{\substack{k \in \{0,\ldots,N\} \\ j \in \{1,\ldots,J\}}} \mathbf{Gap}(T_k|_{A_j}). \tag{3.8}$$

3.3.3 Bounding the Spectral Gap of $\bar{P}_{sc}$

First note that $\bar{P}_{sc}$ is reversible with respect to the probability mass function

$$\pi^*(\sigma) \stackrel{\text{def}}{=} \pi_{sc}[\mathcal{X}_\sigma] = \prod_{k=0}^{N} \phi_k[A_{\sigma_k}] \qquad \forall \sigma \in \Sigma.$$

For any $\sigma, \tau \in \Sigma$, $\bar{P}_{sc}(\sigma, \tau)$ is the conditional probability at stationarity of moving to $\mathcal{X}_\tau$ under P_{sc}, given that the chain is currently in $\mathcal{X}_\sigma$:

$$\bar{P}_{sc}(\sigma, \tau) = \frac{1}{\pi_{sc}[\mathcal{X}_\sigma]} \int_{\mathcal{X}_\sigma} \int_{\mathcal{X}_\tau} \pi_{sc}(x) P_{sc}(x, dy) \lambda_{sc}(dx).$$

We will begin by bounding this probability in terms of the probability of moving to $\mathcal{X}_\tau$ under Q (a swap move) and the probability of moving to $\mathcal{X}_\tau$ under T (an update move). For swap moves, let $\bar{Q}$ be the projection matrix of Q with respect to $\{\mathcal{X}_\sigma : \sigma \in \Sigma\}$. Then for $k \in \{0, \ldots, N-1\}$ we have

$$\bar{P}_{sc}(\sigma, (k, k+1)\sigma) \geq \frac{1}{4} \bar{Q}(\sigma, (k, k+1)\sigma) \qquad \forall \sigma$$

where the right hand side is the conditional probability of swapping $x_{[k]}$ and $x_{[k+1]}$ under Q, and then holding twice. Similarly, for update moves we denote by $\bar{T}$ the projection matrix of T with respect to $\{\mathcal{X}_\sigma : \sigma \in \Sigma\}$, and denote $\sigma_{[i,j]} = (\sigma_0, \ldots, \sigma_{i-1}, j, \sigma_{i+1}, \ldots, \sigma_N)$. Then

$$\bar{P}_{sc}(\sigma, \sigma_{[i,j]}) \geq \frac{1}{4} \bar{T}(\sigma, \sigma_{[i,j]}) \qquad \forall i, j.$$

Therefore the Dirichlet form $\mathcal{E}_{sc}$ of $\bar{P}_{sc}$ evaluated at $f \in L_2(\pi^*)$ satisfies

$$\mathcal{E}_{sc}(f, f) \geq \frac{1}{4} \mathcal{E}_{\bar{Q}}(f, f) + \frac{1}{4} \mathcal{E}_{\bar{T}}(f, f).$$

Recalling that $\bar{T}_0$ is the projection matrix of T_0 with respect to $\mathcal{A}$, note that

$$
\begin{aligned}
&4(N+1)\mathcal{E}_{\bar{T}}(f,f) \\
&= 2(N+1)\sum_{\sigma,\tau\in\Sigma}(f(\sigma)-f(\tau))^2\pi^*(\sigma)\bar{T}(\sigma,\tau) \\
&\geq \sum_{\sigma\in\Sigma}\pi^*(\sigma)\sum_{j=1}^{J}(f(\sigma)-f(\sigma_{[0,j]}))^2\bar{T}_0(\sigma_0,j) \\
&= \sum_{\sigma_{1:N}}\left[\prod_{k=1}^{N}\phi_k[A_{\sigma_k}]\right]\sum_{i=1}^{J}\sum_{j=1}^{J}(f(i,\sigma_{1:N})-f(j,\sigma_{1:N}))^2\phi_0[A_i]\bar{T}_0(i,j) \\
&\geq \sum_{\sigma_{1:N}}\left[\prod_{k=1}^{N}\phi_k[A_{\sigma_k}]\right]\mathbf{Gap}(\bar{T}_0)\sum_{i=1}^{J}\sum_{j=1}^{J}(f(i,\sigma_{1:N})-f(j,\sigma_{1:N}))^2\phi_0[A_i]\phi_0[A_j] \\
&= \mathbf{Gap}(\bar{T}_0)\sum_{\sigma\in\Sigma}\pi^*(\sigma)\sum_{j=1}^{J}(f(\sigma)-f(\sigma_{[0,j]}))^2\phi_0[A_j]
\end{aligned}
$$

where the second inequality is by recognizing the Dirichlet form for $\bar{T}_0$. Therefore

$$
\frac{\mathcal{E}_{sc}(f,f)}{\mathbf{Gap}(\bar{T}_0)} \geq \left[\frac{1}{8}\mathcal{E}_{\bar{Q}}(f,f)+\sum_{\sigma\in\Sigma}\pi^*(\sigma)\sum_{j=1}^{J}(f(\sigma)-f(\sigma_{[0,j]}))^2\frac{\phi_0[A_j]}{16(N+1)}\right]. \tag{3.9}
$$

Now consider a transition kernel T^* constructed as follows: with probability $\frac{1}{2}$ transition according to $\bar{Q}$; otherwise with probability $\frac{1}{2(N+1)}$ draw $\sigma_{[0]}$ according to the distribution $\{\phi_0[A_j] : j = 1,\ldots,J\}$ (i.e. independent samples at the highest temperature); otherwise hold. Note that the Dirichlet form of T^* is precisely four times the right hand side of (3.9). Clearly T^* is also reversible with respect to π^*, so $\bar{P}_{sc}$ and T^* have the same stationary distribution. Therefore

$$
\mathbf{Gap}(\bar{P}_{sc}) \geq \frac{\mathbf{Gap}(T^*)\mathbf{Gap}(\bar{T}_0)}{4}. \tag{3.10}
$$

We will now bound $\mathbf{Gap}(T^*)$ by comparison with another π^*-reversible chain. Define the transition matrix T^{**} which chooses k uniformly from $\{0,\ldots,N\}$ and then

draws σ_k according to the distribution $\{\phi_k[A_j] : j = 1, \ldots, J\}$. Clearly T^{**} moves easily among the elements of Σ, and consequently has a large spectral gap as we will see. By combining (3.10) with a comparison of T^* to T^{**}, we will obtain a lower bound on the spectral gap of $\bar{P}_{sc}$.

Comparison of T^* to T^{**} will be done using Theorem 3.2.1. To simplify notation we write $\phi_k(j)$ as shorthand for $\phi_k[A_j]$ for the remainder of this section. Let j^* be the value of j that maximizes $\phi_N(j)$. For each edge $(\sigma, \sigma_{[i,j]})$ in the graph of T^{**} we define a path $\gamma_{\sigma,\sigma_{[i,j]}}$ in T^* with the following 7 stages:

1. Change σ_0 to j^*
2. Swap that j^* "up" to level i
3. Take the new σ_{i-1} (formerly σ_i) and swap it "down" to level 0
4. Change the value at level 0 to j (from former σ_i)
5. Swap the j at level 0 "up" to level i
6. Swap the j^* that is now at level $i-1$ "down" to level 0
7. Change the value at level 0 to σ_0 (from j^*)

In each path, skip all steps that do not change the state. Using the defined path set, we will obtain an upper bound c^* on the quantity c in Theorem 3.2.1. Since T^* and T^{**} both have stationary distribution π^*, Theorem 3.2.1 then yields $\mathbf{Gap}(T^*) \geq$

$\frac{1}{c^*}\mathbf{Gap}(T^{**})$. To bound c, we will use Propositions 3.3.1 and 3.3.2.

Proposition 3.3.1. *For the above-defined paths,*

$$\frac{\pi^*(\sigma)T^{**}(\sigma,\sigma_{[i,j]})}{\pi^*(\tau_1)T^*(\tau_1,\tau_2)} \leq 4J\,\gamma(\mathcal{A})^{-(J+3)}\delta(\mathcal{A})^{-2} \tag{3.11}$$

for all σ, i, and j, and any edge (τ_1,τ_2) in $\gamma_{\sigma,\sigma_{[i,j]}}$.

Proof. To obtain (3.11), first note that

$$\begin{aligned}\pi^*(\sigma)T^{**}(\sigma,\sigma_{[i,j]}) &= \frac{\phi_i(j)}{N+1}\left[\prod_{k=0}^{N}\phi_k(\sigma_k)\right] \\ &= \frac{1}{N+1}\min\{\pi^*(\sigma),\pi^*(\sigma_{[i,j]})\}\max\{\phi_i(\sigma_i),\phi_i(j)\}.\end{aligned}$$

For any state τ in the path $\gamma_{\sigma,\sigma_{[i,j]}}$ we will find a lower bound on $\pi^*(\tau)$ in terms of $\min\left\{\pi^*(\sigma),\pi^*(\sigma_{[i,j]})\right\}$. Consider the states in the path $\gamma_{\sigma,\sigma_{[i,j]}}$ up to stage 4 (where σ_i is at level 0). We will show that each state τ satisfies $\pi^*(\tau)\geq\pi^*(\sigma)\gamma(\mathcal{A})^{J+2}J^{-1}$. Then by symmetry the states from stage 4 to the end of the path satisfy $\pi^*(\tau)\geq\pi^*(\sigma_{[i,j]})\gamma(\mathcal{A})^{J+2}J^{-1}$.

Any state in stages 1 or 2 of the path from σ to $\sigma_{[i,j]}$ is of the following form for some $l\in\{0,\ldots,i\}$:

$$\tau=(\sigma_1,\ldots,\sigma_l,j^*,\sigma_{l+1},\ldots,\sigma_N).$$

Therefore

$$\begin{aligned}
\pi^*(\tau) &= \pi^*(\sigma)\left[\prod_{k=1}^{l}\frac{\phi_{k-1}(\sigma_k)}{\phi_k(\sigma_k)}\right]\frac{\phi_l(j^*)}{\phi_0(\sigma_0)} \\
&= \pi^*(\sigma)\left[\prod_{k=1}^{l}\prod_{m=1}^{J}\left[\mathrm{I}(\sigma_k = m)\frac{\phi_{k-1}(m)}{\phi_k(m)} + \mathrm{I}(\sigma_k \neq m)\right]\right]\frac{\phi_l(j^*)}{\phi_0(\sigma_0)} \\
&\geq \pi^*(\sigma)\left[\prod_{m=1}^{J}\prod_{k=1}^{N}\min\left\{1, \frac{\phi_{k-1}(m)}{\phi_k(m)}\right\}\right]\frac{\phi_l(j^*)}{\phi_0(\sigma_0)} \\
&\geq \pi^*(\sigma)\gamma(\mathcal{A})^{J+1}J^{-1}
\end{aligned}$$

where the last inequality uses the fact that by definition $\phi_N(j^*) \geq J^{-1}$, so $\phi_k(j^*) \geq \gamma(\mathcal{A})J^{-1}$ for all k and

$$\frac{\phi_l(j^*)}{\phi_0(\sigma_0)} \geq \frac{\gamma(\mathcal{A})J^{-1}}{\phi_0(\sigma_0)} \geq \gamma(\mathcal{A})J^{-1}.$$

Now consider the states in stage 3 of the path, the last of which is also the first state in stage 4. Any such state τ is of the form

$$\tau = (\sigma_1, \ldots, \sigma_l, \sigma_i, \sigma_{l+1}, \ldots, \sigma_{i-1}, j^*, \sigma_{i+1}, \ldots, \sigma_N)$$

for some $l \in \{0, \ldots, i-1\}$. Therefore

$$\begin{aligned}
\pi^*(\tau) &= \pi^*(\sigma)\left[\prod_{k=1}^{l}\frac{\phi_{k-1}(\sigma_k)}{\phi_k(\sigma_k)}\right]\frac{\phi_i(j^*)\phi_l(\sigma_i)}{\phi_0(\sigma_0)\phi_i(\sigma_i)} \\
&\geq \pi^*(\sigma)\gamma(\mathcal{A})^{J+1}J^{-1}\frac{\phi_l(\sigma_i)}{\phi_i(\sigma_i)} \;\geq\; \pi^*(\sigma)\gamma(\mathcal{A})^{J+2}J^{-1}
\end{aligned}$$

where the last step is because $l < i$. Putting the above together, we have that for all τ in the path from σ to $\sigma_{[i,j]}$

$$\pi^*(\tau) \geq \min\{\pi^*(\sigma), \pi^*(\sigma_{[i,j]})\}\gamma(\mathcal{A})^{J+2}J^{-1}.$$

We use this to obtain (3.11) as follows: for any edge (τ_1, τ_2) on the path $\gamma_{\sigma,\sigma_{[i,j]}}$, we have either $\tau_2 = (k, k+1)\tau_1$ for some k, or $\tau_2 = \tau_{1[0,m]}$ for some m. The probability of proposing the swap $\tau_2 = (k, k+1)\tau_1$ according to Q is $\frac{1}{2N}$. Recall that for any $l_1, l_2 \in \{1, \ldots, J\}$, $\delta(\mathcal{A})^2$ is a lower bound for the marginal probability at stationarity of accepting a proposed swap between $x_k \in A_{l_1}$ and $x_{k+1} \in A_{l_2}$. Thus we have $\bar{Q}(\tau_1, \tau_2) \geq \delta(\mathcal{A})^2/(2N)$, and so

$$\begin{aligned}
\frac{\pi^*(\sigma)T^{**}(\sigma, \sigma_{[i,j]})}{\pi^*(\tau_1)T^*(\tau_1, \tau_2)} &= \frac{\pi^*(\sigma)\phi_i(j)}{(N+1)\pi^*(\tau_1)T^*(\tau_1, \tau_2)} \\
&\leq \frac{2\pi^*(\sigma)\phi_i(j)}{(N+1)\pi^*(\tau_1)\bar{Q}(\tau_1, \tau_2)} \leq \frac{4\pi^*(\sigma)\phi_i(j)}{\pi^*(\tau_1)\delta(\mathcal{A})^2} \\
&= \frac{4\min\{\pi^*(\sigma), \pi^*(\sigma_{[i,j]})\}\max\{\phi_i(j), \phi_i(\sigma_i)\}}{\pi^*(\tau_1)\delta(\mathcal{A})^2} \\
&\leq \frac{4J\max\{\phi_i(j), \phi_i(\sigma_i)\}}{\gamma(\mathcal{A})^{J+2}\delta(\mathcal{A})^2} \leq \frac{4J}{\gamma(\mathcal{A})^{J+2}\delta(\mathcal{A})^2}.
\end{aligned} \tag{3.12}$$

In the case that, instead, $\tau_2 = \tau_{1[0,m]}$ for some m, (3.12) becomes

$$\frac{\pi^*(\sigma)T^{**}(\sigma, \sigma_{[i,j]})}{\pi^*(\tau_1)T^*(\tau_1, \tau_2)} = \frac{4\pi^*(\sigma)\phi_i(j)}{\pi^*(\tau_1)\phi_0(m)} \tag{3.13}$$

and there are three possible cases: the edge (τ_1, τ_2) could be stage 1, stage 4, or stage 7 of $\gamma_{\sigma,\sigma_{[i,j]}}$. If it is stage 1, then (3.13) is bounded by

$$\frac{4\pi^*(\sigma)\phi_i(j)}{\pi^*(\sigma)\phi_0(j^*)} \leq \frac{4}{\phi_0(j^*)} \leq \frac{4J}{\gamma(\mathcal{A})}.$$

If the move is stage 4, then (3.13) is bounded by

$$\begin{aligned}
\frac{4\pi^*(\sigma)\phi_i(j)}{\pi^*(\tau_1)\phi_0(j)} &= \frac{4\min\{\pi^*(\sigma), \pi^*(\sigma_{[i,j]})\}\max\{\phi_i(j), \phi_i(\sigma_i)\}}{\min\{\pi^*(\tau_1), \pi^*(\tau_2)\}\max\{\phi_0(j), \phi_0(\sigma_i)\}} \\
&\leq \frac{4}{\gamma(\mathcal{A})}\frac{\min\{\pi^*(\sigma), \pi^*(\sigma_{[i,j]})\}}{\min\{\pi^*(\tau_1), \pi^*(\tau_2)\}} \leq 4J\gamma(\mathcal{A})^{-(J+3)}
\end{aligned}$$

since $\phi_i(j) \leq \frac{\phi_0(j)}{\gamma(\mathcal{A})} \leq \frac{\max\{\phi_0(j),\phi_0(\sigma_i)\}}{\gamma(\mathcal{A})}$ and $\phi_i(\sigma_i) \leq \frac{\phi_0(\sigma_i)}{\gamma(\mathcal{A})} \leq \frac{\max\{\phi_0(j),\phi_0(\sigma_i)\}}{\gamma(\mathcal{A})}$. Finally, if the move is stage 7, then (3.13) is bounded by

$$\frac{4\pi^*(\sigma)\phi_i(j)}{\pi^*(\sigma_{[i,j]})\phi_0(j^*)} = \frac{4\pi^*(\sigma_{[i,j]})\phi_i(\sigma_i)}{\pi^*(\sigma_{[i,j]})\phi_0(j^*)} \leq \frac{4J}{\gamma(\mathcal{A})}.$$

The result (3.11) follows for any edge (τ_1, τ_2) on the path from σ to $\sigma_{[i,j]}$. □

We also have the following result.

Proposition 3.3.2. *For the above-defined paths,*

$$\sum_{\gamma_{\sigma,\sigma_{[i,j]}} \ni (\tau_1,\tau_2)} |\gamma_{\sigma,\sigma_{[i,j]}}| \leq 16(N+1)^2 J^2 \tag{3.14}$$

for any edge (τ_1, τ_2) *in the graph of* T^*.

Proof. We will bound the number of paths $\gamma_{\sigma,\sigma_{[i,j]}}$ that go through any edge (τ_1, τ_2), and the length of any such path.

Consider the set of paths for which the edge is in stage 1 of the path. Then $\tau_1 = \sigma$ and $\tau_2 = \sigma_{[0,j^*]}$, and since $i \in \{0, \ldots, N\}$ and $j \in \{1, \ldots, J\}$, there are no more than $(N+1)J$ such paths. Similarly, there are no more than $(N+1)J$ paths for which the edge is in stage 4 of the path.

Now consider the set of paths for which the edge is in stage 2 of the path. Then we must have $\tau_1 = (\sigma_1, \ldots, \sigma_l, j^*, \sigma_{l+1}, \ldots, \sigma_N)$ for some $l \in \{0, \ldots, i-1\}$ and $\tau_2 = (l, l+1)\tau_1$. σ_0 is unknown but has only J possible values, so with i,j unknown there are no more than $(N+1)J^2$ such paths. Similarly, there are no more than $(N+1)J^2$ paths for which the edge is in stage 3 of the path.

If the edge has $\tau_2 = (k, k+1)\tau_1$ for some k, then it can only be in stages 2,3,5, or 6 of the path, while if $\tau_2 = \tau_{1[0,m]}$ for some m then it can only be in stages 1,4, or 7. Since the edge can be in at most 4 stages, each with at most $(N+1)J^2$ paths, the total number of paths containing any edge is no more than $4(N+1)J^2$. Each of these paths has length at most $4N + 3 < 4(N+1)$, so (3.14) follows. □

Combining Propositions 3.3.1 and 3.3.2, we obtain an upper bound on the constant c in Theorem 3.2.1:

$$c \leq \frac{2^6(N+1)^2J^3}{\gamma(\mathcal{A})^{J+3}\delta(\mathcal{A})^2}$$

and recalling that both T^* and T^{**} have stationary distribution π^*, application of Theorem 3.2.1 yields

$$\mathbf{Gap}(T^*) \geq \frac{\gamma(\mathcal{A})^{J+3}\delta(\mathcal{A})^2}{2^6(N+1)^2J^3}\mathbf{Gap}(T^{**}).$$

Now since T^{**} is a product chain whose ϕ_k-reversible component chains each have spectral gap 1 by definition (2.1), Theorem 3.2.3 gives $\mathbf{Gap}(T^{**}) \geq (N+1)^{-1}$ and we have

$$\mathbf{Gap}(T^*) \geq \frac{\gamma(\mathcal{A})^{J+3}\delta(\mathcal{A})^2}{2^6(N+1)^3J^3}.$$

Then we obtain the bound for $\mathbf{Gap}(\bar{P}_{sc})$ from (3.10):

$$\mathbf{Gap}(\bar{P}_{sc}) \geq \frac{\mathbf{Gap}(T^*)\mathbf{Gap}(\bar{T}_0)}{4} \geq \left(\frac{\gamma(\mathcal{A})^{J+3}\delta(\mathcal{A})^2}{2^8(N+1)^3J^3}\right)\mathbf{Gap}(\bar{T}_0). \tag{3.15}$$

Using (3.7), (3.8), and (3.15) then proves Theorem 3.1.1.

Chapter 4

Multimodal Distributions for which Parallel and Simulated Tempering are Rapidly Mixing

We will give several bimodal distributions for which parallel and simulated tempering are rapidly mixing as implied by Theorem 3.1.1 and Corollary 3.1.1. The first two are symmetric, so that $\gamma(\mathcal{A}) = 1$, and the last is not, so we show that $\gamma(\mathcal{A})$ is polynomially decreasing in the problem size.

4.1 Rapid Mixing on the Mean-Field Ising Model

Recall from Section 2.4.1 the definition of the mean field Ising model and the corresponding proposal kernel S. Taking $N = M$, $\beta_k = k/N$, and T_k equal to Metropolis-Hastings for S with respect to $\phi_k = \pi_{\beta_k}$, it is shown in Madras and Zheng (2003)

that parallel and simulated tempering are rapidly mixing. We will show that this is also a consequence of Theorem 3.1.1 and Corollary 3.1.1.

As in Madras and Zheng (2003), partition $\mathcal{X}$ into $A_1 = \{z \in \mathcal{X} : \sum_i z_i < 0\}$ and $A_2 = \{z \in \mathcal{X} : \sum_i z_i \geq 0\}$. Restricting to M odd, the density π is clearly symmetric with respect to the partition $\{A_1, A_2\}$.

Since π_{β_0} is uniform, $T_0 = S$. Note that S is a product chain as defined in Theorem 3.2.3, composed of M chains that each have spectral gap equal to 1. By Theorem 3.2.3, $\mathbf{Gap}(T_0) = 1/M$. Since T_0 is rapidly mixing, so is $\bar{T}_0$ (Theorem 3.2.2).

It is shown in Madras and Zheng (2003) that the minimum over k and j of $\mathbf{Gap}(T_k|_{A_j})$ is polynomially decreasing. Also note that for any $z \in \mathcal{X}$ and any $k \in \{0, \ldots, N-1\}$,

$$\pi(z)^{\beta_{k+1}-\beta_k} \;=\; \pi(z)^{\frac{1}{M}} \;\in\; \left[\frac{1}{Z^{1/M}}, \frac{1}{Z^{1/M}} \exp\{\alpha'/2\}\right].$$

Therefore

$$\begin{aligned}\frac{\phi_{k+1}(z)}{\phi_k(z)} &= \pi(z)^{\beta_{k+1}-\beta_k} \left(\frac{\sum_{w\in\mathcal{X}} \pi(w)^{\beta_k}}{\sum_{w\in\mathcal{X}} \pi(w)^{\beta_{k+1}}}\right) \\ &\in [\exp\{-\alpha'/2\}, \exp\{\alpha'/2\}]\end{aligned}$$

which implies that

$$\begin{aligned}\sum_{z\in\mathcal{X}} \min\{\phi_k(z), \phi_{k+1}(z)\} &= \sum_{z\in\mathcal{X}} \phi_k(z) \min\left\{1, \frac{\phi_{k+1}(z)}{\phi_k(z)}\right\} \\ &\geq \exp\{-\alpha'/2\}.\end{aligned}$$

Recalling that for π symmetric, the inequality in (3.6) is an equality, $\delta(\mathcal{A})$ is bounded below by a constant for all M. Therefore by Theorem 3.1.1 and Corollary 3.1.1, this set of parallel or simulated tempering chains is rapidly mixing.

4.2 Rapid Mixing on a Symmetric Mixture of Normals in $\mathbb{R}^M$

Recall the definitions from Section 2.4.2. Let 1_M denote the vector of M ones, and I_M denote the $M \times M$ identity matrix. Take any $b > 0$, and consider the following mixture of two normal densities in $\mathbb{R}^M$:

$$\pi(z) = \frac{1}{2} N_M(z; -b1_M, \mathrm{I}_M) + \frac{1}{2} N_M(z; b1_M, \mathrm{I}_M). \tag{4.1}$$

Observe that π is symmetric with respect to the partition of the state space defined by $A_1 = \{z : \sum_i z_i < 0\}$ and $A_2 = \{z : \sum_i z_i \geq 0\}$. We will consider in this section the following approximation to π, extending the same results to π in Section 4.4:

$$\tilde{\pi}(z) \propto \frac{1}{2} N_M(z; -b1_M, \mathrm{I}_M)\mathbf{1}_{A_1}(z) + \frac{1}{2} N_M(z; b1_M, \mathrm{I}_M)\mathbf{1}_{A_2}(z) \tag{4.2}$$

where $\mathbf{1}$ is the indicator function for a set. Recall that the proposal kernel S proposes uniformly on the ball of radius M^{-1} centered at the current state. Metropolis-Hastings for S with respect to the density

$$\tilde{\pi}|_{A_1}(z) \propto N_M(z; -b1_M, \mathrm{I}_M)\mathbf{1}_{A_1}(z)$$

or with respect to

$$\tilde{\pi}|_{A_2}(z) \propto N_M(z; b1_M, \mathrm{I}_M)\mathbf{1}_{A_2}(z)$$

is rapidly mixing in M, as we will show in Section 4.3.4. We will also show that Metropolis-Hastings for S with respect to $\tilde{\pi}$ is torpidly mixing (Section 4.3.2). In Section 4.3 we will show that for any β,

$$\tilde{\pi}_\beta(z) = \frac{N_M(z; -b1_M, \beta^{-1}\mathrm{I}_M)\mathbf{1}_{A_1}(z)}{2\Phi(b\sqrt{M}\beta^{1/2})} + \frac{N_M(z; b1_M, \beta^{-1}\mathrm{I}_M)\mathbf{1}_{A_2}(z)}{2\Phi(b\sqrt{M}\beta^{1/2})}$$

where Φ is the cumulative normal distribution function in one dimension. All these facts will be shown for a weighted normal mixture of which (4.2) is a special case. Set $N = M$ and $\beta_k = M^{-(M-k)/M}$ (a geometric progression), and let T_k be the Metropolis-Hastings kernel for S with respect to $\phi_k = \tilde{\pi}_{\beta_k}$. With these specifications, parallel and simulated tempering are rapidly mixing, as implied by the following additional facts. The quantity $\min_{j,k} \mathbf{Gap}(T_k|_{A_j})$ is polynomially decreasing in M, as we will show in Section 4.3.4. Metropolis-Hastings for S with respect to $\tilde{\pi}_{M^{-1}}$ is rapidly mixing (Section 4.3.3). The quantity $\delta(\mathcal{A})$ is also polynomially decreasing, shown as follows.

Let λ be Lebesgue measure in $\mathbb{R}^M$. Take any M and any $k \in \{0, \ldots, N-1\}$.

Note that $\beta_k/\beta_{k+1} = M^{-1/M}$, so that

$$\begin{aligned}
&\int_{\mathcal{X}} \min\left\{\phi_k(z), \phi_{k+1}(z)\right\} \lambda(dz) \;=\; 2\int_{A_2} \min\left\{\phi_k(z), \phi_{k+1}(z)\right\} \lambda(dz) \\
&= \int_{A_2} \min\left\{ \frac{N_M(z; b1_M, \beta_k^{-1}\mathrm{I}_M)}{\Phi(b\sqrt{M}{\beta_k}^{1/2})}, \frac{N_M(z; b1_M, \beta_{k+1}^{-1}\mathrm{I}_M)}{\Phi(b\sqrt{M}\beta_{k+1}^{1/2})} \right\} \lambda(dz) \\
&\geq \int_{A_2} \min\left\{N_M(z; b1_M, \beta_k^{-1}\mathrm{I}_M), N_M(z; b1_M, \beta_{k+1}^{-1}\mathrm{I}_M)\right\} \lambda(dz) \\
&= (2\pi)^{-M/2} \int_{A_2} \beta_{k+1}^{M/2} \times \\
&\min\left\{ \left(\frac{\beta_k}{\beta_{k+1}}\right)^{M/2} \exp\left\{-\frac{\beta_k}{2}\sum_i (z_i - b)^2\right\}, \exp\left\{-\frac{\beta_{k+1}}{2}\sum_i (z_i - b)^2\right\} \right\} \lambda(dz) \\
&\geq \frac{1}{\sqrt{M}} (2\pi)^{-M/2} \int_{A_2} \beta_{k+1}^{M/2} \exp\left\{-\frac{\beta_{k+1}}{2}\sum_i (z_i - b)^2\right\} \lambda(dz) \\
&= \frac{1}{\sqrt{M}} \int_{A_2} N_M(z; b1_M, \beta_{k+1}^{-1}\mathrm{I}_M) \quad \geq \frac{1}{2\sqrt{M}}.
\end{aligned} \tag{4.3}$$

Therefore $\delta(\mathcal{A})$ is polynomially decreasing in M. By Theorem 3.1.1 and Corollary 3.1.1, parallel and simulated tempering with this N and this set of inverse temperatures are rapidly mixing.

In the case of a normal mixture with unequal weights, we must additionally show that $\gamma(\mathcal{A})$ is polynomially decreasing, and we will do this in Section 4.3. We will also need to use more inverse temperatures in that case in order to show that $\delta(\mathcal{A})$ is polynomially decreasing.

4.3 Rapid Mixing on a Weighted Mixture of Normals in $\mathbb{R}^M$

Recall the definitions from Section 2.4.2. Let 1_M denote the vector of M ones, and I_M denote the $M \times M$ identity matrix. Take any $b > 0$, and any sequence $a_1, a_2, \ldots$ where $a_M \in [1/2, 1)$ for all M. As a generalization of the symmetric normal mixture defined in (4.1), consider the following mixture of two normal densities in $\mathbb{R}^M$:

$$\pi(z) = a_M N_M(z; -b1_M, \mathrm{I}_M) + (1 - a_M) N_M(z; b1_M, \mathrm{I}_M).$$

As in the symmetric case, partition $\mathcal{X}$ into $A_1 = \{z : \sum_i z_i < 0\}$ and $A_2 = \{z : \sum_i z_i \geq 0\}$. For technical reasons, we will use the following approximation to π:

$$\tilde{\pi}(z) \propto a_M N_M(z; -b1_M, \mathrm{I}_M)\mathbf{1}_{A_1}(z) + (1 - a_M) N_M(z; b1_M, \mathrm{I}_M)\mathbf{1}_{A_2}(z) \tag{4.4}$$

and will restrict $a_M/(1-a_M)$ to be exponentially bounded-above, meaning that there is some $c \geq 1$ such that $a_M/(1 - a_M) \leq c^M$ for all M.

Recall that S is the proposal kernel that is uniform on the ball of radius M^{-1} centered at the current state. Metropolis-Hastings for S with respect to the density

$$\tilde{\pi}|_{A_1}(z) \propto N_M(z; -b1_M, \mathrm{I}_M)\mathbf{1}_{A_1}(z)$$

or with respect to

$$\tilde{\pi}|_{A_2}(z) \propto N_M(z; b1_M, \mathrm{I}_M)\mathbf{1}_{A_2}(z)$$

is rapidly mixing in M, as we will show in Section 4.3.4. However, Metropolis-Hastings for S with respect to $\tilde{\pi}$ is torpidly mixing, which we will show in Section 4.3.2.

Note that for any β,

$$\tilde{\pi}_\beta(z) \propto a_M^\beta N_M(z; -b1_M, \beta^{-1}\mathrm{I}_M)\mathbf{1}_{A_1}(z) + (1-a_M)^\beta N_M(z; b1_M, \beta^{-1}\mathrm{I}_M)\mathbf{1}_{A_2}(z) \tag{4.5}$$

Let Φ be the cumulative normal distribution function in one dimension. Consider any normal distribution with covariance $\sigma^2\mathrm{I}_M$ for $\sigma > 0$. Observe that the probability under this normal distribution of any half-space that is Euclidean distance d from the center of the normal distribution at its closest point is $\Phi(-d/\sigma)$. This is due to the independence of the dimensions and can be shown by a rotation and scaling in $\mathbb{R}^M$.

Note that the distance between the half-space A_2 and the point $-b1_M$ is equal to $b\sqrt{M}$, as is the distance between A_1 and $b1_M$. Therefore the normalizing constant of (4.5) is $[a_M^\beta + (1-a_M)^\beta]\Phi(b\sqrt{M}\beta^{1/2})$.

Consider the inverse temperature specification that we used for the symmetric case, with $N = M$ and the set of inverse temperatures $\{M^{-(M-k)/M} : k = 0, \ldots, M\}$. Also recall the inverse temperature specification for the mean-field Ising model, with $N = M$ and the set of inverse temperatures $\{k/M : k = 0, \ldots, M\}$. For the mixture of normals with unequal weights, we use both, taking the set of inverse temperatures $\{M^{-(M-k)/M} : k = 0, \ldots, M\} \cup \{k/M : k = 1, \ldots, M\}$, so that $N = 2M$.

Metropolis-Hastings for S with respect to $\tilde{\pi}_{M^{-1}}$ is rapidly mixing, as we will show in Section 4.3.3. In addition, $\min_{j,k} \mathbf{Gap}(T_k|_{A_j})$ is polynomially decreasing in M, to be proven in Section 4.3.4. The quantities $\gamma(\mathcal{A})$ and $\delta(\mathcal{A})$ are also polynomially

decreasing, shown as follows.

Note that for any β,

$$\tilde{\pi}_\beta[A_1] = \frac{a_M^\beta}{a_M^\beta + (1-a_M)^\beta}$$

which is an increasing function of β since $a_M \geq 1/2$. Therefore

$$\gamma(\mathcal{A}) \;=\; \frac{\phi_0[A_1]}{\phi_N[A_1]} \;\geq\; \frac{1}{2\phi_N[A_1]} \;\geq\; \frac{1}{2}$$

which does not depend on M. Also note that for any $k \in \{0, \ldots, N-1\}$,

$$\begin{aligned}\frac{\phi_{k+1}[A_2]}{\phi_k[A_2]} \geq \frac{\phi_k[A_1]\phi_{k+1}[A_2]}{\phi_{k+1}[A_1]\phi_k[A_2]} \;&=\; \left(\frac{1-a_M}{a_M}\right)^{\beta_{k+1}-\beta_k} \\ &\geq \left(\frac{1-a_M}{a_M}\right)^{1/M} \;\geq\; c^{-1}.\end{aligned}$$

Therefore

$$\begin{aligned}&\frac{\int_{A_2} \min\left\{\phi_k(z), \phi_{k+1}(z)\right\} \lambda(dz)}{\max\{\phi_k[A_2], \phi_{k+1}[A_2]\}} \\ &= \frac{\int_{A_2} \min\left\{\phi_k[A_2]\frac{N_M(z;b1_M,\beta_k^{-1}I_M)}{\Phi(b\sqrt{M}\beta_k^{1/2})}, \phi_{k+1}[A_2]\frac{N_M(z;b1_M,\beta_{k+1}^{-1}I_M)}{\Phi(b\sqrt{M}\beta_{k+1}^{1/2})}\right\} \lambda(dz)}{\max\left\{\phi_k[A_2], \phi_{k+1}[A_2]\right\}} \\ &\geq \frac{\phi_{k+1}[A_2]\int_{A_2} \min\left\{N_M(z;b1_M,\beta_k^{-1}I_M), N_M(z;b1_M,\beta_{k+1}^{-1}I_M)\right\} \lambda(dz)}{\phi_k[A_2]} \\ &\geq c^{-1}\int_{A_2} \min\left\{N_M(z;b1_M,\beta_k^{-1}I_M), N_M(z;b1_M,\beta_{k+1}^{-1}I_M)\right\} \lambda(dz) \\ &\geq \frac{1}{2\sqrt{M}}c^{-1}\end{aligned}$$

where the last inequality is from (4.3), using the fact that $\beta_k/\beta_{k+1} \geq M^{-1/M}$. This argument can be repeated for A_1. Note that

$$\delta(\mathcal{A}) = \min_{k\in\{0,\ldots,N-1\},j} \frac{\int_{A_j} \min\left\{\phi_k(z), \phi_{k+1}(z)\right\} \lambda(dz)}{\max\{\phi_k[A_j], \phi_{k+1}[A_j]\}}.$$

Therefore $\delta(\mathcal{A})$ is polynomially decreasing. By Theorem 3.1.1 and Corollary 3.1.1, parallel and simulated tempering with this N and this set of inverse temperatures are rapidly mixing on the weighted mixture of normals.

4.3.1 Tools for Bounding Spectral Gaps

We will need the following tools to prove the results of this section. These results hold for any transition kernels P and Q defined on a space $\mathcal{X}$ with measure λ.

Theorem 4.3.1. *(Lawler and Sokal 1988) Assume that P is reversible with respect to a distribution μ on $\mathcal{X}$. The conductance of a set $A \subset \mathcal{X}$ for which $0 < \mu(A) < 1$ is defined as:*

$$\Phi_P(A) = \frac{(\mathbf{1}_A, P\mathbf{1}_{A^c})_\mu}{\mu(A)\mu(A^c)}$$

where $\mathbf{1}_A$ is the indicator function of the set A. Also define the conductance of P to be $\Phi_P = \inf_{A \subset \mathcal{X}: 0 < \mu(A) < 1} \Phi_P(A)$. Then

$$\frac{\Phi_P^2}{8} \leq \mathbf{Gap}(P) \leq \Phi_P.$$

Note that since P is reversible with respect to μ, $\Phi_P(A)$ can be rewritten as

$$\Phi_P(A) = \frac{(\mathbf{1}_A, P\mathbf{1}_{A^c})_\mu}{\mu(A)} + \frac{(\mathbf{1}_{A^c}, P\mathbf{1}_A)_\mu}{\mu(A^c)} \tag{4.6}$$

In order to show that T_0 is rapidly mixing and that $\min_{j,k} \mathbf{Gap}(T_k|_{A_j})$ is polynomially decreasing, we will need the following results regarding the mixing of Metropolis-

Hastings. The first regards the mixing of Metropolis-Hastings with respect to a mixture density.

Theorem 4.3.2. *(Madras and Randall 2002) Let π_P be some mixture density on $\mathcal{X}$, so that there exists an $I \in \mathbb{N}$ and a set of weights a_i and densities ψ_i (with respect to λ) for $i \in \{1, \ldots, I\}$ such that for all $z \in \mathcal{X}$*

$$\pi_P(z) = \sum_{i=1}^{I} a_i \psi_i(z).$$

Let P be the Metropolis-Hastings chain for some proposal kernel Q with respect to π_P. For each i let $\mathbf{Gap}_i$ denote the spectral gap of the Metropolis-Hastings chain for Q with respect to ψ_i. Define the overlap quantity

$$d = \min_{i=1,\ldots,I-1} \int \min\{\psi_i(z), \psi_{i+1}(z)\} \lambda(dz).$$

Then

$$\mathbf{Gap}(P) \geq \frac{d}{2(I-1)} \min_{i=1,\ldots,I} a_i \mathbf{Gap}_i.$$

We will also need a result regarding the mixing of Metropolis-Hastings with respect to "similar" densities:

Lemma 4.3.1. *(Madras and Piccioni 1999) Assume that P and Q are Metropolis-Hastings kernels for the same symmetric proposal kernel, with respect to the densities π_P and π_Q, respectively. Also assume that there exists an $a \geq 1$ such that*

$$a^{-1} \leq \frac{\pi_P(z)}{\pi_Q(z)} \leq a$$

for all $z \in \mathcal{X}$ such that $\pi_P(z)$ and $\pi_Q(z)$ do not vanish simultaneously. Then

$$a^{-2}Gap(P) \leq Gap(Q) \leq a^2 Gap(P)$$

The last two results regard the mixing of Metropolis-Hastings with respect to a normal distribution in $\mathbb{R}^M$ as M increases. They are consequences of results in Kannan and Li (1996a) and Kannan and Li (1996b), as shown in Section 4.5.

Theorem 4.3.3. *Take any $M \in \mathbb{N}$ and any $\sigma, \tau > 0$. Let S' be the proposal kernel in $\mathbb{R}^M$ that draws uniformly from the ball of radius σ centered at the current state. Consider the Metropolis-Hastings kernel for S' with respect to any normal density with covariance matrix $\tau^2 I_M$. The conductance Φ_{MH} of the kernel satisfies*

$$\Phi_{MH} \geq \frac{\Phi_{loc}^2 \sigma}{2^{3/2}\tau\sqrt{M\pi}}$$

for a quantity Φ_{loc} that satisfies

$$\Phi_{loc} \geq \min\left\{\exp\left\{-\frac{2\sqrt{M}\sigma^2}{\tau^2}\right\}, \frac{1}{4}\right\}.$$

Theorem 4.3.4. *Take any $M \in \mathbb{N}$ and any $\sigma, \tau > 0$. Consider any normal density in $\mathbb{R}^M$ that has covariance matrix $\tau^2 I_M$, and take the restriction of this density to any half-space that contains the center of the normal distribution. Take the Metropolis-Hastings kernel for S' with respect to this restricted normal density. Then the conductance Φ_{MH} of the kernel satisfies*

$$\Phi_{MH} \geq \frac{\Phi_{loc}^2 \sigma}{2^{3/2}\tau\sqrt{M\pi}}$$

for a quantity Φ_{loc} *that satisfies*

$$\Phi_{loc} \geq \min\left\{\frac{1}{2}\exp\left\{-\frac{2\sqrt{M}\sigma^2}{\tau^2}\right\}, \frac{1}{8}\right\}.$$

4.3.2 Metropolis-Hastings is Torpidly Mixing for $\beta = 1$

Recall the weighted normal mixture $\tilde{\pi}$, the corresponding proposal kernel S, and the sets A_1 and A_2. Consider the Metropolis-Hastings kernel for S with respect to $\tilde{\pi}$. Note that the boundary B_{A_2} of A_2 with respect to the Metropolis-Hastings kernel is the set of $z \in A_2$ such that z is within distance M^{-1} of the hyperplane $\sum_i z_i = 0$.

Recall that the probability under any normal distribution with covariance $\sigma^2 \mathrm{I}_M$ for $\sigma > 0$ of any half-space that is distance d from the center of the normal distribution is $\Phi(-d/\sigma)$, where Φ is the cumulative normal distribution function. Therefore

$$\begin{aligned}
\frac{\tilde{\pi}[B_{A_2}]}{\tilde{\pi}[A_2]} &= \frac{\Phi\left(b\sqrt{M}\right) - \Phi\left(b\sqrt{M} - M^{-1}\right)}{\Phi\left(b\sqrt{M}\right)} \\
&\leq 2\left[\Phi\left(b\sqrt{M}\right) - \Phi\left(b\sqrt{M} - M^{-1}\right)\right] \\
&\leq 2\left[1 - \Phi\left(b\sqrt{M} - M^{-1}\right)\right] \\
&= 2\Phi\left(-b\sqrt{M} + M^{-1}\right).
\end{aligned}$$

For M large enough, this is bounded above by $2\Phi\left(-b\sqrt{M}/2\right)$. Analytic integration shows that for any $a > 0$, $\Phi(-a) \leq N_1(a; 0, 1)/a$. Therefore $\tilde{\pi}[B_{A_2}]/\tilde{\pi}[A_2]$ is exponentially decreasing. Similarly, for B_{A_1} equal to the boundary of A_1 with respect to the Metropolis-Hastings kernel, $\tilde{\pi}[B_{A_1}]/\tilde{\pi}[A_1]$ is exponentially decreasing. Using

the form of conductance in (4.6), the conductance of A_2 is exponentially decreasing, which implies that Metropolis-Hastings is torpidly mixing, by Theorem 4.3.1.

4.3.3 Metropolis-Hastings is Rapidly Mixing for $\beta = M^{-1}$

We will show that T_0 is rapidly mixing, where T_0 is the Metropolis-Hastings kernel for S with respect to $\tilde{\pi}_{M^{-1}}$. This will imply that $\bar{T}_0$ is rapidly mixing, by Theorem 3.2.2. Recall that there is some $c \geq 1$ such that $a_M/(1-a_M) \in [1/2, c^M]$ for all M. Therefore

$$\frac{(a_M)^{M^{-1}}}{(a_M)^{M^{-1}} + (1-a_M)^{M^{-1}}} \in \left[\frac{1}{2}, \frac{1}{1+c^{-1}}\right]$$

and

$$\frac{(1-a_M)^{M^{-1}}}{(a_M)^{M^{-1}} + (1-a_M)^{M^{-1}}} \in \left[\frac{1}{1+c}, \frac{1}{2}\right].$$

Using (4.5),

$$\tilde{\pi}_{M^{-1}}(z) \in \left[\frac{2}{1+c}\tilde{\psi}(z), \frac{2}{1+c^{-1}}\tilde{\psi}(z)\right] \qquad z \in \mathcal{X}$$

where

$$\tilde{\psi}(z) \propto \frac{1}{2}N_M(z; -b1_M, M\mathrm{I}_M)\mathbf{1}_{A_1}(z) + \frac{1}{2}N_M(z; b1_M, M\mathrm{I}_M)\mathbf{1}_{A_2}(z) \qquad z \in \mathcal{X}$$

Note that for $z \in A_1$, we have $N_M(z; -b1_M, \mathrm{I}_M) \geq N_M(z; b1_M, \mathrm{I}_M)$. Similarly, for $z \in A_2$, we have $N_M(z; -b1_M, \mathrm{I}_M) \leq N_M(z; b1_M, \mathrm{I}_M)$. Therefore $\tilde{\psi}$ is within a factor of two of the density

$$\psi(z) \stackrel{\text{def}}{=} \frac{1}{2}N_M(z; -b1_M, M\mathrm{I}_M) + \frac{1}{2}N_M(z; b1_M, M\mathrm{I}_M) \qquad z \in \mathcal{X}$$

Therefore $\tilde{\pi}_{M^{-1}}$ and ψ are within a constant factor of one another. Recall that S has proposal radius M^{-1}. We wish to show that Metropolis-Hastings for S with respect to $\tilde{\pi}_{M^{-1}}$ is rapidly mixing. By Lemma 4.3.1, it is equivalent to show that Metropolis-Hastings for S with respect to ψ is rapidly mixing. By Theorem 4.3.2 it is furthermore sufficient to show that:

1. the overlap d between $N_M(z; -b1_M, M\mathrm{I}_M)$ and $N_M(z; b1_M, M\mathrm{I}_M)$ is polynomially decreasing in M,

2. Metropolis-Hastings for S with respect to $N_M(z; -b1_M, M\mathrm{I}_M)$ is rapidly mixing, and

3. Metropolis-Hastings for S with respect to $N_M(z; b1_M, M\mathrm{I}_M)$ is rapidly mixing.

Consider the Metropolis-Hastings chain for S with respect to $N_M(z; -b1_M, M\mathrm{I}_M)$. Applying Theorem 4.3.3 with $\sigma = M^{-1}$ and $\tau = M^{1/2}$ then shows that this Metropolis-Hastings chain is rapidly mixing, so condition 2 above is satisfied. By symmetry, condition 3 is also satisfied.

Now we will show that the overlap between $N_M(z; -b1_M, M\mathrm{I}_M)$ and $N_M(z; b1_M, M\mathrm{I}_M)$ is constant in M. Recall that for $z \in A_1$, $N_M(z; -b1_M, M\mathrm{I}_M) \geq N_M(z; b1_M, M\mathrm{I}_M)$. Therefore

$$\int_{A_1} \min\{N_M(z; -b1_M, M\mathrm{I}_M), N_M(z; b1_M, M\mathrm{I}_M)\}\lambda(dz)$$

$$= \int_{A_1} N_M(z; b1_M, M\mathrm{I}_M)\lambda(dz) = \Phi(-b)$$

By symmetry,

$$\int \min\{N_M(z; -b1_M, MI_M), N_M(z; b1_M, MI_M)\}\lambda(dz) = 2\Phi(-b)$$

Therefore the overlap between $N_M(z; -b1_M, MI_M)$ and $N_M(z; b1_M, MI_M)$ is constant, so T_0 is rapidly mixing.

4.3.4 Metropolis-Hastings is Rapidly Mixing when Restricted to A_1 or to A_2

We will show that the infimum of the spectral gap of Metropolis-Hastings for S with respect to $\tilde{\pi}_\beta|_{A_2}$ over $\beta \geq M^{-1}$ is polynomially decreasing in M. This implies that regardless of the choice of inverse temperatures, since $\beta_0 = M^{-1}$, the minimum of $\mathbf{Gap}(T_k|_{A_2})$ over k is polynomially decreasing in M. The same then holds for $\mathbf{Gap}(T_k|_{A_1})$. Applying Theorem 4.3.4 with $\sigma = M^{-1}$ and $\tau = \beta^{-1/2}$ shows that

$$\begin{aligned}
\inf_{\beta\in[M^{-1},1]} \Phi_{\text{loc}} &\geq \inf_{\beta\in[M^{-1},1]} \min\left\{\frac{1}{2}\exp\left\{-\frac{2\sqrt{M}\beta}{M^2}\right\}, \frac{1}{8}\right\} \\
&= \min\left\{\frac{1}{2}\exp\left\{-2M^{-3/2}\right\}, \frac{1}{8}\right\} \\
&\geq \min\left\{\frac{1}{2}\exp\left\{-2\right\}, \frac{1}{8}\right\}
\end{aligned}$$

so that the local conductance is bounded below by a constant. Note that σ/τ is polynomially decreasing in M. Therefore by Theorem 4.3.4, the infimum over $\beta \geq M^{-1}$ of the conductance of Metropolis-Hastings is polynomially decreasing. By Theorem 4.3.1, the infimum over $\beta \geq M^{-1}$ of the spectral gap is also polynomially decreasing.

4.4 Rapid Mixing on the Symmetric Normal Mixture, with Tails

Consider the symmetric normal mixture π and its truncation approximation $\tilde{\pi}$ as defined in Section 4.2. We will show that the results of that section for $\tilde{\pi}$ also hold for π. We will need a result showing that when a mixture density is tempered, the result is "close" to a mixture of the tempered densities, meaning the following:

Theorem 4.4.1. *Consider any mixture density π', defined on a space $\mathcal{X}$ with measure λ, so that*

$$\pi'(z) = \sum_{i=1}^{I} a_i \psi_i(z) \qquad z \in \mathcal{X}$$

where $\sum_i a_i = 1$ and the ψ_i are densities with respect to λ. Define the following quantity for any β, $j \in \{1, \ldots, I\}$:

$$p_{\beta,j} = \frac{\int a_j^\beta \psi_j(w)^\beta \lambda(dw)}{\sum_i \int a_i^\beta \psi_i(w)^\beta \lambda(dw)}.$$

Define the following density:

$$\psi_\beta(z) = \sum_{j=1}^{I} p_{\beta,j} \psi_{j\beta}(z)$$

where $\psi_{j\beta}$ is ψ_j tempered by β as defined in Section 2.3. Then

$$\pi'_\beta(z) \in [I^{-2}\psi_\beta(z), I^2 \psi_\beta(z)].$$

Proof. Note that $\psi_\beta(z) = 0$ if and only if $\pi'(z) = 0$. Define the following function for

any $j \in \{1, \ldots, I\}$ and $z \in \mathcal{X}$ such that $\pi'(z) > 0$:

$$f_{\beta,j}(z) = \psi_{j\beta}(z) \left[\frac{\left[\sum_i a_i \psi_i(z) \right]^\beta}{\sum_i a_i^\beta \psi_i(z)^\beta} \right] \left[\frac{\sum_i \int a_i^\beta \psi_i(w)^\beta \lambda(dw)}{\int \left[\sum_i a_i \psi_i(w) \right]^\beta \lambda(dw)} \right].$$

Note that $\pi'_\beta(z)$ is equal to $\sum_j p_{\beta,j} f_{\beta,j}(z)$ for all $z \in \mathcal{X}$ such that $\pi'(z) > 0$. Now observe the following bound:

$$\begin{aligned} \left[\sum_i a_i \psi_i(z) \right]^\beta &\geq \left[\max_i \{a_i \psi_i(z)\} \right]^\beta \\ &= \max_i \left\{ a_i^\beta \psi_i(z)^\beta \right\} \geq \frac{1}{I} \sum_i a_i^\beta \psi_i(z)^\beta \end{aligned}$$

which implies that

$$\int \left[\sum_i a_i \psi_i(w) \right]^\beta \lambda(dw) \geq \frac{1}{I} \int \sum_i a_i^\beta \psi_i(w)^\beta \lambda(dw).$$

Also note the following bound in the other direction:

$$\begin{aligned} \left[\sum_i a_i \psi_i(z) \right]^\beta &\leq \left[I \max_i \{a_i \psi_i(z)\} \right]^\beta = I^\beta \max_i \left\{ a_i^\beta \psi_i(z)^\beta \right\} \\ &\leq I^\beta \sum_i a_i^\beta \psi_i(z)^\beta \leq I \sum_i a_i^\beta \psi_i(z)^\beta \end{aligned}$$

which implies that

$$\int \left[\sum_i a_i \psi_i(w) \right]^\beta \lambda(dw) \leq I \int \sum_i a_i^\beta \psi_i(w)^\beta \lambda(dw)$$

Therefore for any β and any z such that $\pi'(z) > 0$,

$$\frac{\left[\sum_i a_i \psi_i(z) \right]^\beta}{\sum_i a_i^\beta \psi_i(z)^\beta} \in [I^{-1}, I]$$

and

$$\frac{\int \sum_i a_i^\beta \psi_i(w)^\beta \lambda(dw)}{\int \left[\sum_i a_i \psi_i(w)\right]^\beta \lambda(dw)} \in [I^{-1}, I]$$

Therefore

$$\left[\frac{\left[\sum_i a_i \psi_i(z)\right]^\beta}{\sum_i a_i^\beta \psi_i(z)^\beta}\right] \left[\frac{\sum_i \int a_i^\beta \psi_i(w)^\beta \lambda(dw)}{\int \left[\sum_i a_i \psi_i(w)\right]^\beta \lambda(dw)}\right] \in \left[I^{-2}, I^2\right]$$

This proves that for any β and j, $f_{\beta,j}(z) \in [I^{-2}\psi_{j\beta}(z), I^2\psi_{j\beta}(z)]$. Therefore Theorem 4.4.1 holds. □

Let us apply Theorem 4.4.1 to the symmetric mixture of normals given in (4.1). Note that $p_{\beta,1} = p_{\beta,2} = 1/2$ for any β. Therefore for any β,

$$\psi_\beta(z) = \frac{1}{2} N_M(z; -b1_M, \beta^{-1}\mathrm{I}_M) + \frac{1}{2} N_M(z; b1_M, \beta^{-1}\mathrm{I}_M)$$

and

$$\pi_\beta(z) \in \left[2^{-2}\psi_\beta(z), 2^2\psi_\beta(z)\right] \tag{4.7}$$

It is straightforward to show that $\psi_\beta(z)$ is within a factor of two of $\tilde{\pi}_\beta(z)$ for every z and β. Therefore $\pi_\beta(z)$ is within a factor of 2^3 of $\tilde{\pi}_\beta(z)$. In Section 4.2 we showed that for $\tilde{\pi}$, $\min_{k,j} \mathbf{Gap}(T_k|_{A_j})$ is polynomially decreasing in M, T_0 is rapidly mixing, and the overlap $\delta(\mathcal{A})$ is polynomially decreasing. Using Lemma 4.3.1, these conditions also hold for π. Therefore parallel and simulated tempering are rapidly mixing for the symmetric normal mixture π.

4.5 Proof of Metropolis-Hastings Mixing on Normal Densities

We will prove Theorems 4.3.3 and 4.3.4 using the following results from Kannan and Li (1996a) and Kannan and Li (1996b). First, define a real-valued function g on $\mathbb{R}^M$ to be log-concave if it is nonnegative and if, for any $z, w \in \mathbb{R}^M$ and any $\rho \in [0,1]$,

$$g(z)^\rho g(w)^{1-\rho} \leq g(\rho z + (1-\rho)w).$$

Then we have the following result.

Theorem 4.5.1. *(Kannan and Li (1996a), Theorem 3.1) Take any $M \in \mathbb{N}$ and any log-concave function g on $\mathbb{R}^M$. Let f be the density defined as follows:*

$$f(z) \propto g(z) N_M(z; 0, I_M) \qquad z \in \mathbb{R}^M$$

Take any $\sigma > 0$, and define a proposal kernel S' in $\mathbb{R}^M$ that proposes uniformly on the ball of radius σ centered at the current state. Let $B(\sigma, z)$ denote this ball, where z is the current state, and let $vol(B(\sigma, z))$ denote the volume of $B(\sigma, z)$. Consider the Metropolis-Hastings kernel for S' with respect to f. For any $z \in \mathbb{R}^M$ such that $f(z) > 0$, define the following quantity, where λ is Lebesgue measure in $\mathbb{R}^M$:

$$\Phi_{loc}(z) = \frac{\int\limits_{w \in B(\sigma,z)} \min\{f(z), f(w)\}\, \lambda(dw)}{f(z)\ vol(B(\sigma, z))}$$

Define the "local conductance" of the Metropolis-Hastings kernel as

$$\Phi_{loc} = \inf_{z \in \mathbb{R}^M : f(z) > 0} \Phi_{loc}(z).$$

Then the conductance Φ_{MH} of the Metropolis-Hastings kernel satisfies

$$\Phi_{MH} \geq \frac{\Phi_{loc}^2 \sigma}{2^{3/2}\sqrt{M\pi}}.$$

The proof of Theorem 4.5.1 is in Kannan and Li (1996b), as is the following lemma:

Lemma 4.5.1. *(Kannan and Li 1996b) Take any ball B_r of radius r in $\mathbb{R}^M$ that is centered on the surface of another ball B_R of radius R, where $R \geq \frac{\sqrt{M}r}{2}$. Then the volume of their intersection satisfies*

$$vol(B_r \cap B_R) \geq \left(\frac{1}{2} - \frac{r\sqrt{M}}{4R}\right) vol(B_r).$$

The proof of Theorem 4.3.3 uses Theorem 4.5.1 and Lemma 4.5.1 and is closely related to the proof of a similar bound in Kannan and Li (1996b). Consider the context of Theorem 4.3.3. First scale by a factor of τ^{-1} and translate to center the normal density at zero, yielding a standard normal target density. The proposal radius is $\sigma' = \sigma/\tau$ after scaling and translation. Since the acceptance probability of Metropolis-Hastings is invariant to such linear transformations of the space, the conductance Φ_{MH} is unchanged. We will therefore show that Metropolis-Hastings for proposal radius σ', with respect to the standard normal density, satisfies

$$\Phi_{\text{loc}} \geq \min\left\{\exp\left\{-2\sqrt{M}\sigma'^2\right\}, \frac{1}{4}\right\}$$

and

$$\Phi_{\text{MH}} \geq \frac{\Phi_{\text{loc}}^2 \sigma'}{2^{3/2}\sqrt{M\pi}}.$$

We will show this result by applying Theorem 4.5.1 with $g(z) = 1$ for all $z \in \mathcal{X}$, so that $f(z) = N_M(z; 0, \mathrm{I}_M)$. First we bound the local conductance of the chain. Let the current state be denoted z. If $\|z\|_2 \leq \sqrt{M}\sigma'$,

$$\begin{aligned}
\Phi_{\mathrm{loc}}(z) &\geq \frac{\inf_{w \in B(\sigma', z)} f(w)}{f(z)} = \frac{\inf_{w \in B(\sigma', z)} \exp\{-\|w\|_2^2/2\}}{\exp\{-\|z\|_2^2/2\}} \\
&= \frac{\exp\{-(\|z\|_2 + \sigma')^2/2\}}{\exp\{-\|z\|_2^2/2\}} = \exp\{-(\|z\|_2\sigma' + \frac{\sigma'^2}{2})\} \\
&\geq \exp\{-(\sqrt{M}\sigma'^2 + \frac{\sigma'^2}{2})\} \geq \exp\{-2\sqrt{M}\sigma'^2\}.
\end{aligned}$$

If $\|z\|_2 > \sqrt{M}\sigma'$,

$$\begin{aligned}
\Phi_{\mathrm{loc}}(z) &\geq \frac{\int_{w \in B(\sigma', z)} 1(f(w) \geq f(z))\lambda(dw)}{\mathrm{vol}(B(\sigma', z))} \\
&= \frac{\int_{w \in B(\sigma', z)} 1(\|w\|_2 \leq \|z\|_2)\lambda(dw)}{\mathrm{vol}(B(\sigma', z))} \\
&= \frac{\mathrm{vol}(B(\sigma', z) \cap B(\|z\|_2, 0))}{\mathrm{vol}(B(\sigma', z))}. \qquad (4.8)
\end{aligned}$$

Then we can apply Lemma 4.5.1, obtaining

$$\frac{\mathrm{vol}(B(\sigma', z) \cap B(\|z\|_2, 0))}{\mathrm{vol}(B(\sigma', z))} \geq \frac{1}{2} - \frac{\sigma'\sqrt{M}}{4\|z\|_2} \geq \frac{1}{4}$$

Therefore we have proven the bound on Φ_{loc}. Since g is log-concave, we can apply Theorem 4.5.1, giving the desired bound on the conductance of the Metropolis-Hastings chain.

Next we prove Theorem 4.3.4. Consider the restriction of the normal distribution to the half-space that contains the center of the normal distribution. It is straightforward to show that such a restriction reduces the local conductance of the

Metropolis-Hastings kernel by at most a factor of two, relative to the unrestricted normal in Theorem 4.3.3. Application of Theorem 4.5.1 then implies the desired result.

. □

Chapter 5

Upper Bounds on the Convergence Rates of Parallel and Simulated Tempering and Conditions for Torpid Mixing

As we showed in Section 2.1, an upper bound on the spectral gap of a Markov chain implies an upper bound on the rate of convergence to stationarity. In this chapter we will give upper bounds on the spectral gaps of parallel and simulated tempering chains, as well as conditions for torpid mixing.

5.1 Upper Bounds on the Spectral Gaps of Swapping and Simulated Tempering Chains and Conditions for Torpid Mixing

Consider a parallel or simulated tempering chain as defined in Section 2.3. It is typically assumed that if such a chain has high acceptance rates for swap or temperature-

changing moves between all adjacent temperature levels, then it is mixing quickly. We show that this is not necessarily the case: if the target distribution has a subset with low *conductance* for β close to 1 and low *persistence* (defined below), then the tempering chain mixes slowly. In addition, we show that if the inverse temperatures of two adjacent levels are far apart so that the *overlap* (also defined below) of the levels is small, the tempering chain mixes slowly.

For purposes of sampling from continuous π, consider sets $A \subset \mathcal{X}$ that contain a single local mode of π along with the surrounding area of high density. If π has multiple modes, separated by areas of low density, and if the proposal kernel makes only local moves, then the *conductance* of A with respect to Metropolis-Hastings is typically small for β close to 1. The conductance of a set $A \subset \mathcal{X}$ is defined as follows for any transition kernel P that is reversible with respect to a distribution μ on $\mathcal{X}$, where we require that $0 < \mu(A) < 1$:

$$\Phi_P(A) = \frac{(\mathbf{1}_A, P\mathbf{1}_{A^c})_\mu}{\mu(A)\mu(A^c)}$$

and where $\mathbf{1}_A$ is the indicator function of the set A. The conductance of A is an upper bound on the spectral gap of P (Lawler and Sokal 1988). Note that since P is reversible with respect to μ,

$$(\mathbf{1}_A, P\mathbf{1}_{A^c})_\mu = \int_{x \in A} \int_{y \in A^c} \mu(dx) P(x, dy) = (\mathbf{1}_{A^c}, P\mathbf{1}_A)_\mu$$

so we can rewrite $\Phi_P(A)$ as

$$\frac{(\mathbf{1}_A, P\mathbf{1}_{A^c})_\mu}{\mu(A)} + \frac{(\mathbf{1}_{A^c}, P\mathbf{1}_A)_\mu}{\mu(A^c)}. \tag{5.1}$$

In particular, $\Phi_P(A) \leq 2$.

We will first give upper bounds on the spectral gap of any parallel or simulated tempering chain with $\phi_k = \pi_{\beta_k}$ for some N and $\{\beta_k : k = 0, \ldots, N\}$, in terms of an arbitrary subset A of $\mathcal{X}$. The set A can be taken so that $\pi|_A$ is unimodal as described above, but we only require that $0 < \pi[A] < 1$.

Recall the definition of H_β in Section 2.3. The bounds will be in terms of the conductance of A under H_β and the persistence of A under tempering by β, defined as

$$\gamma(A, \beta) = \min\left\{1, \frac{\pi_\beta[A]}{\pi[A]}\right\} \tag{5.2}$$

The persistence measures how much smaller the probability of A is under π_β than under π, if it is smaller. If A has low persistence for small values of β, then a parallel or simulated tempering chain starting in A^c may take a long time to discover A at small β. If A is a unimodal subset of a multimodal distribution then it may have low conductance for β close to 1, so the tempering chain may take a long time to discover A at every inverse temperature, even if the probability of A is high under π. This leads to slow mixing of the tempering chain, as we will see. It also contradicts the general presumption that if the simulated or parallel tempering chain is monitored to ensure that the swapping or level-changing acceptance rates are high, then the chain is mixing quickly.

Recall $\mathcal{B}$ from Section 2.3. Using the conductance $\Phi_{H_\beta}(A)$ and the persistence $\gamma(A, \beta)$, we have the following result.

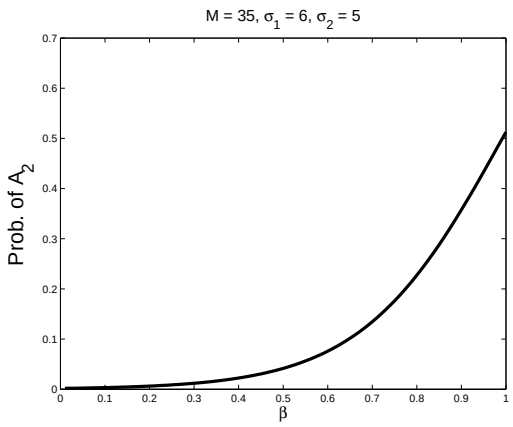

Figure 5.1: The probability of $A = A_2$ under $\tilde{\pi}_\beta$ as a function of β, for the approximated mixture of normals in Section 6.1 with $M = 35$, $\sigma_1 = 6$, and $\sigma_2 = 5$.

Theorem 5.1.1. *Consider a parallel tempering chain P_{pt} that uses either of the swapping schemes* **SC1** *or* **SC2**, *or a simulated tempering chain P_{st} that changes levels using scheme* **ST1** *(Section 2.3), such that $\phi_k = \pi_{\beta_k}$ and $T_k = H_{\beta_k}$ for some N and some $\{\beta_k : k = 0, \ldots, N\}$. For any $A \subset \mathcal{X}$ such that $0 < \pi[A] < 1$,* $\mathbf{Gap}(P_{pt})$ *and* $\mathbf{Gap}(P_{st})$ *are bounded above by*

$$6 \sup_{\beta \in \mathcal{B}} \left\{ \gamma(A, \beta) \Phi_{H_\beta}(A) \right\} \qquad \textit{and} \qquad 192 \left[\sup_{\beta \in \mathcal{B}} \left\{ \gamma(A, \beta) \Phi_{H_\beta}(A) \right\} \right]^{1/4}$$

respectively.

This is a consequence of a more general result for any swapping or simulated tempering chain (ϕ_k not necessarily tempered versions of π). In this case, the persistence of

A is defined for each level k as

$$\gamma(A,k) = \min\left\{1, \frac{\phi_k[A]}{\phi_N[A]}\right\}. \tag{5.3}$$

This quantity is related in the following way to $\gamma(\{A_j\})$ for a partition $\{A_j : j = 1,\ldots,J\}$ of $\mathcal{X}$, as defined in Section 3.1. If $\phi_k[A_j]$ is a monotonic function of k for each j, then

$$\gamma(\{A_j\}) = \min_{j,k} \gamma(A_j,k).$$

Next consider the conductance $\Phi_{T_k}(A)$. We will show that it is related to the *projection matrix* $\bar{T}_k$ for T_k with respect to the partition $\{A, A^c\}$, as defined in Section 3.1 and in Madras and Randall (2002). The transition $\bar{T}_k$ is a 2×2 matrix; the spectral gap of such a matrix is given by the sum of the off-diagonal elements. This sum is precisely $\Phi_{T_k}(A)$, written in the form (5.1), so that $\mathbf{Gap}(\bar{T}_k) = \Phi_{T_k}(A)$.

Using $\Phi_{T_k}(A)$ and $\gamma(A,k)$, for any $A \subset \mathcal{X}$ such that $0 < \phi_k[A] < 1$ for all $k \in \{0,\ldots,N\}$, $\mathbf{Gap}(P_{sc})$ and $\mathbf{Gap}(P_{st})$ are bounded above by

$$6\max_k\{\gamma(A,k)\Phi_{T_k}(A)\} \qquad \text{and} \qquad 192\left[\max_k\{\gamma(A,k)\Phi_{T_k}(A)\}\right]^{1/4} \tag{5.4}$$

respectively. Note that these bounds directly imply Theorem 5.1.1. The bound for P_{sc} is proven in Section 5.2, and that for P_{st} is proven in Section 5.3.

Recall from Section 2.1 that torpid mixing of a chain means that the spectral gap of the transition kernel is exponentially decreasing in the problem size. Theorem 5.1.1 implies that if there is some inverse temperature $\beta^a \in \mathcal{B}$ such that the conductance of A is exponentially decreasing in the problem size for $\beta \in [\beta^a, 1]$, and such that

the persistence of A is exponentially decreasing for $\beta \in [0, \beta^a) \cap \mathcal{B}$, then parallel and simulated tempering are torpidly mixing. In Section 6.1 we will specify a mixture of normals with unequal covariances in $\mathbb{R}^M$. We will that show for an approximation of the normal mixture ($\tilde{\pi}$) formed by truncation of the overlapping parts of the tails, the persistence of a set A containing the narrower mode is exponentially decreasing for $\beta < \frac{1}{2}$. Figure 5.1 shows $\tilde{\pi}_\beta[A]$ as a function of β for $M = 35$. It is clear from the figure that for $\beta < \frac{1}{2}$, $\tilde{\pi}_\beta[A]$ is much smaller than $\pi[A]$. This effect becomes more extreme as M increases, leading to exponentially small persistence for $\beta < \frac{1}{2}$. We will also show that for the mixture $\tilde{\pi}$ the conductance of the same set A is exponentially decreasing for $\beta \geq \frac{1}{2}$. Therefore Theorem 5.1.1 will imply the torpid mixing of parallel and simulated tempering for this mixture. The untruncated version of the normal mixture is addressed at the end of this section.

For a general target distribution, even if every subset has high persistence for small values of β, having a subset with low persistence within an intermediate β-interval is enough to cause slow mixing by creating a bottleneck in the parallel or simulated tempering chain. This is because a proposed move between a small β and a large β' typically has a very low probability of being accepted. The probability of acceptance of a proposed move in the simulated tempering chain from inverse temperature β to inverse temperature β', conditional on $z \in A$, will be called the overlap of π_β and $\pi_{\beta'}$ with respect to A, and is given by

$$\delta(A, \beta, \beta') = \frac{\int_A \min\{\pi_\beta(z), \pi_{\beta'}(z)\}\, \lambda(dz)}{\pi_\beta[A]}.$$

More generally, for any swapping chain or simulated tempering chain define the overlap of levels k and l with respect to A to be:

$$\delta(A,k,l) = \frac{\int_A \min\{\phi_k(z), \phi_l(z)\}\,\lambda(dz)}{\phi_k[A]}. \tag{5.5}$$

The quantity $\delta(A,k,l)$ is related as follows to the overlap of $\{\phi_k : k = 0,\ldots,N\}$ with respect to a partition $\{A_j, j = 1,\ldots,J\}$ of $\mathcal{X}$, as defined in Section 3.1:

$$\delta(\{A_j\}) = \min_{|k-l|=1,j} \delta(A_j,k,l).$$

The quantity $\delta(\{A_j\})$ is in turn a lower bound on the overlap quantity defined in Zheng (2003). Using $\delta(A,k,l)$, $\gamma(A,k)$, and the conductance $\Phi_{T_k}(A)$, we have the following result, which is proven for P_{sc} in Section 5.2 and for P_{st} in Section 5.3.

Theorem 5.1.2. *Consider a swapping chain P_{sc} that uses either of the swapping schemes* **SC1** *or* **SC2**, *or a simulated tempering chain P_{st} that changes levels using scheme* **ST1** *(Section 2.3). For any $A \subset \mathcal{X}$ such that $0 < \phi_k[A] < 1$ for all k, and for any $k^* \in \{1,\ldots,N\}$,*

$$\mathbf{Gap}(P_{sc}) \le 12 \max_{k\ge k^*, l<k^*} \{\gamma(A,k)\max\{\Phi_{T_k}(A), \delta(A,k,l), \delta(A^c,k,l)\}\}$$

and

$$\mathbf{Gap}(P_{st}) \le 192 \left[\max_{k\ge k^*, l<k^*} \{\gamma(A,k)\max\{\Phi_{T_k}(A), \delta(A,k,l)\}\}\right]^{1/4}.$$

For the case where the ϕ_k are tempered versions of π, the bounds in Theorem 5.1.2 show that adjacent inverse temperatures must be chosen close enough for the overlaps

to be large, since it is clear that if there is some level k^* such that the overlap of any pair of levels $k \geq k^*$ and $l < k^*$ with respect to A and A^c is exponentially decreasing, and such that the conductance of A is exponentially decreasing for $k \geq k^*$, then the parallel or simulated tempering chain is torpidly mixing. This is the case for the mean-field Ising model with fixed inverse temperatures, as we will show in Section 6.3.

Although the bounds in Theorem 5.1.2 are for a specific N and set of densities ϕ_k, if the ϕ_k are tempered versions of π then the bounds can be generalized to not depend on the number and choice of inverse temperatures:

Corollary 5.1.1. *Consider a parallel tempering chain P_{pt} that uses either of the swapping schemes* **SC1** *or* **SC2**, *or a simulated tempering chain P_{st} that changes levels using scheme* **ST1** *(Section 2.3), such that $\phi_k = \pi_{\beta_k}$ and $T_k = H_{\beta_k}$ for some N and $\{\beta_k : k = 0, \ldots, N\}$. Take any inverse temperature $\beta^b \in \mathcal{B}$ such that $\beta^b > \inf\{\beta \in \mathcal{B}\}$ and any $A \subset \mathcal{X}$ such that $0 < \pi[A] < 1$. Regardless of the choice of N and $\{\beta_k\}$,*

$$\mathbf{Gap}(P_{pt}) \leq 12 \sup_{\substack{\beta \in [\beta^b, 1] \\ \beta' \in [0, \beta^b) \cap \mathcal{B}}} \left\{ \gamma(A, \beta) \max\left\{ \Phi_{H_\beta}(A), \delta(A, \beta, \beta'), \delta(A^c, \beta, \beta') \right\} \right\}$$

and

$$\mathbf{Gap}(P_{st}) \leq 192 \left[\sup_{\substack{\beta \in [\beta^b, 1] \\ \beta' \in [0, \beta^b) \cap \mathcal{B}}} \left\{ \gamma(A, \beta) \max\left\{ \Phi_{H_\beta}(A), \delta(A, \beta, \beta') \right\} \right\} \right]^{1/4}.$$

This is a corollary of Theorems 5.1.1 and 5.1.2, verified by setting $k^* = \min\{k :$

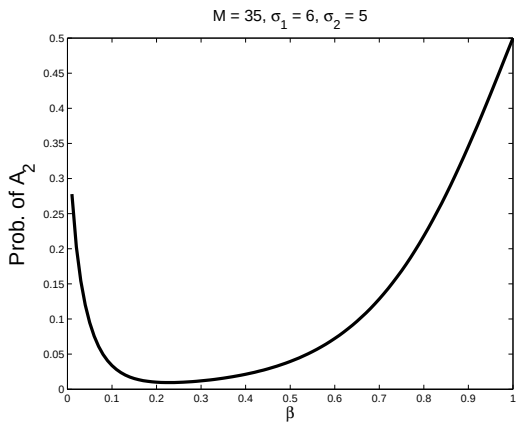

Figure 5.2: The probability of $A = A_2$ under π_β as a function of β, for the mixture of normals in Section 6.1 with $M = 35$, $\sigma_1 = 6$, and $\sigma_2 = 5$.

$\beta_k \geq \beta^b\}$. If $k^* = 0$ then use Theorem 5.1.1; otherwise, use Theorem 5.1.2. Note that if there is some inverse temperature $\beta^a > \beta^b$ such that the conductance of A is exponentially decreasing in the problem size for $\beta \in [\beta^a, 1]$, such that the persistence of A is exponentially decreasing for $\beta \in [\beta^b, \beta^a)$, and such that the overlap of β and β' with respect to A and A^c is exponentially decreasing for $\beta \in [\beta^a, 1]$ and $\beta' \in [0, \beta^b) \cap \mathcal{B}$, then parallel and simulated tempering are torpidly mixing. This latter condition on the overlap is presumably the case in most problems of interest, for which a range of intermediate β values are necessary to interpolate between small and large values of β. Therefore having a set A with low conductance for β close to 1 and low persistence for β in some intermediate β-interval leads to slow mixing of parallel and simulated tempering. Note that this is possible since $\pi_\beta[A]$ is not necessarily a monotonic function of β; if we include the tails of the normal mixture

from Figure 5.1, we obtain Figure 5.2. In order to show torpid mixing with the tails included, we would need to use Corollary 5.1.1 rather than Theorem 5.1.1, since the set A has low persistence within an intermediate β-interval and higher persistence for small β, as shown in Figure 5.2.

5.2 Proof of the Upper Bounds on the Spectral Gap of a Swapping Chain

We will prove the bounds in (5.4) and Theorem 5.1.2 for the swapping chain. Recall the context of those results. Also recall from Section 2.3 that $P_{sc} = QTQ$. Note that by the definition of the spectral gap given in (2.1), $\mathbf{Gap}(QTQ) = 8\mathbf{Gap}(\frac{1}{8}QTQ + \frac{7}{8}I)$. By Lemma 3.2.1, $\mathbf{Gap}(\frac{1}{8}QTQ + \frac{7}{8}I) \leq \mathbf{Gap}((\frac{1}{2}T + \frac{1}{2}Q)^3)$. By Lemma 3.2.2, $\mathbf{Gap}((\frac{1}{2}T + \frac{1}{2}Q)^3) \leq 3\mathbf{Gap}(\frac{1}{2}T + \frac{1}{2}Q)$. Therefore

$$\mathbf{Gap}(P_{sc}) \leq 24\mathbf{Gap}(\frac{1}{2}T + \frac{1}{2}Q). \tag{5.6}$$

Take any $A \subset \mathcal{X}$ such that $0 < \phi_k[A] < 1$ for all k, and any $k^* \in \{0, \ldots, N\}$. We will set $k^* = 0$ to prove the bound in (5.4) and $k^* > 0$ to prove Theorem 5.1.2. Define $B = \{x \in \mathcal{X}_{sc} : \forall k \geq k^*, x_{[k]} \in A^c\}$. The spectral gap of $(\frac{1}{2}T + \frac{1}{2}Q)$ is bounded above by the conductance of B for $(\frac{1}{2}T + \frac{1}{2}Q)$:

$$\begin{aligned}\mathbf{Gap}(\frac{1}{2}T + \frac{1}{2}Q) &\leq \frac{(\mathbf{1}_B, (\frac{1}{2}T + \frac{1}{2}Q)\mathbf{1}_{B^c})_{\pi_{sc}}}{\pi_{sc}[B]\pi_{sc}[B^c]} \\ &= \frac{(\mathbf{1}_B, T\mathbf{1}_{B^c})_{\pi_{sc}}}{2\pi_{sc}[B]\pi_{sc}[B^c]} + \frac{(\mathbf{1}_B, Q\mathbf{1}_{B^c})_{\pi_{sc}}}{2\pi_{sc}[B]\pi_{sc}[B^c]}.\end{aligned} \tag{5.7}$$

Observe that $\pi_{sc}[B^c] = 1 - \prod_{k \geq k^*} \phi_k[A^c]$, so that for any $k \geq k^*$,

$$\pi_{sc}[B^c] \geq \max\{\phi_k[A], \phi_N[A]\}.$$

Therefore

$$\begin{aligned}
\frac{(\mathbf{1}_B, T\mathbf{1}_{B^c})_{\pi_{sc}}}{\pi_{sc}[B]\pi_{sc}[B^c]} &= \frac{1}{\pi_{sc}[B^c]} \frac{1}{2(N+1)} \sum_{k \geq k^*} \frac{(\mathbf{1}_{A^c}, T_k \mathbf{1}_A)_{\phi_k}}{\phi_k[A^c]} \\
&\leq \frac{1}{2\pi_{sc}[B^c]} \max_{k \geq k^*} \left\{ \frac{(\mathbf{1}_{A^c}, T_k \mathbf{1}_A)_{\phi_k}}{\phi_k[A^c]} \right\} \\
&\leq \frac{1}{2} \max_{k \geq k^*} \left\{ \frac{1}{\max\{\phi_k[A], \phi_N[A]\}} \frac{(\mathbf{1}_{A^c}, T_k \mathbf{1}_A)_{\phi_k}}{\phi_k[A^c]} \right\} \\
&= \frac{1}{2} \max_{k \geq k^*} \{\gamma(A,k) \Phi_{T_k}(A)\}. \qquad (5.8)
\end{aligned}$$

To show the bounds in (5.4), take $k^* = 0$. In this case $(\mathbf{1}_B, Q\mathbf{1}_{B^c})_{\pi_{sc}} = 0$, so combining (5.6), (5.7) and (5.8) proves the bound in (5.4) for the swapping chain. To show Theorem 5.1.2, take $k^* > 0$. First consider the swapping scheme **SC1**. Note that

$$\begin{aligned}
&\frac{(\mathbf{1}_B, Q\mathbf{1}_{B^c})_{\pi_{sc}}}{\pi_{sc}[B]\pi_{sc}[B^c]} \\
&= \frac{1}{\pi_{sc}[B^c]} \sum_{k \geq k^*, l < k^*} \frac{1}{(N+1)^2} \frac{\int_{z \in A^c} \int_{w \in A} \min\{\phi_k(z)\phi_l(w), \phi_k(w)\phi_l(z)\} \lambda(dw)\lambda(dz)}{\phi_k[A^c]} \\
&\leq \frac{1}{4\pi_{sc}[B^c]} \max_{k \geq k^*, l < k^*} \frac{\int_{z \in A^c} \int_{w \in A} \min\{\phi_k(z)\phi_l(w), \phi_k(w)\phi_l(z)\} \lambda(dw)\lambda(dz)}{\phi_k[A^c]} \\
&\leq \frac{1}{4} \max_{k \geq k^*, l < k^*} \frac{\phi_k[A]}{\max\{\phi_k[A], \phi_N[A]\}} \frac{\int_{z \in A^c} \int_{w \in A} \min\{\phi_k(z)\phi_l(w), \phi_k(w)\phi_l(z)\} \lambda(dw)\lambda(dz)}{\phi_k[A]\phi_k[A^c]} \\
&= \frac{1}{4} \max_{k \geq k^*, l < k^*} \gamma(A,k) \frac{\int_{z \in A^c} \int_{w \in A} \min\{\phi_k(z)\phi_l(w), \phi_k(w)\phi_l(z)\} \lambda(dw)\lambda(dz)}{\phi_k[A]\phi_k[A^c]}.
\end{aligned}$$

Consider the case where $\phi_l[A^c] < \phi_k[A^c]$. Then for any k, l,

$$\begin{aligned}
&\frac{\int\limits_{z\in A^c}\int\limits_{w\in A}\min\{\phi_k(z)\phi_l(w),\phi_k(w)\phi_l(z)\}\lambda(dw)\lambda(dz)}{\phi_k[A]\phi_k[A^c]} \\
&\leq \frac{\int\limits_{z\in A^c}\int\limits_{w\in A}\min\{\phi_l(w),\phi_k(w)\}\left[\phi_k(z)+\phi_l(z)\right]\lambda(dw)\lambda(dz)}{\phi_k[A]\phi_k[A^c]} \\
&= \frac{(\phi_k[A^c]+\phi_l[A^c])\int\limits_{w\in A}\min\{\phi_l(w),\phi_k(w)\}\lambda(dw)}{\phi_k[A]\phi_k[A^c]} \\
&\leq 2\frac{\int\limits^{w\in A}\min\{\phi_l(w),\phi_k(w)\}\lambda(dw)}{\phi_k[A]} = 2\delta(A,k,l).
\end{aligned}$$

If $\phi_l[A^c] \geq \phi_k[A^c]$, exchanging the roles of A and A^c yields an upper bound of $2\delta(A^c, k, l)$. Therefore

$$\frac{(\mathbf{1}_B, Q\mathbf{1}_{B^c})_{\pi_{sc}}}{\pi_{sc}[B]\pi_{sc}[B^c]} \leq \frac{1}{2}\max_{k\geq k^*, l<k^*}\left[\gamma(A,k)\max\left\{\delta(A,k,l),\delta(A^c,k,l)\right\}\right]. \tag{5.9}$$

Combining (5.6), (5.7), (5.8), and (5.9), we have that for $k^* > 0$, $\mathbf{Gap}(P_{sc})$ is bounded above by

$$12\max\left\{\max_{k\geq k^*}\gamma(A,k)\Phi_{T_k}(A), \max_{k\geq k^*, l<k^*}\gamma(A,k)\max\{\delta(A,k,l),\delta(A^c,k,l)\}\right\}$$

which implies Theorem 5.1.2 for the swapping chain. Note that the above proof holds with only slight modification for the swapping scheme where swaps between only adjacent levels are proposed, rather than swaps between arbitrary levels (scheme **SC2**).

5.3 Proof of the Upper Bounds on the Spectral Gap of a Simulated Tempering Chain

We will prove the bounds in (5.4) and Theorem 5.1.2 for the simulated tempering chain. Recall the context of those results. Take $k^* \in \{0, \ldots, N\}$; we will set $k^* = 0$ to prove (5.4) and $k^* > 0$ to prove Theorem 5.1.2. Define $D = \{k \in \{0, \ldots, N\} : k \geq k^*, \gamma(A, k) \geq \frac{1}{2}\Phi_{T_k}(A)\}$ and $B = \{(z, k) \in \mathcal{X}_{st} : k \in D, z \in A\}$. Note that $N \in D$ so $D \neq \emptyset$. Just as in the proof for the swapping chain, we have

$$\mathbf{Gap}(P_{st}) \leq 12 \left[\frac{(\mathbf{1}_B, T'\mathbf{1}_{B^c})_{\pi_{st}}}{\pi_{st}[B]\pi_{st}[B^c]} + \frac{(\mathbf{1}_B, Q'\mathbf{1}_{B^c})_{\pi_{st}}}{\pi_{st}[B]\pi_{st}[B^c]} \right]. \tag{5.10}$$

Note that for any I and set of constants $\{a_i : i = 1, \ldots, I\}$ and $\{b_i : i = 1, \ldots, I\}$, we have that $\sum_i a_i / \sum_i b_i \leq \max_i\{a_i/b_i\}$. Therefore

$$\begin{aligned}
\frac{(\mathbf{1}_B, T'\mathbf{1}_{B^c})_{\pi_{st}}}{\pi_{st}[B]\pi_{st}[B^c]} &= \frac{(\mathbf{1}_B, T'\mathbf{1}_{B^c})_{\pi_{st}}}{\pi_{st}[B]} + \frac{(\mathbf{1}_B, T'\mathbf{1}_{B^c})_{\pi_{st}}}{\pi_{st}[B^c]} \\
&= \frac{\frac{1}{2(N+1)} \sum_{k \in D} (\mathbf{1}_A, T_k \mathbf{1}_{A^c})_{\phi_k}}{\frac{1}{N+1} \sum_{k \in D} \phi_k[A]} + \frac{\frac{1}{2(N+1)} \sum_{k \in D} (\mathbf{1}_A, T_k \mathbf{1}_{A^c})_{\phi_k}}{\frac{1}{N+1} \sum_{k \in D} \phi_k[A^c] + \frac{|D^c|}{N+1}} \\
&\leq \frac{\sum_{k \in D} (\mathbf{1}_A, T_k \mathbf{1}_{A^c})_{\phi_k}}{2 \sum_{k \in D} \phi_k[A]} + \frac{\sum_{k \in D} (\mathbf{1}_A, T_k \mathbf{1}_{A^c})_{\phi_k}}{2 \sum_{k \in D} \phi_k[A^c]} \\
&\leq \max_{k \in D} \frac{(\mathbf{1}_A, T_k \mathbf{1}_{A^c})_{\phi_k}}{2\phi_k[A]} + \max_{k \in D} \frac{(\mathbf{1}_A, T_k \mathbf{1}_{A^c})_{\phi_k}}{2\phi_k[A^c]} \\
&\leq \max_{k \in D} \frac{(\mathbf{1}_A, T_k \mathbf{1}_{A^c})_{\phi_k}}{2\phi_k[A]\phi_k[A^c]} + \max_{k \in D} \frac{(\mathbf{1}_A, T_k \mathbf{1}_{A^c})_{\phi_k}}{2\phi_k[A]\phi_k[A^c]} \\
&= \max_{k \in D} \Phi_{T_k}(A).
\end{aligned} \tag{5.11}$$

To show Theorem 5.1.2 we will take $k^* > 0$. In this case $|D| < N + 1$. To show the bounds in (5.4) we will take $k^* = 0$. Then if $|D| = N + 1$, we have

$(\mathbf{1}_B, Q'\mathbf{1}_{B^c}) = 0$. In addition if $|D| = N+1$ then for every k we have $\Phi_{T_k}(A)/2 \leq \gamma(A,k)$, so $\Phi_{T_k}(A)^2/4 \leq \gamma(A,k)\Phi_{T_k}(A)/2$. Also note that $\gamma(A,k)\Phi_{T_k}(A) \leq 2$ so $[\gamma(A,k)\Phi_{T_k}(A)/2]^{1/2} \leq [\gamma(A,k)\Phi_{T_k}(A)/2]^{1/4}$. Therefore if $|D| = N+1$ then (5.10) and (5.11) together imply that

$$\begin{aligned}\mathbf{Gap}(P_{st}) &\leq 12 \max_k \Phi_{T_k}(A) \leq 24 \left[\frac{\max_k \gamma(A,k)\Phi_{T_k}(A)}{2}\right]^{1/2} \\ &\leq 24 \left[\frac{\max_k \gamma(A,k)\Phi_{T_k}(A)}{2}\right]^{1/4}. \end{aligned} \tag{5.12}$$

This result implies the bound in (5.4) for the case where $|D| = N+1$. From now on we consider the other case, where $|D| < N+1$.

Define

$$c = \frac{\sum_{k\in D} \phi_k[A]}{|D| \max_{k\in D} \phi_k[A]}.$$

Note that $\pi_{st}[B^c] \geq \frac{|D^c|}{N+1}$. Therefore

$$\begin{aligned}\frac{(\mathbf{1}_B, Q'\mathbf{1}_{B^c})_{\pi_{st}}}{\pi_{st}[B]\pi_{st}[B^c]} &= \frac{1}{\pi_{st}[B^c]} \sum_{k\in D} \left(\frac{\phi_k[A]}{\sum_{i\in D}\phi_i[A]}\right) \sum_{l\in D^c} \frac{\delta(A,k,l)}{2(N+1)} \\ &\leq \frac{|D||D^c|}{2(N+1)\pi_{st}[B^c]} \max_{k\in D, l\in D^c} \left\{ \left(\frac{\phi_k[A]}{\sum_{i\in D}\phi_i[A]}\right) \delta(A,k,l) \right\} \\ &\leq \frac{|D|}{2} \max_{k\in D, l\in D^c} \left\{ \left(\frac{\phi_k[A]}{\sum_{i\in D}\phi_i[A]}\right) \delta(A,k,l) \right\} \\ &= \frac{|D| \max_{k\in D}\phi_k[A]}{2\sum_{k\in D}\phi_k[A]} \max_{k\in D, l\in D^c} \left\{ \left(\frac{\phi_k[A]}{\max_{i\in D}\phi_i[A]}\right) \delta(A,k,l) \right\} \\ &\leq \frac{1}{2c} \max_{k\in D, l\in D^c} \{\gamma(A,k)\delta(A,k,l)\}. \end{aligned}$$

Using this result and (5.11),

$$\max\left\{\frac{(\mathbf{1}_B, T'\mathbf{1}_{B^c})_{\pi_{st}}}{\pi_{st}[B]\pi_{st}[B^c]}, \frac{(\mathbf{1}_B, Q'\mathbf{1}_{B^c})_{\pi_{st}}}{\pi_{st}[B]\pi_{st}[B^c]}\right\}$$
$$\leq \max\left\{\max_{k\in D}\Phi_{T_k}(A), \frac{1}{2c}\max_{k\in D, l\in D^c}\gamma(A,k)\delta(A,k,l)\right\}. \quad (5.13)$$

Combining (5.10) and (5.13), we have that

$$\mathbf{Gap}(P_{st}) \leq 24\max\left\{\max_{k\in D}\Phi_{T_k}(A), \frac{1}{c}\max_{k\in D, l\in D^c}\gamma(A,k)\delta(A,k,l)\right\}. \quad (5.14)$$

Now define $\tilde{k}$ to be the $k \in D$ that maximizes $\phi_k[A]$. Also define $C = \{(z,k) \in \mathcal{X}_{st} : k = \tilde{k}, z \in A\}$. Since $0 < |D| < N+1$, we must have $N > 0$, so $\pi_{st}[C^c] \geq \frac{N}{N+1} \geq \frac{1}{2}$. Note that

$$\mathbf{Gap}(P_{st}) \leq 12\left[\frac{(\mathbf{1}_C, T'\mathbf{1}_{C^c})_{\pi_{st}}}{\pi_{st}[C]\pi_{st}[C^c]} + \frac{(\mathbf{1}_C, Q'\mathbf{1}_{C^c})_{\pi_{st}}}{\pi_{st}[C]\pi_{st}[C^c]}\right]. \quad (5.15)$$

Also note that

$$\begin{aligned}\frac{(\mathbf{1}_C, T'\mathbf{1}_{C^c})_{\pi_{st}}}{\pi_{st}[C]\pi_{st}[C^c]} &\leq \frac{2(\mathbf{1}_C, T'\mathbf{1}_{C^c})_{\pi_{st}}}{\pi_{st}[C]} \\ &= \frac{(\mathbf{1}_A, T_{\tilde{k}}\mathbf{1}_{A^c})_{\phi_{\tilde{k}}}}{\phi_{\tilde{k}}[A]} \\ &\leq \Phi_{T_{\tilde{k}}}(A) \leq \max_{k\in D}\Phi_{T_k}(A). \quad (5.16)\end{aligned}$$

Note that $\gamma(A,\tilde{k}) = 1$. Therefore

$$\begin{aligned}
\frac{(\mathbf{1}_C, Q'\mathbf{1}_{C^c})_{\pi_{st}}}{\pi_{st}[C]\pi_{st}[C^c]} &\leq \frac{2(\mathbf{1}_C, Q'\mathbf{1}_{C^c})_{\pi_{st}}}{\pi_{st}[C]} \\
&= \frac{1}{N+1}\sum_{l\in D, l\neq\tilde{k}} \delta(A,\tilde{k},l) + \frac{1}{N+1}\sum_{l\in D^c}\delta(A,\tilde{k},l) \\
&= \frac{1}{N+1}\sum_{l\in D, l\neq\tilde{k}} \delta(A,\tilde{k},l) + \frac{1}{N+1}\sum_{l\in D^c}\gamma(A,\tilde{k})\delta(A,\tilde{k},l) \\
&\leq \frac{1}{N+1}\sum_{l\in D, l\neq\tilde{k}} \frac{\phi_l[A]}{\phi_{\tilde{k}}[A]} + \frac{1}{N+1}\sum_{l\in D^c}\gamma(A,\tilde{k})\delta(A,\tilde{k},l) \\
&\leq c + \frac{1}{N+1}\sum_{l\in D^c}\gamma(A,\tilde{k})\delta(A,\tilde{k},l) \\
&\leq 2\max\{c, \max_{l\in D^c}\gamma(A,\tilde{k})\delta(A,\tilde{k},l)\} \\
&\leq 2\max\{c, \max_{k\in D, l\in D^c}\gamma(A,k)\delta(A,k,l)\}. \qquad (5.17)
\end{aligned}$$

Combining (5.15), (5.16), and (5.17), we obtain

$$\mathbf{Gap}(P_{st}) \leq 48\max\left\{\max_{k\in D}\Phi_{T_k}(A), c, \max_{k\in D, l\in D^c}\gamma(A,k)\delta(A,k,l)\right\}. \qquad (5.18)$$

Consider the case where

$$c < \left[\max_{k\in D, l\in D^c}\gamma(A,k)\delta(A,k,l)\right]^{1/2}. \qquad (5.19)$$

Since $\Phi_{T_k}(A) \leq 2$ we have $\Phi_{T_k}(A)/2 \leq [\Phi_{T_k}(A)/2]^{1/2}$. Therefore in the case where (5.19) holds, (5.18) implies that

$$\begin{aligned}
\mathbf{Gap}(P_{st}) &\leq 48\max\left\{\max_{k\in D}\Phi_{T_k}(A), \left[\max_{k\in D, l\in D^c}\gamma(A,k)\delta(A,k,l)\right]^{1/2}\right\} \\
&\leq 96\left[\max\left\{\max_{k\in D}\Phi_{T_k}(A), \max_{k\in D, l\in D^c}\gamma(A,k)\delta(A,k,l)\right\}\right]^{1/2} \qquad (5.20)
\end{aligned}$$

If instead we have

$$c \geq \left[\max_{k \in D, l \in D^c} \gamma(A,k)\delta(A,k,l) \right]^{1/2},$$

then (5.14) implies (5.20).

Note that for any k, l,

$$\gamma(A,k)\delta(A,k,l) \leq \min\left\{1, \frac{\phi_k[A]}{\phi_N[A]}\right\} \min\left\{1, \frac{\phi_l[A]}{\phi_k[A]}\right\} \leq \gamma(A,l)$$

where the final inequality can be shown by considering the case where $\phi_l[A] > \phi_k[A]$ separately from the case where $\phi_l[A] \leq \phi_k[A]$. For every $k \in D$, we have $\Phi_{T_k}(A)/2 \leq \gamma(A,k)$ so $\Phi_{T_k}(A)^2/4 \leq \gamma(A,k)\Phi_{T_k}(A)/2$. For all $l \in D^c$ such that $l \geq k^*$, we have $\gamma(A,l) < \Phi_{T_l}(A)/2$ so $\gamma(A,l)^2 < \gamma(A,l)\Phi_{T_l}(A)/2$. Consider (5.20) for the case where $k^* = 0$. Then

$$\begin{aligned} \mathbf{Gap}(P_{st}) &\leq 96 \left[\max\left\{ \max_{k \in D} \Phi_{T_k}(A), \max_{l \in D^c} \gamma(A,l) \right\} \right]^{1/2} \\ &\leq 192 \left[\max_{k} \gamma(A,k)\Phi_{T_k}(A) \right]^{1/4} \end{aligned} \tag{5.21}$$

Equation (5.21) is the bound in (5.4) for simulated tempering.

Now take $k^* > 0$ to show Theorem 5.1.2. Then (5.20) yields

$$\begin{aligned} \mathbf{Gap}(P_{st}) &\leq 96 \left[\max\left\{ \max_{k \in D} \Phi_{T_k}(A), \max_{l \geq k^*, l \in D^c} \gamma(A,l), \max_{k \in D, l < k^*} \gamma(A,k)\delta(A,k,l) \right\} \right]^{1/2} \\ &\leq 192 \left[\max\left\{ \left[\max_{k \geq k^*} \gamma(A,k)\Phi_{T_k}(A) \right]^{1/2}, \max_{k \in D, l < k^*} \gamma(A,k)\delta(A,k,l) \right\} \right]^{1/2} \\ &\leq 192 \left[\max\left\{ \max_{k \geq k^*} \gamma(A,k)\Phi_{T_k}(A), \max_{k \geq k^*, l < k^*} \gamma(A,k)\delta(A,k,l) \right\} \right]^{1/4} \end{aligned}$$

which implies Theorem 5.1.2 for simulated tempering.

Chapter 6

Multimodal Distributions for which Parallel and Simulated Tempering are Torpidly Mixing

In this chapter we will use Theorems 5.1.1 and 5.1.2 to show the torpid mixing of parallel and simulated tempering on several multimodal distributions.

6.1 Torpid Mixing on a Mixture of Normals with Unequal Variances in $\mathbb{R}^M$

Recall the definitions from Section 2.4.2. Let 1_M denote the vector of M ones, and I_M denote the $M \times M$ identity matrix. Consider the following mixture of two normal densities in $\mathbb{R}^M$ with unequal covariances:

$$\pi(z) = \frac{1}{2} N_M(z; -1_M, \sigma_1^2 \mathrm{I}_M) + \frac{1}{2} N_M(z; 1_M, \sigma_2^2 \mathrm{I}_M)$$

where $0 < \sigma_2 < \sigma_1$. Recall that S is the proposal kernel that is uniform on the ball of radius M^{-1} centered at the current state. For technical reasons we will use the following approximation to π, where $A_1 = \{z \in \mathbb{R}^M : \sum_i z_i < 0\}$ and $A_2 = \{z \in \mathbb{R}^M : \sum_i z_i \geq 0\}$:

$$\tilde{\pi}(z) \propto \frac{1}{2} N_M(z; -1_M, \sigma_1^2 \mathrm{I}_M) \mathbf{1}_{A_1}(z) + \frac{1}{2} N_M(z; 1_M, \sigma_2^2 \mathrm{I}_M) \mathbf{1}_{A_2}(z). \tag{6.1}$$

Metropolis-Hastings for S with respect to the density

$$\tilde{\pi}|_{A_1}(z) \propto N_M(z; -1_M, \sigma_1^2 \mathrm{I}_M) \mathbf{1}_{A_1}(z)$$

or with respect to

$$\tilde{\pi}|_{A_2}(z) \propto N_M(z; 1_M, \sigma_2^2 \mathrm{I}_M) \mathbf{1}_{A_2}(z)$$

is rapidly mixing in M, as implied by Theorem 4.3.4. However, Metropolis-Hastings for S with respect to $\tilde{\pi}$ is torpidly mixing in M, as we will show. We will also show that parallel and simulated tempering are torpidly mixing, regardless of the number and choice of temperatures. This is in contrast to the case where $\sigma_1 = \sigma_2$, for which rapidly mixing parallel and simulated tempering chains exist, as shown in Section 4.2.

First, calculate $\tilde{\pi}_\beta[A_2]$ as follows. Let Φ be the cumulative normal distribution function in one dimension. Consider any normal distribution in $\mathbb{R}^M$ with covariance $\sigma^2 \mathrm{I}_M$ for some $\sigma > 0$. Recall from Section 4.3 that the probability under this normal distribution of any half-space that is Euclidean distance d away from the center at its closest point is $\Phi(-d/\sigma)$. Note that the distance between the set A_2 and the point

-1_M is $\sqrt{M}$. Let λ denote Lebesgue measure. Then

$$\begin{aligned}
&\int_{A_1} N_M(z; -1_M, \sigma_1^2 \mathrm{I}_M)^\beta \lambda(dz) \\
&= (2\pi)^{-M\beta/2} \sigma_1^{-M\beta} \int_{A_1} \exp\left\{ -\frac{\beta}{2\sigma_1^2} \sum_i (z_i + 1)^2 \right\} \lambda(dz) \\
&= (2\pi)^{(M/2)(1-\beta)} \sigma_1^{M(1-\beta)} \beta^{-M/2} \int_{A_1} N_M(z; -1_M, \frac{\sigma_1^2}{\beta} \mathrm{I}_M) \lambda(dz) \\
&= (2\pi)^{(M/2)(1-\beta)} \sigma_1^{M(1-\beta)} \beta^{-M/2} \Phi\left(\frac{\sqrt{M}\beta^{1/2}}{\sigma_1} \right).
\end{aligned}$$

Similarly,

$$\int_{A_2} N_M(z; 1_M, \sigma_2^2 \mathrm{I}_M)^\beta \lambda(dz) = (2\pi)^{(M/2)(1-\beta)} \sigma_2^{M(1-\beta)} \beta^{-M/2} \Phi\left(\frac{\sqrt{M}\beta^{1/2}}{\sigma_2} \right).$$

Therefore

$$\frac{\tilde{\pi}_\beta[A_2]}{\tilde{\pi}_\beta[A_1]} = \left(\frac{\sigma_2}{\sigma_1} \right)^{M(1-\beta)} \frac{\Phi\left(\frac{\sqrt{M}\beta^{1/2}}{\sigma_2} \right)}{\Phi\left(\frac{\sqrt{M}\beta^{1/2}}{\sigma_1} \right)}.$$

Recall the definition of $\mathcal{B}$ from Section 2.3. Note that for the mixture $\tilde{\pi}$ we have $\mathcal{B} = (0, 1]$. We will apply Theorem 5.1.1 with $A = A_2$, showing that parallel tempering is torpidly mixing on the mixture $\tilde{\pi}$. Define the inverse temperaure $\beta^a = \frac{1}{2}$. Observe that for any inverse temperature β, $\Phi\left(\frac{\sqrt{M}\beta^{1/2}}{\sigma_1} \right) \geq \frac{1}{2}$. Therefore

$$\begin{aligned}
\sup_{\beta \in (0, \beta^a)} \tilde{\pi}_\beta[A_2] \leq \sup_{\beta \in (0, \beta^a)} \frac{\tilde{\pi}_\beta[A_2]}{\tilde{\pi}_\beta[A_1]} &\leq 2 \sup_{\beta \in (0, \beta^a)} \left(\frac{\sigma_2}{\sigma_1} \right)^{M(1-\beta)} \\
&= 2 \left(\frac{\sigma_2}{\sigma_1} \right)^{M(1-\beta^a)}
\end{aligned}$$

which is exponentially decreasing. Also note that $\tilde{\pi}[A_2] > \frac{1}{2}$. Therefore

$$\sup_{\beta \in [0, \beta^a) \cap \mathcal{B}} \gamma(A_2, \beta) \leq \sup_{\beta \in [0, \beta^a) \cap \mathcal{B}} \tilde{\pi}_\beta[A_2] / \tilde{\pi}[A_2] \tag{6.2}$$

is exponentially decreasing.

Define the boundary B_{A_2} of A_2 to be the set of $z \in A_2$ such that it is possible to move to A_1 via one move according to the proposal kernel S. Note that B_{A_2} is equal to the set of $z \in A_2$ such that z is within distance M^{-1} of the hyperplane $\sum_i z_i = 0$. Therefore

$$\begin{aligned}
\sup_{\beta \in [\beta^a, 1]} \frac{\tilde{\pi}_\beta[B_{A_2}]}{\tilde{\pi}_\beta[A_2]} &= \sup_{\beta \in [\beta^a, 1]} \frac{\Phi\left(\frac{\sqrt{M}\beta^{1/2}}{\sigma_2}\right) - \Phi\left(\frac{(\sqrt{M}-M^{-1})\beta^{1/2}}{\sigma_2}\right)}{\Phi\left(\frac{\sqrt{M}\beta^{1/2}}{\sigma_2}\right)} \\
&\leq 2 \sup_{\beta \in [\beta^a, 1]} \left\{ \Phi\left(\frac{\sqrt{M}\beta^{1/2}}{\sigma_2}\right) - \Phi\left(\frac{(\sqrt{M}-M^{-1})\beta^{1/2}}{\sigma_2}\right) \right\} \\
&\leq 2 \sup_{\beta \in [\beta^a, 1]} \left\{ 1 - \Phi\left(\frac{(\sqrt{M}-M^{-1})\beta^{1/2}}{\sigma_2}\right) \right\} \\
&= 2 \sup_{\beta \in [\beta^a, 1]} \left\{ \Phi\left(-\frac{(\sqrt{M}-M^{-1})\beta^{1/2}}{\sigma_2}\right) \right\} \\
&= 2\Phi\left(-\frac{(\sqrt{M}-M^{-1})(\beta^a)^{1/2}}{\sigma_2}\right).
\end{aligned}$$

For $M > 1$, this is bounded above by

$$2\Phi\left(-\frac{\sqrt{M}(\beta^a)^{1/2}}{2\sigma_2}\right). \tag{6.3}$$

Analytic integration shows that for any $a > 0$, $\Phi(-a) \leq N_1(a; 0, 1)/a$. Therefore (6.3) is exponentially decreasing. Similarly, for B_{A_1} equal to the boundary of A_1 with respect to the Metropolis-Hastings kernel,

$$\sup_{\beta \in [\beta^a, 1]} \frac{\tilde{\pi}_\beta[B_{A_1}]}{\tilde{\pi}_\beta[A_1]}$$

is exponentially decreasing. Recall that for any β, H_β is the Metropolis-Hastings kernel for S with respect to $\tilde{\pi}_\beta$. Then using the form of the conductance given in

(5.1),

$$\sup_{\beta \in [\beta^a, 1]} \Phi_{H_\beta}(A_2) \tag{6.4}$$

is exponentially decreasing. In particular, $\Phi_{H_\beta}(A_2)$ is exponentially decreasing for $\beta = 1$, so Metropolis-Hastings with respect to $\tilde{\pi}$ is torpidly mixing. Using the fact that (6.2) and (6.4) are exponentially decreasing, Theorem 5.1.1 implies that parallel and simulated tempering are also torpidly mixing.

6.2 Torpid Mixing on the Mean-Field Potts Model for $q \geq 3$

Recall the mean-field Potts model and the associated proposal kernel S as defined in Section 2.4.1. For $q \geq 3$ define $\alpha_c = \frac{2(q-1)\ln(q-1)}{q-2}$, and for $q = 2$ define $\alpha_c = 2$. The value $\alpha = \alpha_c$ is called the critical value, and we will see that the asymptotic properties of π are dramatically different for $\alpha = \alpha_c$, $\alpha < \alpha_c$, and $\alpha > \alpha_c$.

Metropolis-Hastings for S with respect to the mean-field Potts density with $q \geq 3$ and $\alpha \geq \alpha_c$ is torpidly mixing, as we will show. We will also use Theorem 5.1.1 to show that both parallel and simulated tempering are torpidly mixing for $q \geq 3$ and $\alpha \geq \alpha_c$. This builds on the results of Bhatnagar (2007), who showed the torpid mixing of parallel and simulated tempering for the mean-field Potts model with $q = 3$ and $\alpha = \alpha_c$. We use the same cut of the state space as Bhatnagar (2007), since it is a cut with low conductance for β close to 1. Unlike Bhatnagar (2007), we prove the torpid mixing result by showing that the persistence of one of the cut sets is

exponentially decreasing for small inverse temperatures.

For $k \in \{1, \ldots, q\}$ define $\sigma_k(z) = \sum_i \mathbf{1}(z_i = k)$. Note that π can be written

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

so the marginal distribution of the vector σ is

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

Let $a = (a_1, \ldots, a_q) = \sigma/M$ be the proportion of sites in each color. As in Gore and Jerrum (1999), we use Stirling's formula to write $\binom{M}{\sigma_1, \ldots, \sigma_q}$ as follows:

$$\binom{M}{\sigma_1, \ldots, \sigma_q} = \exp\left\{ \left(-\sum_{k=1}^{q} a_k \ln a_k \right) M + \Delta(a) \right\} \tag{6.5}$$

where $x \ln x$ for $x = 0$ is defined to be $\lim_{x \to 0^+} x \ln x = 0$ and where $\Delta(a)$ is an error term satisfying

$$\sup_a |\Delta(a)| = O(\ln M). \tag{6.6}$$

As in Gore and Jerrum (1999), use (6.5) to rewrite ρ as follows:

$$\rho(\sigma) \propto \exp\{ f_\alpha(a) M + \Delta(a) \}$$

where

$$f_\alpha(a) = \sum_{k=1}^{q} g_\alpha(a_k)$$

and $g_\alpha(x) = \frac{\alpha}{2} x^2 - x \ln x$. Note that f_α does not depend on M.

It is shown in Gore and Jerrum (1999) that any local maximum a^* of the function f_α takes the form $a^* = (x, \frac{1-x}{q-1}, \ldots, \frac{1-x}{q-1})$ or a permutation thereof for some $x \in [\frac{1}{q}, 1)$

that satisfies $g'_\alpha(x) = g'_\alpha(\frac{1-x}{q-1})$. Gore and Jerrum (1999) also show that for $q \geq 3$ and $\alpha = \alpha_c$, the local maxima occur for $x = \frac{1}{q}$ and $x = \frac{q-1}{q}$. Define $m_1 = (\frac{1}{q}, \ldots, \frac{1}{q})$, $m_2 = (\frac{q-1}{q}, \frac{1}{q(q-1)}, \ldots, \frac{1}{q(q-1)})$, and $m_3 = (\frac{1}{q(q-1)}, \frac{q-1}{q}, \frac{1}{q(q-1)}, \ldots, \frac{1}{q(q-1)})$. Note that

$$f_{\alpha_c}(m_1) = f_{\alpha_c}(m_2)$$

and that for any a, permuting the elements of a does not change the value of $f_\alpha(a)$. Therefore for $q \geq 3$ the $q+1$ local maxima of the function f_{α_c} are also global maxima. We will additionally need the following result:

Proposition 6.2.1. *For any $q \geq 3$ and $\alpha < \alpha_c$, f_α has a unique global maximum at m_1. For $\alpha > \alpha_c$, any global maximum of the function f_α takes the form $(x, \frac{1-x}{q-1}, \ldots, \frac{1-x}{q-1})$ or a permutation thereof for some $x \in \left[\frac{q-1}{q}, 1\right)$.*

Proof. For fixed a, the derivative of $f_\alpha(a)$ with respect to α is $\sum_k a_k^2/2$. Note that $\sum_k a_k^2$ has a unique global minimum at $a = m_1$. Let $m_{1,k}$ denote the kth element of m_1. Then for any $\alpha < \alpha_c$ and any $a \neq m_1$,

$$f_\alpha(a) - f_\alpha(m_1) = f_{\alpha_c}(a) - f_{\alpha_c}(m_1) + \frac{\alpha - \alpha_c}{2}\left[\sum_k a_k^2 - \sum_k m_{1,k}^2\right] < 0 \qquad (6.7)$$

so f_α has a unique global maximum at m_1.

Now consider $\alpha > \alpha_c$. For $x \in [0, 1]$ define $\hat{f}_\alpha(x) = f_\alpha(x, \frac{1-x}{q-1}, \ldots, \frac{1-x}{q-1})$. Note that for $a = (x, \frac{1-x}{q-1}, \ldots, \frac{1-x}{q-1})$, we have $\sum_k a_k^2 = x^2 + \frac{(1-x)^2}{q-1}$, which is a strictly increasing

function of x for $x \in [\frac{1}{q}, 1)$. Therefore for $x \in [\frac{1}{q}, \frac{q-1}{q})$,

$$\hat{f}_\alpha(x) - \hat{f}_\alpha(\frac{q-1}{q}) = \\ \hat{f}_{\alpha_c}(x) - \hat{f}_{\alpha_c}(\frac{q-1}{q}) + \frac{\alpha - \alpha_c}{2}\left[x^2 + \frac{(1-x)^2}{q-1} - \frac{(q-1)^2}{q^2} - \frac{1}{q^2(q-1)}\right] < 0.$$

This implies that $(x, \frac{1-x}{q-1}, \ldots, \frac{1-x}{q-1})$ is not a global maximum of f_α. The result follows.

□

Asymptotically, the distribution of $a(z)$ concentrates near the global maxima of $f_\alpha(a)$, meaning the following:

Proposition 6.2.2. *(Gore and Jerrum 1999) For any $q \in \{2, 3, \ldots\}$, $\alpha \geq 0$ and $\epsilon > 0$ define $C_{\alpha,\epsilon}$ to be the set of a such that a is less than Euclidean distance ϵ from at least one of the global maxima of f_α. Then $Pr(a(z) \in C^c_{\alpha,\epsilon})$ is exponentially decreasing in M. The probability that $a(z)$ is within distance ϵ of a particular global maximum of f_α decreases at most polynomially in M.*

Proof. This is stated in Gore and Jerrum (1999) for $\alpha = \alpha_c$; however, their argument extends to any α as follows. Let a^* be any global maximum of f_α. They observe that since f_α is continuous,

$$\sup_{a \in C^c_{\alpha,\epsilon}} f_\alpha(a) < f_\alpha(a^*). \tag{6.8}$$

Note that for any fixed a such that $a_k > 0$ for all k, adding or subtracting at most $1/M$ from a_k for each k changes $\exp\{f_\alpha(a)M\}$ by a factor that is bounded in M.

Recall that any global maximum a^* of $f_\alpha(a)$ satisfies $a_k^* > 0$ for all k. For every M there is some valid σ-vector σ^* such that σ_k^*/M is in the interval $[a_k^* - \frac{1}{M}, a_k^* + \frac{1}{M}]$ for each k. For this σ^*, $\exp\{f_\alpha(\frac{\sigma^*}{M})M\}$ and $\exp\{f_\alpha(a^*)M\}$ differ by a factor that is bounded in M. Using (6.8), $\rho(\sigma^*)$ is exponentially larger than $\max_{\sigma:\sigma/M \in C^c_{\alpha,\epsilon}} \rho(\sigma)$. Since there are a polynomial ($< M^{q-1}$) number of valid σ-vectors, $\Pr(a(z) \in C^c_{\alpha,\epsilon})$ is exponentially decreasing. Using the same facts, the probability that $a(z)$ is less than distance ϵ from a particular global maximum of f_α decreases at most polynomially in M, proving Proposition 6.2.2. □

As in Bhatnagar (2007), define $A = \{z : \sigma_1(z) > \frac{M}{2}\}$. Note that both A and A^c are nonempty. Then we have the following result.

Proposition 6.2.3. *For any fixed $q \geq 3$ and $\alpha \geq \alpha_c$, $\pi[A]$ and $\pi[A^c]$ decrease at most polynomially in M. For any $q \geq 3$ and $\alpha < \alpha_c$, $\pi[A]$ is exponentially decreasing in M. Furthermore, for any $\tau \in (0, \alpha_c)$, $\sup_{\alpha < \alpha_c - \tau} \pi[A]$ is also exponentially decreasing.*

Proof. First consider the case where $\alpha \geq \alpha_c$. By Proposition 6.2.1, there is a global maximum of f_α at $(x, \frac{1-x}{q-1}, \ldots, \frac{1-x}{q-1})$ for some $x \in [\frac{q-1}{q}, 1)$, and another at $(\frac{1-x}{q-1}, x, \frac{1-x}{q-1}, \ldots, \frac{1-x}{q-1})$. Since $q \geq 3$, if $a(z)$ is within distance $1/12$ of the first global maximum then $z \in A$, and if $a(z)$ is within distance $1/12$ of the second global maximum, then $z \in A^c$. By Proposition 6.2.2, $\pi[A]$ and $\pi[A^c]$ are therefore decreasing at most polynomially in M.

Now consider $\alpha < \alpha_c$. Then Propositions 6.2.1 and 6.2.2 imply that $\pi[A]$ is exponentially decreasing, since $\{a : a_1 > \frac{1}{2}\} \subset C^c_{\alpha,1/12}$. Recall from the proof of Proposition 6.2.1 that for any a, the derivative of $f_\alpha(a)$ with respect to α is $\sum_k a_k^2/2$, which has a unique global minimum at m_1. Now take any $\tau \in (0, \alpha_c)$. Note that

$$\begin{aligned}\sup_{\alpha<\alpha_c-\tau} \sup_{a:a_1>\frac{1}{2}} [f_\alpha(a) - f_\alpha(m_1)] &= \sup_{a:a_1>\frac{1}{2}} [f_{\alpha_c-\tau}(a) - f_{\alpha_c-\tau}(m_1)] \\ &\leq \sup_{a\in C^c_{\alpha_c-\tau,1/12}} [f_{\alpha_c-\tau}(a) - f_{\alpha_c-\tau}(m_1)] < 0\end{aligned}$$

where the last inequality uses (6.8). By the same argument as for Proposition 6.2.2, $\sup_{\alpha<\alpha_c-\tau} \pi[A]$ is exponentially decreasing. □

We also have the following result:

Proposition 6.2.4. *Recall the proposal kernel S, which changes the color of a single site. For $q \geq 3$ there exists some $\tau \in (0, \alpha_c)$ such that the supremum over $\alpha \geq \alpha_c - \tau$ of the conductance of A under Metropolis-Hastings for S with respect to π is exponentially decreasing.*

Proof. Since $q \geq 3$, $\{a : a_1 \in (\frac{5}{12}, \frac{1}{2}]\} \subset C^c_{\alpha_c,1/12}$. Using (6.8),

$$\sup_{a:a_1\in(\frac{5}{12},\frac{1}{2}]} [f_{\alpha_c}(a) - f_{\alpha_c}(m_2)] \leq \sup_{a\in C^c_{\alpha_c,1/12}} [f_{\alpha_c}(a) - f_{\alpha_c}(m_2)] < 0.$$

Since f_α is a continuous function of α and of a, there exists some $\tau > 0$ such that

$$c \stackrel{\text{def}}{=} \sup_{a:a_1\in(\frac{5}{12},\frac{1}{2}]} [f_{\alpha_c-\tau}(a) - f_{\alpha_c-\tau}(m_2)] < 0.$$

Recall that for any a, the derivative of $f_\alpha(a)$ with respect to α is $\sum_k a_k^2/2$. Note that for all a such that $a_1 \in (\frac{5}{12}, \frac{1}{2}]$, we have $\sum_k a_k^2 \leq 2(\frac{1}{2})^2 \leq (\frac{q-1}{q})^2 + (q-1)(\frac{1}{q(q-1)})^2 = \sum_k m_{2,k}^2$. Therefore for any such a, $[f_\alpha(a) - f_\alpha(m_2)]$ is a decreasing function of α. As a result,

$$\sup_{\alpha \geq \alpha_c - \tau} \sup_{a: a_1 \in (\frac{5}{12}, \frac{1}{2}]} [f_\alpha(a) - f_\alpha(m_2)] = c < 0. \tag{6.9}$$

Consider the boundary of A^c with respect to the Metropolis-Hastings kernel, meaning the set of $z \in A^c$ such that it is possible to move to A in one step via the Metropolis-Hastings kernel. This boundary is $B = \{z : \sigma_1(z) \in \{\frac{M-1}{2}, \frac{M}{2}\}\}$. Note that for M large enough, $z \in B \Rightarrow a_1(z) \in (\frac{5}{12}, \frac{1}{2}]$. There is some valid σ-vector σ^* such that $\frac{\sigma_k^*}{M} \in [m_{2,k} - \frac{1}{M}, m_{2,k} + \frac{1}{M}]$ for all k. For M large enough, if $\sigma(z) = \sigma^*$ then $z \in A$. Using (6.9) and by the same argument as for Proposition 6.2.2,

$$\sup_{\alpha \geq \alpha_c - \tau} \pi[B]/\pi[A]$$

is exponentially decreasing. Replacing m_2 by m_3 in (6.9) does not change the value of the left hand side, so it is still strictly negative. There is some valid σ-vector σ^* such that $\frac{\sigma_k^*}{M} \in [m_{3,k} - \frac{1}{M}, m_{3,k} + \frac{1}{M}]$ for all k. For M large enough, if $\sigma(z) = \sigma^*$ then $z \in A^c$. By the same argument as for Proposition 6.2.2,

$$\sup_{\alpha \geq \alpha_c - \tau} \pi[B]/\pi[A^c]$$

is exponentially decreasing. Therefore the supremum over $\alpha \geq \alpha_c - \tau$ of the conductance of A is exponentially decreasing. □

Now consider any $q \geq 3$ and $\alpha \geq \alpha_c$. Proposition 6.2.4 implies that Metropolis-Hastings with respect to π is torpidly mixing. Observe that for any β, the density π_β is equal to the mean-field Potts density with parameter $\alpha\beta$. Recall that H_β is the Metropolis-Hastings kernel for S with respect to π_β. Take the value of τ from Proposition 6.2.4. Define the inverse temperature $\beta^a = \alpha_c/\alpha - \tau/\alpha$. Propositions 6.2.3 and 6.2.4 imply that

$$\sup_{\beta \in [\beta^a, 1]} \Phi_{H_\beta}(A)$$

and

$$\sup_{\beta \in [0, \beta^a)} \gamma(A, \beta) \leq \sup_{\beta \in [0, \beta^a)} \frac{\pi_\beta[A]}{\pi[A]}$$

are exponentially decreasing. Therefore Theorem 5.1.1 implies that parallel and simulated tempering are torpidly mixing.

6.3 Torpid Mixing on the Mean-Field Ising Model using Fixed Temperatures

The mean-field Ising model is the mean-field Potts model with $q = 2$. Recall the definitions from Section 6.2 for the mean-field Potts model. It is straightforward to show that for $q = 2$ and $\alpha > \alpha_c$, the conductance of A under Metropolis-Hastings for S with respect to the mean-field Potts (Ising) density π is exponentially decreasing. Therefore Metropolis-Hastings for S with respect to π is torpidly mixing. Madras and Zheng (2003) show that parallel and simulated tempering with $N = M$ and $\beta_k = k/N$ are rapidly mixing for the mean-field Ising model. We will show that if N

and $\{\beta_k : k = 0, \ldots, N\}$ are fixed in M then parallel and simulated tempering are torpidly mixing for $\alpha > \alpha_c$. We will need the following result:

Proposition 6.3.1. *For $\alpha \leq \alpha_c$, f_α has a unique global maximum at $a = (\frac{1}{2}, \frac{1}{2})$. For $\alpha > \alpha_c$ the global maxima occur at $(x, 1-x)$ and $(1-x, x)$ for some $x > \frac{1}{2}$ that is strictly increasing with α.*

Proof. Recall from Section 6.2 that any local maximum a^* of the function f_α takes the form $a^* = (x, 1-x)$ or $a^* = (1-x, x)$ for some $x \in [\frac{1}{2}, 1)$ that satisfies $g'_\alpha(x) = g'_\alpha(1-x)$. This is trivially the case for $x = \frac{1}{2}$. Restricting to $x > \frac{1}{2}$ and rearranging we obtain that if $g'_\alpha(x) = g'_\alpha(1-x)$, then

$$\alpha = \frac{\ln(x) - \ln(1-x)}{2x - 1} \tag{6.10}$$

The right hand side of (6.10) is a strictly increasing function of x. It approaches $\alpha_c = 2$ as $x \to \frac{1}{2}^+$ and approaches infinity as $x \to 1^-$. Therefore for $\alpha \leq \alpha_c$, there is no $x > \frac{1}{2}$ that satisfies (6.10), so $a = (\frac{1}{2}, \frac{1}{2})$ is the unique maximum of f_α. For $\alpha > \alpha_c$ there is exactly one value of $x > \frac{1}{2}$ that satisfies (6.10), and that value is strictly increasing in α. Recall the definition of $\hat{f}_\alpha$ from the proof of Theorem 6.2.1. It is straightforward to verify that for $\alpha > \alpha_c$, $\hat{f}_\alpha$ is convex at $x = \frac{1}{2}$. Therefore the global maxima of f_α occur at $(x, 1-x)$ and $(1-x, x)$ for some $x > \frac{1}{2}$ that is strictly increasing in α. □

Now consider any α_1, α_2 such that $\alpha_c < \alpha_2$ and $\alpha_1 < \alpha_2$. If $\alpha_1 \leq \alpha_c$, let $x_1 = \frac{1}{2}$; otherwise, let x_1 be the value of x in Proposition 6.3.1 for α_1. Let x_2 be the value of x in Proposition 6.3.1 for α_2, so that $x_1 < x_2$. Let $\epsilon = |x_2 - x_1|/2$. Recalling the definition of $C_{\alpha,\epsilon}$ from Proposition 6.2.2, $C_{\alpha_1,\epsilon} \cap C_{\alpha_2,\epsilon} = \emptyset$. Letting π and π' be the mean-field Ising density at α_1 and α_2 respectively, Proposition 6.2.2 implies that $\pi[\{z : a(z) \in C^c_{\alpha_1,\epsilon}\}]$ and $\pi'[\{z : a(z) \in C^c_{\alpha_2,\epsilon}\}]$ are exponentially decreasing. Therefore $\sum_z \min\{\pi(z), \pi'(z)\}$ is exponentially decreasing.

Parallel and simulated tempering with $N = 0$ are equivalent to Metropolis-Hastings with respect to π, so they are torpidly mixing for $\alpha > \alpha_c$. Now consider the case where $N > 0$. Note that for $l \in \{0, \ldots, N-1\}$, π_{β_l} is the mean field Ising model with parameter $\alpha\beta_l$ and that $\pi_{\beta_N} = \pi$ is the mean-field Ising model with parameter α. Therefore with β_l fixed in M, $\sum_z \min\{\pi_{\beta_l}(z), \pi_{\beta_N}(z)\}$ is exponentially decreasing. Note that $\pi[A] \in [\frac{1}{4}, \frac{3}{4}]$ for all M. Therefore $\delta(A, N, l)$ and $\delta(A^c, N, l)$ are exponentially decreasing. By Theorem 5.1.2 with $k^* = N$, parallel and simulated tempering are torpidly mixing.

Chapter 7

Conclusions

We have given lower and upper bounds on the spectral gaps of parallel and simulated tempering. These imply lower and upper bounds on the rate of convergence to stationarity of these algorithms. We have used these bounds to obtain conditions for rapid and torpid mixing of parallel and simulated tempering. We have then given a number of distributions for which these conditions imply rapid or torpid mixing. These distributions include mixtures of normal distributions, approximations of which commonly occur in statistical inference, and the Potts model, which is used in statistical physics, in statistical image analysis, and for modeling of spatial random effects.

The lower and the upper bounds on the spectral gaps use closely related quantities. The overlap is common to both sets of bounds, and has been characterized previously

in a slightly different form. The persistence occurs in the upper bound and a closely related quantity occurs in the lower bound; these two quantities are equal in the examples that we have analyzed here. The importance of the persistence for the convergence of tempering algorithms has not previously been documented.

We have seen that parallel and simulated tempering are rapidly mixing on a weighted mixture of normals with identity covariance matrices. However, they are torpidly mixing on a mixture of normals where one covariance matrix is a multiple of the other, so that one of the normal densities has an exponentially higher maximum than the other. This suggests that if one mode is much narrower and higher than the other modes of a target distribution, then parallel and simulated tempering are slow to find this mode, and that if the modes are about equally narrow and high then parallel and simulated tempering mix quickly among the modes. This intuition holds for the mean-field Potts model as well; a disordered mode occurs for more than two colors, but does not occur for two colors. If it occurs, the disordered mode consists of many configurations of low probability, while an ordered mode consists of a few configurations of high probability. In this case, the ordered mode is much narrower and higher than the disordered mode, and parallel and simulated tempering are torpidly mixing. When there are only two colors (the mean-field Ising model) the disordered mode does not occur, and parallel and simulated tempering are rapidly mixing. For both the mean-field Potts model with $q \geq 3$ and the normal mixture with unequal covariances, there is a set containing the taller and narrower mode

that has exponentially decreasing persistence. We have shown that if the persistence of some subset is exponentially decreasing, then parallel and simulated tempering are torpidly mixing, and that if the persistence is monotonic and large enough at the highest temperature for every subset, then parallel and simulated tempering are rapidly mixing (assuming that the other conditions hold). We suspect that the persistence of a subset containing a single mode is a measure of the "spikiness" of the mode relative to the other modes, and that the above monotonicity condition is unnecessary. If this is true, then parallel and simulated tempering are rapidly mixing if and only if there is no mode that is exponentially "spikier" than another mode of the distribution (assuming that the other conditions are satisfied).

We have also shown that for distributions for which Metropolis-Hastings is torpidly mixing, the overlap of the parallel or simulated tempering chain must be polynomially decreasing in the problem size in order to have rapid mixing. This reinforces previous results on the importance of the overlap to the convergence of the parallel or simulated tempering chain.

Exact values for the spectral gap, numeric approximations thereof, or even tight bounds are very difficult to obtain for algorithms of interest and target distributions that are more than a few states. It this thesis we obtain order-of-complexity bounds in terms of relevant quantities. The relevant quantities can also be difficult to estimate for complex target densities–possibly as difficult as the original sampling problem. However, the bounds show what the relevant quantities are and how they affect the

magnitude of the spectral gap. We use the bounds to obtain the order of complexity of the spectral gap and of the relevant quantities as a function of the problem size for a number of examples. We then categorize these examples into distributions for which the spectral gap is exponentially decreasing (torpid mixing) and distributions for which the spectral gap is polynomially decreasing (rapid mixing).

The example densities that we use are simplified approximations of densities of interest. However, the normal examples may be able to be extended to create necessary and sufficient conditions for rapid mixing on finite mixtures of normal distributions. This would be very valuable since many target densities in statistics are well-approximated by finite mixtures of normals. In this way, the examples here of rapid and torpid mixing might be able to be extended to larger classes of distributions for which rapid or torpid mixing is implied by our bounds.

We conjecture that the conditions in this thesis for rapid and torpid mixing hold for any tempering-based sampling algorithm; that these properties are due to the use of tempered versions of π rather than to the specific construction of the algorithm. Tempering-based sampling algorithms include the "evolutionary Monte Carlo" algorithm of Liang and Wong (2000) and the "equi-energy" sampler of Kou et al. (2006). Supporting this hypothesis, we have been able to extend some of our torpid mixing results to the equi-energy sampler, despite its non-Markovian construction.

Bibliography

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004), *Hierarchical Modeling and Analysis for Spatial Data*, Chapman and Hall.

Bhatnagar, N. (2007), "Annealing and tempering for sampling and counting," Ph.D. thesis, Georgia Institute of Technology.

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 (SODA)*, pp. 478–487.

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

Caracciolo, S., Pelissetto, A., and Sokal, A. D. (1992), "Two remarks on simulated tempering," Unpublished manuscript.

Cowles, M. K. and Carlin, B. P. (1996), "Markov chain Monte Carlo convergence diagnostics: A comparative review," *Journal of the American Statistical Association*, 91, 883–904.

Diaconis, P. and Saloff-Coste, L. (1993), "Comparison theorems for reversible Markov chains," *Annals of Applied Probability*, 3, 696–730.

Diaconis, P. and Saloff-Coste, L. (1996), "Logarithmic Sobolev inequalities for finite Markov chains," *Annals of Applied Probability*, 6, 695–750.

Diaconis, P. and Stroock, D. (1991), "Geometric bounds for eigenvalues of Markov chains," *Annals of Applied Probability*, 1, 36–61.

Gelfand, A. E. and Smith, A. F. M. (1990), "Sampling-based approaches to calculating marginal densities," *Journal of the American Statistical Association*, 85, 398–409.

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.

Geweke, J. (1992), "Evaluating the accuracy of sampling-based approaches to calculating posterior moments," in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford: Clarendon Press.

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, Fairfax Station, VA: Interface Foundation of North America, pp. 156–163.

Geyer, C. J. and Thompson, E. A. (1995), "Annealing Markov chain Monte Carlo with applications to ancestral inference," *J. Amer. Statist. Assoc.*, 90, 909–920.

Gilks, W., Best, N., and Tan, K. (1995), "Adaptive rejection Metropolis sampling within Gibbs sampling," *Applied Statistics (Series C)*, 44, 455–472.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., eds. (1996), *Markov Chain Monte Carlo in Practice*, New York: Chapman and Hall.

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. (1995), "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, 82, 711–732.

Green, P. J. and Richardson, S. (2002), "Hidden Markov models and disease mapping," *J. Amer. Statist. Assoc.*, 97, 1055–1070.

Jerrum, M. and Sinclair, A. (1996), *Approximation Algorithms for NP-hard Problems*, Boston, MA: PWS Publishing, chapter 12, The Markov chain Monte Carlo method: an approach to approximate counting and integration.

Kannan, R. and Li, G. (1996a), "Sampling according to the multivariate normal density," in *Proceedings of the IEEE Symposium on Foundations of Computer Science*, pp. 204–213.

Kannan, R. and Li, G. (1996b), "Sampling according to the multivariate normal density, preliminary version," Unpublished manuscript.

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.

Lauritzen, S. (1996), *Graphical Models*, Oxford: Clarendon Press.

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. (2000), "Evolutionary Monte Carlo: Applications to C_p model sampling and change point problem," *Statistica Sinica*, 10, 317–342.

Liang, F. and Wong, W. H. (2001), "Real-parameter evolutionary Monte Carlo with applications to Bayesian mixture models," *Journal of the American Statistical Association*, 96, 653–666.

Madras, N. and Piccioni, M. (1999), "Importance sampling for families of distributions," *Annals of Applied Probability*, 9, 1202–1225.

Madras, N. and Randall, D. (2002), "Markov chain decomposition for convergence rate analysis," *The Annals of Applied Probability*, 12, 581–606.

Madras, N. and Slade, G. (1993), *The Self-Avoiding Walk*, Boston: Birkhauser.

Madras, N. and Zheng, Z. (2003), "On the swapping algorithm," *Random Structures and Algorithms*, 1, 66–97.

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

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), "Equation of state calculations by fast computing machines," *J. Chem. Phys.*, 21, 1087–1092.

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

Neal, R. M. (1996), *Bayesian Learning for Neural Networks*, New York: Springer-Verlag.

Predescu, C., Predescu, M., and Ciobanu, C. V. (2004), "The incomplete beta function law for parallel tempering sampling of classical canonical systems," *J. Chem. Phys.*, 120, 4119–4128.

Robert, C. and Casella, G. (1999), *Monte Carlo Statistical Methods*, New York: Springer-Verlag.

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 Tweedie, R. L. (2001), "Geometric L^2 and L^1 convergence are equivalent for reversible Markov chains," *Journal of Applied Probability*, 38A, 37–41.

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

Swendsen, R. H. and Wang, J. (1987), "Nonuniversal dynamics in Monte Carlo simulation," *Physical Review Letters*, 58, 86–88.

Tanner, M. A. and Wong, W. H. (1987), "The calculation of posterior distributions by data augmentation," *Journal of the American Statistical Association*, 82, 528–540.

Tierney, L. (1994), "Markov chains for exploring posterior distributions," *Annals of Statistics*, 22, 1701–1728.

Yuen, W. K. (2001), "Application of geometric bounds to convergence rates of Markov chains and Markov processes on $\mathbb{R}^n$," Ph.D. thesis, University of Toronto.

Zheng, Z. (2003), "On swapping and simulated tempering algorithms," *Stochastic Processes and their Applications*, 104, 131–154.

Biography

Dawn Woodard was born July 10, 1979 in Honolulu, Hawaii. She graduated from Stanford University with honors and distinction in Mathematical and Computational Science in 2001, and has won awards including the James B. Duke and University Scholarships at Duke and the Gertrude Cox Scholarship from the American Statistical Association. She has previously published in the area of statistical assessment of physician performance ("Performance Assessment for Radiologists Interpreting Screening Mammography" in Stastistics in Medicine, 2007). She has also presented at two International Society of Bayesian analysis meetings as well as a Breast Cancer Surveillance Consortium meeting and the Third Workshop on Monte Carlo Methods. She has worked for a number of statistical software companies including SAS and Insightful throughout undergraduate and graduate school, and has interned at Hewlett-Packard Laboratories.