Electronic Journal of Probability

Vol. 14 (2009), Paper no. 29, pages 780–804.

Journal URL
http://www.math.washington.edu/~ejpecp/

Sufficient Conditions for Torpid Mixing of Parallel and Simulated Tempering

Dawn B. Woodard,* Scott C. Schmidler,† Mark Huber‡

Abstract

We obtain upper bounds on the spectral gap of Markov chains constructed by parallel and simulated tempering, and provide a set of sufficient conditions for torpid mixing of both techniques. Combined with the results of [22], these results yield a two-sided bound on the spectral gap of these algorithms. We identify a *persistence* property of the target distribution, and show that it can lead unexpectedly to slow mixing that commonly used convergence diagnostics will fail to detect. For a multimodal distribution, the persistence is a measure of how "spiky", or tall and narrow, one peak is relative to the other peaks of the distribution. We show that this persistence phenomenon can be used to explain the torpid mixing of parallel and simulated tempering on the ferromagnetic mean-field Potts model shown previously. We also illustrate how it causes torpid mixing of tempering on a mixture of normal distributions with unequal covariances in $\mathbb{R}^M$, a previously unknown result with relevance to statistical inference problems. More generally, anytime a multimodal distribution includes both very narrow and very wide peaks of comparable probability mass, parallel and simulated tempering are shown to mix slowly.

Key words: Markov chain, rapid mixing, spectral gap, Metropolis algorithm.

*School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14853 (email dbw59@cornell.edu).

†Department of Statistical Science, Duke University, Durham, NC 27708 (email schmidler@stat.duke.edu).

‡Department of Mathematics, Duke University, Durham, NC 27708 (email mhuber@math.duke.edu).

AMS 2000 Subject Classification: Primary 65C40; Secondary: 60J2.

Submitted to EJP on September 4, 2008, final version accepted 2009.

1 Introduction

Parallel and simulated tempering [4; 13; 5] are Markov chain simulation algorithms commonly used in statistics, statistical physics, and computer science for sampling from multimodal distributions, where standard Metropolis-Hastings algorithms with only local moves typically converge slowly. Tempering-based sampling algorithms are designed to allow movement between modes (or "energy wells") by successively flattening the target distribution. Although parallel and simulated tempering have distinct constructions, they are known to have closely related mixing times; Zheng [24] bounds the spectral gap of simulated tempering below by a multiple of that of parallel tempering.

Madras and Zheng [12] first showed that tempering could be rapidly mixing on a target distribution where standard Metropolis-Hastings is torpidly mixing, doing so for the particular case of the mean-field Ising model from statistical physics. "Rapid" and "torpid" here are formalizations of the relative terms "fast" and "slow", and are defined in Section 2. However, Bhatnagar and Randall [2] show that for the more general ferromagnetic mean-field Potts model with $q \geq 3$, tempering is torpidly mixing for any choice of temperatures.

Woodard et al. [22] generalize the mean-field Ising example of [12] to give conditions which guarantee rapid mixing of tempering algorithms on general target distributions. They apply these conditions to show rapid mixing for an example more relevant to statistics, namely a weighted mixture of normal distributions in $\mathbb{R}^M$ with identity covariance matrices. In [22] the authors partition the state space into subsets on which the target distribution is unimodal. The conditions for rapid mixing of the tempering chain are that Metropolis-Hastings is rapidly mixing when restricted to any one of the unimodal subsets, that Metropolis-Hastings mixes rapidly among the subsets at the highest temperature, that the overlap between distributions at adjacent temperatures is decreasing at most polynomially in the problem size, and that an additional quantity γ (related to the persistence quantity of the current paper) is at most polynomially decreasing. These conditions follow from a lower bound on the spectral gaps of parallel and simulated tempering for general target distributions given in [22].

Here we provide complementary results, showing several ways in which the violation of these conditions implies torpid mixing of Markov chains constructed by parallel and simulated tempering. Most importantly, we identify a *persistence* property of distributions and show that the existence of any set with low conductance at low temperatures (e.g. a unimodal subset of a multimodal distribution) and having small persistence (as defined in Section 3 with interpretation in Section 5), guarantees tempering will mix slowly for any choice of temperatures. This result is troubling as this mixing problem will not be detected by standard convergence diagnostics (see Section 6).

We arrive at these results by deriving upper bounds on the spectral gaps of parallel and simulated tempering for arbitrary target distributions (Theorem 3.1 and Corollary 3.1). Combining with the lower bound in [22] then yields a two-sided bound.

In Section 4.2 we show that this persistence phenomenon can explain the torpid mixing of tempering techniques on the mean-field Potts model. The original result [2] uses a "bad cut" which partitions the space into two sets that have significant probability at temperature one, such that the boundary has low probability at all temperatures. We show that one of these partition sets has low persistence, also implying torpid mixing. We then show the persistence phenomenon for a mixture of normal distributions with unequal covariances in $\mathbb{R}^M$ (Section 4.1), thereby proving that tempering is torpidly mixing on this example. In typical cases such as these, the low-conductance set is a unimodal subset of a multimodal distribution. Then the persistence measures how "spiky", or narrow, this peak is relative to the other peaks of the distribution; this is described in Section 5, where we show that whenever the target distribution includes both very narrow and very wide peaks of comparable probability mass, simulated and parallel tempering mix slowly.

2 Preliminaries

Let $(\mathscr{X}, \mathscr{F}, \lambda)$ be a σ-finite measure space with countably generated σ-algebra $\mathscr{F}$. Often $\mathscr{X} = \mathbb{R}^M$ and λ is Lebesgue measure, or $\mathscr{X}$ is countable with counting measure λ . When we refer to an arbitrary subset $A \subset \mathscr{X}$, we implicitly assume $A \in \mathscr{F}$. Let P be a Markov chain transition kernel on $\mathscr{X}$, defined as in [19], which operates on distributions μ on the left and complex-valued functions f on the right, so that for $x \in \mathscr{X}$,

$$(\mu P)(dx) = \int \mu(dy) P(y, dx) \quad \text{and} \quad (Pf)(x) = \int f(y) P(x, dy).$$

If $\mu P = \mu$ then μ is called a stationary distribution of P. Define the inner product $(f, g)_\mu = \int \overline{f(x)} g(x) \mu(dx)$ and denote by $L_2(\mu)$ the set of complex-valued functions f such that $(f, f)_\mu < \infty$. P is *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 μ-reversible, it follows that μ is a stationary distribution of P. We will be primarily interested in distributions μ having a density π with respect to λ, in which case 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 aperiodic and ϕ-irreducible as defined in [16], μ-reversible, and nonnegative definite, then the Markov chain with transition kernel P converges in distribution to μ at a rate related to the *spectral gap*:

$$\textbf{Gap}(P) = \inf_{\substack{f \in L_2(\mu) \\ \text{Var}_\mu(f) > 0}} \left(\frac{\mathscr{E}(f, f)}{\text{Var}_\mu(f)} \right) \tag{1}$$

where $\mathscr{E}(f, f) = (f, (I - P)f)_\mu$ is a Dirichlet form, and $\text{Var}_\mu(f) = (f, f)_\mu - (f, 1)_\mu^2$ is the variance of f. It can easily be shown that $\textbf{Gap}(P) \in [0, 1]$ (for P not nonnegative definite, $\textbf{Gap}(P) \in [0, 2]$).

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 P as its transition kernel, define the rate of convergence to stationarity as:

$$r = \inf_{\mu_0} \lim_{n\to\infty} \frac{-\ln(\|\mu_0 P^n - \mu\|_2)}{n} \tag{2}$$

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 $-\ln(1-\mathbf{Gap}(P))$, where we define $-\ln(0) = \infty$; for every μ_0 that has a density $\pi_0 \in L_2(\mu)$,

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

and r is the largest quantity for which this holds for all such μ_0. These are facts from functional analysis (see e.g. [23; 11; 17]). Analogous results hold if the chain is started deterministically at x_0 for μ-a.e. $x_0 \in \mathscr{X}$, rather than drawn randomly from a starting distribution μ_0 [17]. Therefore for a particular such starting distribution μ_0 or fixed starting state x_0, the number of iterations n until the L_2-distance to stationarity is less than some fixed $\epsilon > 0$ is $O(r^{-1}\ln(\|\mu_0 - \mu\|_2))$. Similarly, [11] 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 $O(N_0 r^{-1}\ln(\|\mu_0 - \mu\|_2))$.

The quantity $r = -\ln(1-\mathbf{Gap}(P))$ is monotonically increasing with $\mathbf{Gap}(P)$; therefore lower (upper) bounds on $\mathbf{Gap}(P)$ correspond to lower (upper) bounds on r. 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 $\mathbf{Gap}(P)$ (and thus r) is exponentially decreasing as a function of the problem size, we call P *torpidly mixing*. When $\mathbf{Gap}(P)$ (and thus r) 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; polynomial factors are expected to be eventually dominated by increases in computing power due to Moore's law, while exponential factors are presumed to cause a persistent computational problem.

2.1 Metropolis-Hastings

The Metropolis-Hastings algorithm provides a common way of constructing a transition kernel that is π-reversible for a specified density π on a space $\mathscr{X}$ with measure λ. Start with a "proposal" kernel $P(w, dz)$ having density $p(w,\cdot)$ with respect to λ for all $w \in \mathscr{X}$, and define the Metropolis-Hastings kernel as follows: Draw a "proposal" move $z \sim P(w,\cdot)$ from current state w, accept z with probability

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

and otherwise remain at w. The resulting kernel is π-reversible.

2.2 Parallel and Simulated Tempering

If the Metropolis-Hastings proposal kernel moves only locally in the space, and if π is multimodal, 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. For any *inverse temperature* $\beta \in [0,1]$ such that $\int \pi(z)^\beta \lambda(dz) < \infty$, define

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

For any z and w in the support of π, the ratio $\pi_\beta(z)/\pi_\beta(w)$ monotonically approaches one as β decreases, flattening the resulting density. For any β, define T_β to be the Metropolis-Hastings chain with respect to π_β, or more generally assume that we have some way to specify a π_β-reversible transition kernel for each β, and call this kernel T_β.

Parallel tempering. Let $\mathscr{B} = \left\{\beta \in [0,1] : \int \pi(z)^\beta \lambda(dz) < \infty\right\}$. The parallel tempering algorithm [4] simulates parallel Markov chains T_{β_k} at a sequence of inverse temperatures $\beta_0 < \ldots < \beta_N = 1$ with $\beta_0 \in \mathscr{B}$. The inverse temperatures are commonly specified in a geometric progression, and Predescu et al. [15] 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 $\mathscr{X}_{pt} = \mathscr{X}^{N+1}$ and stationary density

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

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 move to guarantee nonnegative definiteness. The update move T chooses k uniformly from $\{0, \ldots, N\}$ and updates $x_{[k]}$ according to T_{β_k}:

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

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

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

PT1. sample k, l uniformly from $\{0, \ldots, N\}$ and propose exchanging the value of $x_{[k]}$ with that of $x_{[l]}$. Accept the proposed state, denoted $(k,l)x$, according to the Metropolis criteria preserving π_{pt}:

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

PT2. sample k uniformly from $\{0,\ldots,N-1\}$ and propose exchanging $x_{[k]}$ and $x_{[k+1]}$, accepting with probability $\rho(x,(k,k+1)x)$.

Both T and either form of Q are π_{pt}-reversible by construction, and nonnegative definite due to their 1/2 holding probability. Therefore the parallel tempering chain defined by $P_{pt} = QTQ$ is nonnegative definite and π_{pt}-reversible, and so the convergence of P_{pt}^n to π_{pt} may be bounded using the spectral gap of P_{pt}.

The above construction holds for any densities ϕ_k that are not necessarily tempered versions of π, by replacing T_{β_k} by any ϕ_k-reversible kernel T_k; the densities ϕ_k may be specified in any convenient way subject to $\phi_N = \pi$. The resulting chain is called a *swapping chain*, with $\mathscr{X}_{sc}$, λ_{sc}, P_{sc} and π_{sc} denoting its state space, measure, transition kernel, and stationary density respectively. Just as for parallel tempering, a swapping chain can be defined using swaps between adjacent levels only, or between arbitrary levels, and the two constructions will be denoted **SC2** and **SC1**, analogously to **PT2** and **PT1** for parallel tempering. Although the terms "parallel tempering" and "swapping chain" are used interchangeably in the computer science literature, we follow the statistics literature in reserving parallel tempering for the case of tempered distributions, and use swapping chain to refer to the more general case.

Simulated tempering. An alternative to simulating parallel chains is to augment a single chain by an inverse temperature index k to create states $(z,k) \in \mathscr{X}_{st} = \mathscr{X} \otimes \{0,\ldots,N\}$ with stationary density

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

The resulting *simulated tempering* chain [13; 5] alternates two types of moves: T' samples $z \in \mathscr{X}$ according to T_k, conditional on k, while Q' attempts to change k via one of the following schemes:

ST1. propose a new temperature 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 probability $\min\left\{1, \frac{\phi_l(z)}{\phi_k(z)}\right\}$, rejecting if $l = -1$ or $N+1$.

As before, a holding probability of 1/2 is added to both T' and Q'; the transition kernel of simulated tempering is defined as $P_{st} = Q'T'Q'$. For a lack of separate terms, we use "simulated tempering" to mean any such chain P_{st}, regardless of whether or not the densities ϕ_k are tempered versions of π.

3 Upper Bounds on the Spectral Gaps of Swapping and Simulated Tempering Chains

The parallel and simulated tempering algorithms described in Section 2.2 are designed to sample from multimodal distributions. Thus when simulating these chains, it is typically assumed that if the temperature swaps between all pairs of adjacent temperatures are occurring at a reasonable rate, then the chain is mixing well. However, Bhatnagar and Randall [2] show that parallel tempering is torpidly mixing for the ferromagnetic mean-field Potts model with $q \geq 3$ (Section 4.2), indicating that tempering does not work for all target distributions. It is therefore of significant practical interest to characterize properties of distributions which may make them amenable to, or inaccessible to, sampling using tempering algorithms.

In this Section we provide conditions for general target distributions π under which rapid mixing fails to hold. In particular, we identify a previously unappreciated property we call the *persistence*, and show that if the target distribution has a subset with low conductance for β close to one and low persistence for values of β within some intermediate β-interval, then the tempering chain mixes slowly. Somewhat more obviously, the tempering chain will also mix slowly if the inverse temperatures are spaced too far apart so that the overlap of adjacent tempered distributions is small.

Consider sets $A \subset \mathscr{X}$ 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 will be small at low temperatures ($\beta \approx 1$). The conductance of a set $A \subset \mathscr{X}$ with $0 < \mu(A) < 1$ is defined as:

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

for P any μ-reversible kernel on $\mathscr{X}$, where $\mathbf{1}_A$ is the indicator function of A. $\Phi_P(A)$ provides an upper bound on $\mathbf{Gap}(P)$ [9]. Note that P reversible implies $(\mathbf{1}_A, P\mathbf{1}_{A^c})_\mu = (\mathbf{1}_{A^c}, P\mathbf{1}_A)_\mu$, so

$$\Phi_P(A) = \frac{(\mathbf{1}_A, P\mathbf{1}_{A^c})_\mu}{\mu(A)} + \frac{(\mathbf{1}_{A^c}, P\mathbf{1}_A)_\mu}{\mu(A^c)} \tag{3}$$

and so $\Phi_P(A) \leq 2$.

We will obtain upper bounds on the spectral gap of a parallel or simulated tempering chain in terms of an arbitrary subset A of $\mathscr{X}$. Conceptually the case where $\pi|_A$ (the restriction of π to A) is unimodal as described above is the most insightful, but the bounds hold for all $A \subset \mathscr{X}$ such that $0 < \pi[A] < 1$.

The bounds will involve the conductance of A under the chain T_β defined in Section 2.2, as well as the *persistence* of A under tempering by β. For any $A \subset \mathscr{X}$ such that $0 < \pi[A] < 1$ and any density ϕ on $\mathscr{X}$, we define the quantity

$$\gamma(A, \phi) = \min\left\{1, \frac{\phi[A]}{\pi[A]}\right\} \tag{4}$$

and define the persistence of A with respect to π_β as $\gamma(A,\pi_\beta)$, also to be denoted by the shorthand $\gamma(A,\beta)$. The persistence measures the decrease in the probability of A between π and π_β. 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 high temperatures (β near zero). If A is a unimodal subset of a multimodal distribution, then it typically has low conductance for low temperatures ($\beta \approx 1$), so the tempering chain may take a long time to discover A at *all* temperatures even when $\pi[A]$ is large. This leads to slow mixing, and contradicts the common assumption in practice that if swapping acceptance rates between temperatures are high, the chain is mixing quickly. A key point is that, due to the low persistence of the set, this problem does *not* manifest as low conductance of the high-temperature chain which may well be rapidly mixing on π_β. Nevertheless, it *does* lead to slow mixing. This contradicts the common assumption in practice that if the highest temperature is rapidly mixing, and swapping acceptance rates between temperatures are high, then the tempering chain is rapidly mixing.

Even if every subset $A \subset \mathscr{X}$ has large persistence for high temperatures, it is possible for some subset to have low persistence within an intermediate temperature-interval. This causes slow mixing by creating a bottleneck in the tempering chain, since swaps between non-adjacent β and β' typically have very low acceptance probability. The acceptance probability of such a swap in simulated tempering, given that $z \in A$, is given by the *overlap* of π_β and $\pi_{\beta'}$ with respect to A. The overlap of two distributions ϕ and ϕ' with respect to a set $A \subset \mathscr{X}$ is given by [22]:

$$\delta(A,\phi,\phi') = \phi[A]^{-1} \int_A \min\{\phi(z),\phi'(z)\}\,\lambda(dz) \tag{5}$$

which is not symmetric. When considering tempered distributions π_β we will use the shorthand $\delta(A,\beta,\beta') = \delta(A,\pi_\beta,\pi_{\beta'})$.

The most general results are given for any swapping or simulated tempering chain with a set of densities ϕ_k not necessarily tempered versions of π. For any level $k \in \{0,\dots,N\}$, let $\gamma(A,k)$ and $\delta(A,k,l)$ be shorthand for $\gamma(A,\phi_k)$ and $\delta(A,\phi_k,\phi_l)$, respectively.

The following result, involving the overlap $\delta(A,k,l)$, the persistence $\gamma(A,k)$, and the conductance $\Phi_{T_k}(A)$, is proven in the Appendix:

Theorem 3.1. *Let P_{sc} be a swapping chain using scheme* **SC1** *or* **SC2***, and P_{st} a simulated tempering chain using scheme* **ST1**. *For any $A \subset \mathscr{X}$ such that $0 < \phi_k[A] < 1$ for all k, and for any $k^* \in \{0,\dots,N\}$, we have*

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

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

where for $k^* = 0$ we take this to mean:

$$\mathbf{Gap}(P_{sc}) \le 12 \max_k \{\gamma(A,k)\Phi_{T_k}(A)\}$$
$$\mathbf{Gap}(P_{st}) \le 192 \big[\max_k \{\gamma(A,k)\Phi_{T_k}(A)\} \big]^{1/4}.$$

One can obtain an alternative bound for the swapping chain by combining the bound for simulated tempering with the results of [24]. However, the alternative bound has a superfluous factor of N so we prefer the one given here.

For the case where tempered distributions $\phi_k = \pi_{\beta_k}$ are used, the bounds in Theorem 3.1 show that the inverse temperatures β_k must be spaced densely enough to allow sufficient overlap between adjacent temperatured distributions. If there is an $A \subset \mathscr{X}$ and a level k^* such that the overlap $\delta(A,k,l)$ is exponentially decreasing in M for every pair of levels $l < k^*$ and $k \ge k^*$, and the conductance $\Phi_{T_{\beta_k}}(A)$ of A is exponentially decreasing for $k \ge k^*$, then the tempering chain is torpidly mixing. An example is given in Section 4.3.

The bounds in Theorem 3.1 are given for a specific choice of densities $\{\phi_k\}_{k=0}^N$. When tempered densities are used, the bounds can be stated independent of the number and choice of inverse temperatures:

Corollary 3.1. *Let P_{pt} be a parallel tempering chain using scheme* **PT1** *or* **PT2***, and let P_{st} be a simulated tempering chain using scheme* **ST1***, with densities ϕ_k chosen as tempered versions of π. For any $A \subset \mathscr{X}$ such that $0 < \pi[A] < 1$, and any $\beta^* \ge \inf\{\beta \in \mathscr{B}\}$, we have*

$$\mathbf{Gap}(P_{pt}) \le 12 \sup_{\substack{\beta \in [\beta^*,1]\cap\mathscr{B} \\ \beta' \in [0,\beta^*)\cap\mathscr{B}}} \Big\{\gamma(A,\beta) \max\Big\{\Phi_{T_\beta}(A), \delta(A,\beta,\beta'), \delta(A^c,\beta,\beta')\Big\}\Big\}$$
$$\mathbf{Gap}(P_{st}) \le 192 \Bigg[\sup_{\substack{\beta \in [\beta^*,1]\cap\mathscr{B} \\ \beta' \in [0,\beta^*)\cap\mathscr{B}}} \Big\{\gamma(A,\beta) \max\Big\{\Phi_{T_\beta}(A), \delta(A,\beta,\beta')\Big\}\Big\} \Bigg]^{1/4}.$$

where for $\beta^* = \inf\{\beta \in \mathscr{B}\}$ we take this to mean:

$$\mathbf{Gap}(P_{pt}) \le 12 \sup_{\beta \in \mathscr{B}} \Big\{\gamma(A,\beta)\Phi_{T_\beta}(A)\Big\}$$
$$\mathbf{Gap}(P_{st}) \le 192 \big[\sup_{\beta \in \mathscr{B}} \Big\{\gamma(A,\beta)\Phi_{T_\beta}(A)\Big\} \big]^{1/4}.$$

This is a corollary of Theorem 3.1, verified by setting $k^* = \min\{k : \beta_k \ge \beta^*\}$.

Recall from Section 2 that torpid mixing of a Markov chain means that the spectral gap of the transition kernel is exponentially decreasing in the problem size. Then Corollary 3.1 implies the following result:

Corollary 3.2. *Assume that there exist inverse temperatures $\beta^* < \beta^{**}$ such that:*

1. *the conductance* $\sup_{\beta\in[\beta^{**},1]} \Phi_{T_\beta}(A)$ *is exponentially decreasing,*

2. *the persistence* $\sup_{\beta\in[\beta^{*},\beta^{**})\cap\mathscr{B}} \gamma(A,\beta)$ *is exponentially decreasing, and*

3. $\beta^* = \inf\{\beta \in \mathscr{B}\}$ *or the overlap* $\sup_{\substack{\beta\in[\beta^{**},1] \\ \beta'\in[0,\beta^*)\cap\mathscr{B}}} \max\{\delta(A,\beta,\beta'),\delta(A^c,\beta,\beta')\}$ *is exponentially decreasing.*

Then parallel and simulated tempering are torpidly mixing.

In Sections 4.1 and 4.2 we will give two examples where we use this corollary with $\beta^* = \inf\{\beta \in \mathscr{B}\}$ to show torpid mixing of parallel and simulated tempering. For this choice of β^*, condition 3 is automatically satisfied. Condition 3 is presumed to hold for most problems of interest, even when $\beta^* > \inf\{\beta \in \mathscr{B}\}$; otherwise, intermediate β values would not be needed at all. Thus the existence of a set A (e.g. with $\pi|_A$ unimodal) with low conductance for β close to 1, and low persistence for β in some intermediate β-interval, induces slow mixing of parallel and simulated tempering. It is possible to have a set A with low persistence in some intermediate β-interval and higher persistence for small β, since $\pi_\beta[A]$ is not necessarily a monotonic function of β (e.g. $\mathscr{X} = \{1,2,3\}$, $\pi = (0.01, 0.8, 0.19)$, and $A = \{1,2\}$).

The quantities in the upper bounds of this section are closely related to the quantities in the lower bounds on the spectral gaps of parallel and simulated tempering given in Woodard et al. [22]. The overlap quantity $\delta(\{A_j\})$ used by Woodard et al. [22] for an arbitrary partition $\{A_j\}_{j=1}^J$ of $\mathscr{X}$ is simply given by

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

The quantity $\gamma(\{A_j\})$ defined in [22] is related to the persistence of the current paper. If $\phi_k[A_j]$ is a monotonic function of k for each j, then

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

In addition, the conductance $\Phi_{T_k}(A)$ of the current paper is exactly the spectral gap of the *projection matrix* $\bar{T}_k$ for T_k with respect to the partition $\{A,A^c\}$, as defined in [22]. Since $\bar{T}_k$ is a 2×2 matrix, its spectral gap is given by the sum of the off-diagonal elements, which is precisely $\Phi_{T_k}(A)$ written in the form (3).

The lower bound given in [22] is, for any partition $\{A_j\}_{j=1}^J$ of $\mathscr{X}$,

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

where $T_k|_{A_j}$ is the restriction of the kernel T_k to the set A_j. Note the upper and lower bounds are stated for arbitrary sets and partitions respectively, and so also hold for the inf over sets A and sup over partitions $\{A_j\}$, respectively. The lower bound shows that if there is a partition $\{A_j\}$ of the space such that $\gamma(\{A_j\})$ is large and such that Metropolis-Hastings restricted to any one of the sets A_j is rapidly mixing, and if Metropolis-Hastings is rapidly mixing at the highest temperature and the overlap $\delta(\{A_j\})$ of adjacent levels is high, then the tempering chains P_{sc} and P_{st} are rapidly mixing. The conditions on $\gamma(\{A_j\})$ and the overlap are the important ones, since the other two conditions are typically satisfied for multimodal distributions of interest. By comparison, Theorem 3.1 shows that both the persistence $\gamma(A_j,k)$ and the overlap $\delta(A_j,k,l)$ must be large for each j in order to have rapid mixing (by setting $A = A_j$). Although the persistence $\gamma(A_j,k)$ is closely related to the quantity $\gamma(\{A_j\})$, the two are not identical so we do not have a single set of necessary and sufficient conditions for rapid mixing. However, our results suggest that the bounds in the current paper and in [22] contain the important quantities and no unnecessary quantities.

4 Examples

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

Consider sampling from a target distribution given by a mixture of two normal densities in $\mathbb{R}^M$:

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

where $N_M(z;\nu,\Sigma)$ denotes the multivariate normal density for $z \in \mathbb{R}^M$ with mean vector ν and $M \times M$ covariance matrix Σ, and 1_M and I_M denote the vector of M ones and the $M \times M$ identity matrix, respectively. Let S be the proposal kernel that is uniform on the ball of radius M^{-1} centered at the current state. When $\sigma_1 = \sigma_2$, Woodard et al. [22] have given an explicit construction of parallel and simulated tempering chains that is rapidly mixing. Here we consider the case $\sigma_1 \neq \sigma_2$, assuming without loss of generality that $\sigma_1 > \sigma_2$.

For technical reasons, we will use the following truncated approximation to π, where $A_1 = \{z \in \mathbb{R}^M : \sum_i z_i < 0\}$ and $A_2 = \{z \in \mathbb{R}^M : \sum_i z_i \geq 0\}$:

$$\tilde{\pi}(z) \propto \frac{1}{2}N_M(z; -1_M, \sigma_1^2 \mathrm{I}_M)\mathbf{1}_{A_1}(z) + \frac{1}{2}N_M(z; 1_M, \sigma_2^2 \mathrm{I}_M)\mathbf{1}_{A_2}(z). \tag{6}$$

Figure 1 shows $\tilde{\pi}_\beta[A_2]$ as a function of β for $M = 35$. It is clear that for $\beta < \frac{1}{2}$, $\tilde{\pi}_\beta[A_2]$ is much smaller than $\tilde{\pi}[A_2]$. This effect becomes more extreme as M increases, so that the persistence of A_2 is exponentially decreasing for $\beta < \frac{1}{2}$, as we will show. We will also show that the conductance of A_2 under Metropolis-Hastings for S with respect to $\tilde{\pi}_\beta$ is exponentially decreasing for $\beta \geq \frac{1}{2}$, implying the torpid mixing of parallel and simulated tempering.

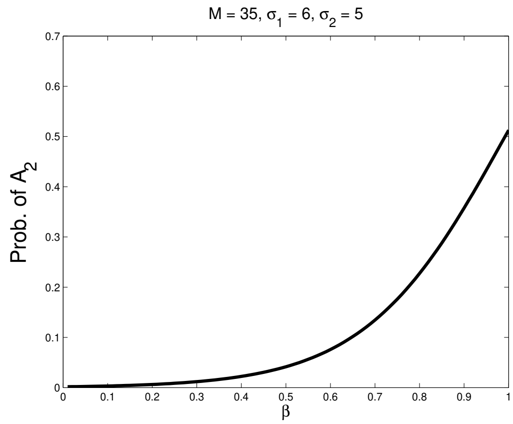

Figure 1: The probability of A_2 under $\tilde{\pi}_\beta$ as a function of β, for the mixture of normals with $M = 35$, $\sigma_1 = 6$, and $\sigma_2 = 5$.

The Metropolis-Hastings chains for S with respect to the densities restricted to each individual mode

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

are rapidly mixing in M, as implied by results in Kannan and Li [8] (details are given in Woodard [21]). As we will see however, Metropolis-Hastings for S with respect to $\tilde{\pi}$ itself is torpidly mixing in M. In addition, we will show that parallel and simulated tempering are also torpidly mixing for this target distribution for any choice of temperatures.

First, calculate $\tilde{\pi}_\beta[A_2]$ as follows. Let F 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 $\sigma > 0$. 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 $F(-d/\sigma)$. This is due to the independence of the dimensions and can be shown by a rotation and scaling in $\mathbb{R}^M$.

The distance between the half-space A_2 and the point -1_M is equal to $\sqrt{M}$. Therefore

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

and similarly

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

Therefore

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

Recall the definition of $\mathscr{B}$ from Section 2.2; for the mixture $\tilde{\pi}$, we have $\mathscr{B} = (0,1]$. We will apply Corollary 3.2 with $A = A_2$, $\beta^* = 0$, and $\beta^{**} = \frac{1}{2}$ to show that parallel and simulated tempering are torpidly mixing on the mixture $\tilde{\pi}$.

Looking first at the persistence $\gamma(A_2, \beta)$, since $F\big(\frac{(M\beta)^{1/2}}{\sigma_1}\big) > \frac{1}{2}$ we have

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

which is exponentially decreasing in M. Therefore since $\tilde{\pi}[A_2] > \frac{1}{2}$,

$$\sup_{\beta\in[0,\beta^{**})\cap\mathscr{B}} \gamma(A_2,\beta) \le \sup_{\beta\in[0,\beta^{**})\cap\mathscr{B}} \frac{\tilde{\pi}_\beta[A_2]}{\tilde{\pi}[A_2]} < 2 \sup_{\beta\in[0,\beta^{**})\cap\mathscr{B}} \tilde{\pi}_\beta[A_2] \tag{7}$$

is also exponentially decreasing.

Turning now to the conductance $\Phi_{T_\beta}(A_2)$, define the boundary ∂A_2 of A_2 with respect to the Metropolis-Hastings kernel T_β as the set of $z \in A_2$ such that it is possible to move to A_1 via one move according to T_β. Then ∂A_2 contains only $z \in A_2$ within distance M^{-1} of A_1. Therefore

$$\begin{aligned}
\sup_{\beta\in[\beta^{**},1]} \frac{\tilde{\pi}_\beta[\partial A_2]}{\tilde{\pi}_\beta[A_2]} &= \sup_{\beta\in[\beta^{**},1]} \left\{ \frac{F\big(\frac{(M\beta)^{\frac{1}{2}}}{\sigma_2}\big) - F\big(\frac{(M^{\frac{1}{2}}-M^{-1})\beta^{\frac{1}{2}}}{\sigma_2}\big)}{F\big(\frac{(M\beta)^{\frac{1}{2}}}{\sigma_2}\big)} \right\} \\
&\le 2 \sup_{\beta\in[\beta^{**},1]} \left\{ F\big(\frac{(M\beta)^{\frac{1}{2}}}{\sigma_2}\big) - F\big(\frac{(M^{\frac{1}{2}}-M^{-1})\beta^{\frac{1}{2}}}{\sigma_2}\big) \right\} \\
&\le 2 \sup_{\beta\in[\beta^{**},1]} \left\{ 1 - F\big(\frac{(M^{\frac{1}{2}}-M^{-1})\beta^{\frac{1}{2}}}{\sigma_2}\big) \right\} \\
&= 2 \sup_{\beta\in[\beta^{**},1]} \left\{ F\big(-\frac{(M^{\frac{1}{2}}-M^{-1})\beta^{\frac{1}{2}}}{\sigma_2}\big) \right\} \\
&= 2F\big(-\frac{(M^{\frac{1}{2}}-M^{-1})(\beta^{**})^{\frac{1}{2}}}{\sigma_2}\big).
\end{aligned}$$

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

$$2F\Big(-\frac{(M\beta^{**})^{\frac{1}{2}}}{2\sigma_2}\Big). \tag{8}$$

Analytic integration shows for any $a > 0$ that $F(-a) \leq N_1(a; 0, 1)/a$. Therefore 8 is exponentially decreasing in M. Analogously, for the boundary ∂A_1 of A_1 with respect to the Metropolis-Hastings kernel,

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

is exponentially decreasing. Therefore the conductance

$$\sup_{\beta \in [\beta^{**}, 1]} \Phi_{T_\beta}(A_2) \tag{9}$$

is exponentially decreasing. In particular, $\Phi_{T_\beta}(A_2)$ is exponentially decreasing for $\beta = 1$, so the standard Metropolis-Hastings chain is torpidly mixing. Using the above facts that (7) and (9) are exponentially decreasing, Corollary 3.2 implies that parallel and simulated tempering are also torpidly mixing for any number and choice of temperatures.

4.2 Small Persistence for the Mean-Field Potts Model

The Potts model is a type of discrete Markov random field which arises in statistical physics, spatial statistics, and image processing [1; 3; 7]. We consider the ferromagnetic mean-field Potts model with $q \geq 2$ colors and M sites, having distribution:

$$\pi(z) \propto \exp\left\{\frac{\alpha}{2M}\sum_{i,j} \mathbf{1}(z_i = z_j)\right\} \qquad \text{for} \qquad z \in \{1, \ldots, q\}^M$$

with interaction parameter $\alpha \geq 0$. The mean-field Potts model exhibits a phase transition phenomenon similar to the more general Potts model, 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.

We will use the proposal kernel S that changes the color of a single site, where the site and color are drawn uniformly at random. It is well-known that Metropolis-Hastings for S with respect to π is torpidly mixing for $\alpha \geq \alpha_c$ [6]. Bhatnagar and Randall [2] show that parallel and simulated tempering are also torpidly mixing on the mean-field Potts model with $q = 3$ and $\alpha = \alpha_c$ (their argument may extend to $q \geq 3$ and $\alpha \geq \alpha_c$). Here we show that this torpid mixing can be explained using the persistence phenomenon described in Section 3. We use the same cut of the state space as do Bhatnagar and Randall [2], since it has low conductance for β close to 1. Our torpid mixing explanation will be stated for $q \geq 3$ and $\alpha \geq \alpha_c$. Our initial definitions will be given for $q \geq 2$ to allow us to address the case $q = 2$ in Section 4.3.

Define $\sigma(z) = (\sigma_1(z), \ldots, \sigma_q(z))$ to be the vector of sufficient statistics, where $\sigma_k(z) = \sum_i \mathbf{1}(z_i = k)$. Then π can be written as

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

and the marginal distribution of σ is given by

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

For $q \geq 3$ define the "critical" parameter value $\alpha_c = \frac{2(q-1)\ln(q-1)}{q-2}$; for $q = 2$ set $\alpha_c = 2$. Let $a = (a_1, \ldots, a_q) = \sigma/M$ be the proportion of sites in each color. Using Stirling's formula, Gore and Jerrum [6] write $\binom{M}{\sigma_1,\ldots,\sigma_q}$ as:

$$\binom{M}{\sigma_1, \ldots, \sigma_q} = \exp\left\{-M\sum_{k=1}^{q} a_k \ln a_k + \Delta(a)\right\} \tag{10}$$

where $\Delta(a)$ is an error term satisfying

$$\sup_a |\Delta(a)| = O(\ln M). \tag{11}$$

Gore and Jerrum [6] apply (10) to rewrite ρ as:

$$\rho(\sigma) \propto \exp\{f_\alpha(a)M + \Delta(a)\} \qquad \text{where} \qquad f_\alpha(a) = \sum_{k=1}^{q} g_\alpha(a_k)$$

and $g_\alpha(x) = \frac{\alpha}{2}x^2 - x\ln x$. Observe that f_α does not depend on M. It is also shown in [6] that any local maximum of f_α is of the form $m = (x, \frac{1-x}{q-1}, \ldots, \frac{1-x}{q-1})$ for some $x \in [\frac{1}{q}, 1)$ satisfying $g'_\alpha(x) = g'_\alpha(\frac{1-x}{q-1})$, or a permutation thereof (the apostrophe denoting the first derivative). Gore and Jerrum also show that at $\alpha = \alpha_c$ the local maxima occur for $x = \frac{1}{q}$ and $x = \frac{q-1}{q}$.

Letting $m^1 = (\frac{1}{q}, \ldots, \frac{1}{q})$, $m^2 = (\frac{q-1}{q}, \frac{1}{q(q-1)}, \ldots, \frac{1}{q(q-1)})$, and m^3 equal to m^2 with the first two elements permuted, note that

$$f_{\alpha_c}(m^1) = f_{\alpha_c}(m^2)$$

and that for any a, $f_\alpha(a)$ is invariant under permutation of the elements of a. Therefore the $q+1$ local maxima of the function f_{α_c} are also global maxima (for $q = 2$ there is a single global maximum). For M large enough the $q+1$ global maxima of f_{α_c} correspond to $q+1$ local maxima of $\rho(\sigma)$; Figure 2 shows the 4 modes of $\rho(\sigma)$ for the case $q = 3$.

We will additionally need the following results. The proofs are given in the thesis by Woodard [20].

Proposition 4.1. *For any $q \geq 3$ and $\alpha < \alpha_c$, f_α has a unique global maximum at m^1, while for $\alpha > \alpha_c$ every global maximum of f_α is of the form $(x, \frac{1-x}{q-1}, \ldots, \frac{1-x}{q-1})$ for some $x \in \left[\frac{q-1}{q}, 1\right)$, or a permutation thereof.*

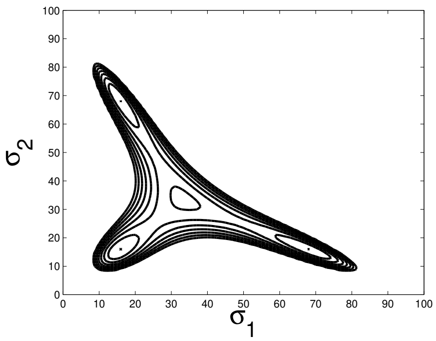

Figure 2: A contour plot of the marginal distribution $\rho(\sigma)$ of the sufficient statistic vector σ, as a function of σ_1 and σ_2, for the mean-field Potts model with $q = 3$, $M = 100$, and $\alpha \approx \alpha_c$.

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

Proposition 4.2. *(Gore and Jerrum 1999) For any fixed $q \geq 2$, $\alpha \geq 0$ and $\epsilon > 0$, let*

$$C_{\alpha,\epsilon} = \{a : \|a - m\| < \epsilon \text{ for some } m \in \mathscr{M}\}$$

where $\mathscr{M}$ are the global maxima of f_α and $\|\|$ indicates Euclidean distance. Then $Pr(a(z) \in C^c_{\alpha,\epsilon})$ is exponentially decreasing in M, while for any specific $m \in \mathscr{M}$, $Pr(\|a(z) - m\| < \epsilon)$ decreases at most polynomially in M.

Gore and Jerrum state this result for $\alpha = \alpha_c$, but their argument can be extended in a straightforward manner; details are given in [20].

As in Bhatnagar and Randall [2], define the set $A = \{z : \sigma_1(z) > \frac{M}{2}\}$. Then we have the following two results, also shown in [20].

Proposition 4.3. *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 $q \geq 3$ and $\tau \in (0, \alpha_c)$, $\sup_{\alpha < \alpha_c - \tau} \pi[A]$ is also exponentially decreasing.*

Proposition 4.4. *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 is exponentially decreasing.*

Now consider any $q \geq 3$ and $\alpha \geq \alpha_c$. For any β, the density π_β is equal to the mean-field Potts density with parameter $\alpha\beta$. Recall that T_β is the Metropolis-Hastings kernel for S with respect to π_β. Take the value of τ from Proposition 4.4. Define the inverse temperature $\beta^{**} = \alpha_c/\alpha - \tau/\alpha$. Propositions 4.3 and 4.4 imply that

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

and

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

are exponentially decreasing. Therefore Corollary 3.2 can be used to show the torpid mixing of parallel and simulated tempering on the mean-field Potts model with $q \geq 3$ and $\alpha \geq \alpha_c$, for any number and choice of inverse temperatures.

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

Consider the mean-field Ising model, which is simply the mean-field Potts model from Section 4.2 with $q = 2$. Recall the definitions from that section. Madras and Zheng [12] show that parallel and simulated tempering with $N = M$ and $\beta_k = k/N$ are rapidly mixing on the mean-field Ising model, while Metropolis-Hastings is torpidly mixing for $\alpha > \alpha_c$. As a demonstration of the importance of the overlap quantity in Theorem 3.1, we show here that if instead the number N of temperatures does not grow with M, then parallel and simulated tempering are torpidly mixing. We will need the following result, proven in the thesis [20]:

Proposition 4.5. *For $q = 2$ and $\alpha \leq \alpha_c$, f_α has a unique global maximum at $a = (\frac{1}{2}, \frac{1}{2})$. For $q = 2$ and $\alpha > \alpha_c$ the global maxima occur at $(x, 1-x)$ and $(1-x, x)$ for some $x > \frac{1}{2}$ that is strictly increasing 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 4.5 for α_1. Let x_2 be the value of x in Proposition 4.5 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 4.2, $C_{\alpha_1,\epsilon} \cap C_{\alpha_2,\epsilon} = \emptyset$. Letting π and π' be the mean-field Ising model density at α_1 and α_2 respectively, Proposition 4.2 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 $\pi_{\beta_N} = \pi$ 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 3.1 with $k^* = N$, parallel and simulated tempering are torpidly mixing.

5 Interpretation of Persistence

As described in Section 3, tempering algorithms mix slowly when there is a set $A \subset \mathscr{X}$ which has low conductance under the low-temperature ($\beta = 1$) chain and has small persistence for some range of

β-values. Here small persistence means $\pi_\beta[A]/\pi[A]$ near zero. To understand how the existence of such a set depends on the properties of π, we can rewrite this ratio as:

$$\begin{aligned}\frac{\pi_\beta[A]}{\pi[A]} &= \frac{\int_A \pi(z)^\beta \lambda(dz) / \int_{\mathscr{X}} \pi(z)^\beta \lambda(dz)}{\int_A \pi(z)\lambda(dz) / \int_{\mathscr{X}} \pi(z)\lambda(dz)} \\ &= \frac{\int_A \pi(z)^\beta \lambda(dz) / \int_A \pi(z)\lambda(dz)}{\int_{\mathscr{X}} \pi(z)^\beta \lambda(dz) / \int_{\mathscr{X}} \pi(z)\lambda(dz)} \\ &= \frac{E_{\pi|_A}(\pi(Z)^{\beta-1})}{E_\pi(\pi(Z)^{\beta-1})} \end{aligned} \tag{12}$$

where $\pi|_A$ is the restriction of π to A. Here $E_\pi(\pi(Z)^{\beta-1})$ denotes the expected value of the random variable $W = \pi(Z)^{\beta-1}$ where Z has distribution π.

Let Z_1 and Z_2 be random variables with distributions $\pi|_A$ and π, respectively, and define random variables $W_1 = \pi(Z_1)^{\beta-1}$ and $W_2 = \pi(Z_2)^{\beta-1}$. One way in which the ratio (12) may be smaller than one is if W_2 stochastically dominates W_1, or equivalently if the random variable $Y_1 = \pi(Z_1)$ stochastically dominates $Y_2 = \pi(Z_2)$. This means that within the set A the probability mass is concentrated in places where the density is high relative to those places where mass is concentrated for the rest of the space A^c. For example, if π consists of two peaks, one in A and the other in A^c, and $\pi(A) = \pi(A^c)$, then loosely speaking the peak in A is more “spiky”, or tall and narrow, than the peak in A^c.

As a concrete example, consider π an equal mixture of two trivariate normal distributions, with component means $\mu_1 = (10, 10, 10)$ and $\mu_2 = -\mu_1$ and covariance matrices $\Sigma_1 = 2I$ and $\Sigma_2 = 10I$. Define the set $A = \{z \in \mathbb{R}^3 : \sum_i z_i \geq 0\}$, which contains almost all of the probability mass of the first component and almost no mass from the second component. Figure 3 shows the cumulative distribution functions of the random variables Y_1 and Y_2 defined above, where it can be seen that Y_1 stochastically dominates Y_2. Intuitively this is because π has two peaks, one primarily in A and the other in A^c, with the first taller and more narrow than the other. As shown above, this stochastic dominance implies that the persistence $\gamma(A, \beta)$ is less than one for any $\beta \in (0, 1)$.

More generally, the persistence of a set A can be less than one whenever Y_1 tends to be larger than Y_2, in the sense that the transformation $W_1 = Y_1^{\beta-1}$ has a smaller expectation than $W_2 = Y_2^{\beta-1}$. Again this occurs when the probability mass within A is concentrated in regions of high density relative to the regions where mass concentrates in A^c. Again, if π consists of two peaks, one in A and one in A^c, and $\pi(A) = \pi(A^c)$, then informally speaking the peak in A is taller and more narrow than the peak in A^c.

Now take the more interesting case where π consists of multiple peaks of comparable probability, some of which are much taller than others; then the tallest peaks are also the narrowest peaks. Define A to contain one of these tall, narrow peaks. Since there are other peaks of the distribution that are much lower and wider, and none that are much taller and narrower, the expectation of

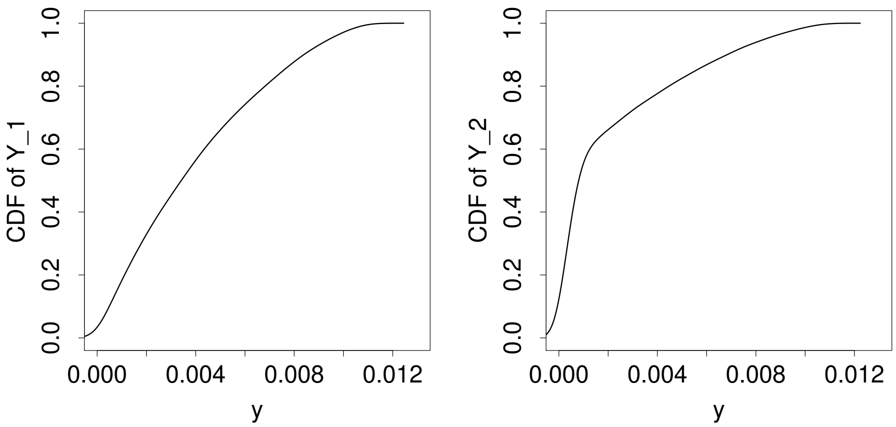

Figure 3: The cumulative distribution functions of Y_1 (left) and Y_2 (right), for the trivariate normal mixtures example.

W_1 is much smaller than that of W_2. The persistence of A is therefore small, and since A is a set having low conductance at low temperatures, the results in Section 3 imply that parallel and simulated tempering mix slowly. Here we mean slow mixing in a relative sense, that the smaller the persistence the slower the mixing, when other factors are held constant.

6 Conclusion

We have seen that if the multimodal target distribution has very wide peaks and very narrow peaks of comparable probability, then parallel and simulated tempering mix slowly. This means that if the simulated tempering chain is initialized in one of the wide peaks, or for parallel tempering if every level of the tempering chain is initialized in a wide peak, then the tempering chain will take a very large number of iterations to discover the narrow peaks of the distribution.

During application of simulated or parallel tempering, the acceptance rate of swap or temperature change moves is monitored, as are standard Markov chain convergence diagnostics. If the convergence diagnostics do not detect a problem, and if the acceptance rate for swap or temperature changes is high, then the tempering chain is presumed to be mixing well among the modes of the target distribution. However, we have shown that small persistence can cause slow mixing even when the acceptance rate for swaps or temperature changes, as measured by the quantity δ, is large. Additionally, standard Markov chain convergence diagnostics will rarely detect the problem; convergence diagnostics based on the history of the chain cannot detect the fact that there are undiscovered modes, unless they take into account some specialized knowledge about the distribution. Widely-used convergence diagnostics, such as time-series plots and autocorrelation plots, make few assumptions about the target distribution; these convergence diagnostics cannot infer features of the distribution in parts of the space that have not been explored. Even the Gelman-Rubin diagnostic, which is specifically designed to detect lack of convergence due to multimodality, works very poorly when some modes have a much smaller "basin of attraction" than others [21]. This is typically the case for the narrow peaks with which we are concerned.

When there are undiscovered modes, inferences based on samples from the tempering chain can be inaccurate. Practitioners should therefore be cautious about inferences that have been obtained using parallel and simulated tempering, just as for Metropolis-Hastings, and not presume that all the modes of the distribution have been discovered.

This slow mixing result is not surprising, since narrow peaks that have a small basin of attraction are extremely difficult to find in a large space. This has been called the "needle in a haystack" or "witch's hat" problem in the statistics literature, where it is recognized as causing difficulty for Metropolis-Hastings and Gibbs samplers [14]. We suspect that the problem of approximately sampling from a multimodal distribution containing very narrow peaks at unknown locations can be shown to be NP-complete (this question is addressed in [18]). If so, then parallel and simulated tempering fail in exactly the same situation that all other sampling methods would fail, namely for high-dimensional multimodal distributions with some very narrow peaks.

Acknowledgement

We thank the referees for their helpful suggestions. Additionally, the first author would like to thank Prof. Michael Lavine for helpful conversations regarding the Potts model example. The third author is partially supported by NSF grant DMS-05-48153.

References

[1] Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall, Boca Raton, FL.

[2] 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*. 478–487. MR2291087

[3] 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.

[4] 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*, E. Keramidas, Ed. Interface Foundation of North America, Fairfax Station, VA, 156–163.

[5] 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.

[6] Gore, V. K. and Jerrum, M. R. (1999). The Swendsen-Wang process does not always mix rapidly. *J. of Statist. Physics 97*, 67–85. MR1733467

[7] Green, P. J. and Richardson, S. (2002). Hidden Markov models and disease mapping. *J. Amer. Statist. Assoc. 97*, 1055–1070. MR1951259

[8] Kannan, R. and Li, G. (1996). Sampling according to the multivariate normal density. In *Proceedings of the IEEE Symposium on Foundations of Computer Science*. 204–213. MR1450618

[9] 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. MR0930082

[10] Madras, N. and Randall, D. (2002). Markov chain decomposition for convergence rate analysis. *Annals of Applied Probability 12*, 581–606. MR1910641

[11] Madras, N. and Slade, G. (1993). *The Self-Avoiding Walk*. Birkhauser, Boston. MR1197356

[12] Madras, N. and Zheng, Z. (2003). On the swapping algorithm. *Random Structures and Algorithms 1*, 66–97. MR1943860

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

[14] Matthews, P. (1993). A slowly mixing Markov chain with implications for Gibbs sampling. *Statistics and Probability Letters 17*, 231–236. MR1229942

[15] 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.

[16] Roberts, G. O. and Rosenthal, J. S. (2004). General state space Markov chains and MCMC algorithms. *Probability Surveys 1*, 20–71. MR2095565

[17] 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. MR1915532

[18] Schmidler, S. C. and Woodard, D. B. (2008). Computational complexity and Bayesian analysis. In preparation.

[19] Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics 22*, 1701–1728. MR1329166

[20] WOODARD, D. B. (2007a). Conditions for rapid and torpid mixing of parallel and simulated tempering on multimodal distributions. Ph.D. thesis, Duke University.

[21] WOODARD, D. B. (2007b). Detecting poor convergence of posterior samplers due to multimodality. Discussion Paper 2008-05, Duke University, Dept. of Statistical Science. http://ftp.isds.duke.edu/pub/WorkingPapers/08-05.html.

[22] WOODARD, D. B., SCHMIDLER, S. C., AND HUBER, M. (2009). Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. In press, *Annals of Applied Probability*.

[23] 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.

[24] ZHENG, Z. (2003). On swapping and simulated tempering algorithms. *Stochastic Processes and their Applications 104*, 131–154. MR1956476

Appendix A: Proof of the Spectral Gap Bounds

We will prove the bound in Theorem 3.1 for the swapping chain. The proof for simulated tempering is similar; see [20] for details. We will use the following results, which hold for any transition kernels P and Q that are reversible with respect to distributions μ_P and μ_Q on a space $\mathscr{X}$ with countably generated σ-algebra $\mathscr{F}$.

Lemma 6.1. *Let $\mu_P = \mu_Q$. If $Q(x, A\backslash\{x\}) \leq P(x, A\backslash\{x\})$ for every $x \in \mathscr{X}$ and every $A \subset \mathscr{X}$, then* $\mathbf{Gap}(Q) \leq \mathbf{Gap}(P)$.

Proof. As in Madras and Randall [10], write $\mathbf{Gap}(P)$ and $\mathbf{Gap}(Q)$ 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 the result is immediate. □ □

Lemma 6.2. *(Madras and Zheng 2003)*

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

Although Madras and Zheng [12] state Lemma 6.2 for finite state spaces, their proof extends easily to general spaces.

To prove Theorem 3.1, start by noting that by the definition of the spectral gap, $\mathbf{Gap}(P_{sc}) = \mathbf{Gap}(QTQ) = 8\mathbf{Gap}(\frac{1}{8}QTQ + \frac{7}{8}I)$ for a swapping chain P_{sc} as defined in Section 2.2. By Lemma 6.1, $\mathbf{Gap}(\frac{1}{8}QTQ + \frac{7}{8}I) \leq \mathbf{Gap}((\frac{1}{2}T + \frac{1}{2}Q)^3)$. By Lemma 6.2, $\mathbf{Gap}((\frac{1}{2}T + \frac{1}{2}Q)^3) \leq 3\mathbf{Gap}(\frac{1}{2}T + \frac{1}{2}Q)$. Therefore

$$\mathbf{Gap}(P_{sc}) \leq 24\mathbf{Gap}(\frac{1}{2}T + \frac{1}{2}Q). \tag{13}$$

Take any $A \subset \mathscr{X}$ such that $0 < \phi_k[A] < 1$ for all k, and any $k^* \in \{0, \ldots, N\}$. Define the set $B = \{x \in \mathscr{X}_{sc} : \forall k \geq k^*, x_{[k]} \in A^c\}$ for which all low-temperature chains are in A^c, so $\pi_{sc}[B^c] = 1 - \prod_{k \geq k^*} \phi_k[A^c]$. $\mathbf{Gap}(\frac{1}{2}T + \frac{1}{2}Q)$ is bounded above by the conductance of B under $(\frac{1}{2}T + \frac{1}{2}Q)$:

$$\begin{aligned}\mathbf{Gap}(\frac{1}{2}T + \frac{1}{2}Q) &\leq \Phi_{(\frac{1}{2}T+\frac{1}{2}Q)}(B) \\ &= \frac{1}{2}\Phi_T(B) + \frac{1}{2}\Phi_Q(B).\end{aligned} \tag{14}$$

For any $k \geq k^*$ we have $\pi_{sc}[B^c] \geq \max\{\phi_k[A], \phi_N[A]\}$; therefore

$$\begin{aligned}\Phi_T(B) &= \frac{1}{\pi_{sc}[B^c]}\Pr(\text{moving to } B^c \text{ via } T| \text{ in } B) \\ &= \frac{1}{\pi_{sc}[B^c]} \frac{1}{2(N+1)} \sum_{k \geq k^*} \frac{(\mathbf{1}_{A^c}, T_k \mathbf{1}_A)_{\phi_k}}{\phi_k[A^c]} \\ &\leq \frac{1}{2\pi_{sc}[B^c]} \max_{k \geq k^*} \left\{ \frac{(\mathbf{1}_{A^c}, T_k \mathbf{1}_A)_{\phi_k}}{\phi_k[A^c]} \right\} \\ &\leq \frac{1}{2} \max_{k \geq k^*} \left\{ \frac{1}{\max\{\phi_k[A], \phi_N[A]\}} \frac{(\mathbf{1}_{A^c}, T_k \mathbf{1}_A)_{\phi_k}}{\phi_k[A^c]} \right\} \\ &= \frac{1}{2} \max_{k \geq k^*} \left\{ \gamma(A, k) \Phi_{T_k}(A) \right\}.\end{aligned} \tag{15}$$

First consider $k^* = 0$. In this case $(\mathbf{1}_B, Q\mathbf{1}_{B^c}) = 0$, so combining (13-15) yields Theorem 3.1. Now consider $k^* > 0$. For swapping scheme **SC1**, we have

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

Consider k, l such that $\phi_l[A^c] < \phi_k[A^c]$; then,

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

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

$$\Phi_Q(B) \leq \frac{1}{2}\max_{k\geq k^*, l<k^*}\left[\gamma(A,k)\max\{\delta(A,k,l),\delta(A^c,k,l)\}\right]. \tag{16}$$

Combining (13-16), we get 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 3.1 for the swapping chain that uses scheme **SC1**. With only minor modification, this proof also applies to the swapping scheme **SC2**. □