Torpid Mixing of Simulated Tempering on the Potts Model

Nayantara Bhatnagar* Dana Randall†

Abstract

Simulated tempering and swapping are two families of sampling algorithms in which a parameter representing temperature varies during the simulation. The hope is that this will overcome bottlenecks that cause sampling algorithms to be slow at low temperatures. Madras and Zheng demonstrate that the swapping and tempering algorithms allow efficient sampling from the low-temperature mean-field Ising model, a model of magnetism, and a class of symmetric bimodal distributions [10]. Local Markov chains fail on these distributions due to the existence of bad cuts in the state space.

Bad cuts also arise in the q-state Potts model, another fundamental model for magnetism that generalizes the Ising model. Glauber (local) dynamics and the Swendsen-Wang algorithm have been shown to be prohibitively slow for sampling from the Potts model at some temperatures [1, 2, 6]. It is reasonable to ask whether tempering or swapping can overcome the bottlenecks that cause these algorithms to converge slowly on the Potts model.

We answer this in the negative, and give the first example demonstrating that tempering can mix slowly. We show this for the 3-state ferromagnetic Potts model on the complete graph, known as the mean-field model. The slow convergence is caused by a first-order (discontinuous) phase transition in the underlying system. Using this insight, we define a variant of the swapping algorithm that samples efficiently from a class of bimodal distributions, including the mean-field Potts model.

1 Introduction

The standard approach to sampling via Markov chain Monte Carlo algorithms is to connect the state space of configurations Ω via a graph called the Markov kernel. The *Metropolis algorithm* proscribes transition probabilities to the edges of the kernel so that the chain will converge to any desired distribution [14]. Unfortunately, for some natural choices of the Markov kernel, the Metropolis Markov chain can converge exponentially slowly. Statistical mechanics offers a wealth of sampling problems for which these methods are often applied; it is now well-known that phase transitions in the underlying systems can cause local Markov chains to require exponential time to reach equilibrium [1].

A particular example of this phenomenon is observed on the Potts model. In the q-state Potts model, vertices of an underlying graph are colored with one of q colors. In the ferromagnetic case, vertices connected by an edge in the graph prefer to have the same color. The strength of this preference is a function of the temperature: at high temperature the correlation is negligible, while at low temperatures the effect is strong. At low enough temperatures, local Markov chains that change the color one vertex at a time will be prohibitively slow [1]. This is because to move from a configuration that is predominantly red to one that is predominantly blue, the chain will have to go through highly unlikely configurations where no color dominates.

When the underlying graph in the Potts model is the complete graph, it is known as the *mean-field* or *Curie-Weiss* model. Mean-field models are important because, despite their simplicity, they capture key features present in more complicated graphs. Moreover, for natural problems such as the mean-field Potts and Ising models, there remain obstacles to sampling efficiently, even on the complete graph. Gore and Jerrum showed that the Swendsen-Wang algorithm, a method for sampling that often succeeds in circumventing bottlenecks in the state space, fails on the mean-field Potts model for $q \geq 3$ near the critical temperature (where the phase transition occurs) [6]. Subsequently, Cooper et al. considered the mean-field Ising model ($q = 2$) and showed that Swendsen-Wang is fast everywhere, except possibly near the critical point, where it remains unresolved. [2].

1.1 Tempering, swapping, and annealing.

Simulated annealing provides the insight that varying a parameter representing temperature during a simulation can be a key to designing efficient algorithms [8]. Annealing is intended for optimization problems when direct methods are likely to get trapped in local minima and never find the global optimum. Similarly, simulated tempering and swapping are intended for *sampling* when direct methods are slow [5, 11].

For the tempering and swapping algorithms, we allow a chain to modify the temperature and interpolate between $M + 1$ different distributions $\pi_0, ..., \pi_M$. At the lowest

*College of Computing, Georgia Institute of Technology, Atlanta GA. Email: nand@cc.gatech.edu. Supported in part by NSF CCR-0105639.

†College of Computing and School of Mathematics, Georgia Institute of Technology, Atlanta GA. Email: randall@cc.gatech.edu. Supported in part by NSF CCR-0105639 and an Alfred P. Sloan research fellowship. Part of this work was done while visiting Microsoft Research and the Isaac Newton Institute for Mathematical Sciences in Cambridge, UK.

temperature, π_M is the goal distribution from which we wish to generate samples; at the highest temperature, π_0 is typically less interesting, but the rate of convergence is fast. A Markov chain that keeps modifying the distribution, interpolating between these two extremes, may produce useful samples efficiently. Despite the extensive use of simulated tempering and swapping in practice, there has been little formal analysis. A notable exception is work by Madras and Zheng [10] showing that swapping converges quickly for two simple, symmetric distributions, including the mean-field Ising model.

1.2 Results. In this work, we show that for the mean field Potts model, tempering and swapping require exponential time to converge to equilibrium. The slow convergence of the tempering chain on the Potts model is caused by a first-order (discontinuous) phase transition. In contrast, the Ising model studied by Madras and Zheng has a second-order (continuous) phase transition, which distinguishes why tempering works for one model and not the other.

In addition, we give the first Markov chain algorithm that is provably rapidly mixing on the Potts model. Traditionally, swapping is implemented by defining a set of interpolating distributions where a parameter corresponding to temperature is varied. We make use of the fact that there is greater flexibility in how we define the set of interpolants. Finally, our analysis extends the arguments of Madras and Zheng showing that swapping is fast on symmetric distributions so as to include asymmetric generalizations.

2 Preliminaries

2.1 The q-state Potts model. The Potts model was defined by R.B. Potts in 1952 to study ferromagnetism and anti-ferromagnetism [15]. The interactions between particles are modeled by an underlying graph with edges between particles that influence each other. Each of the n vertices of the underlying graph G is assigned one of q different *spins* (or colors). A *configuration* $\sigma = (q_1, \cdots, q_n)$ is an assignment of spins to the vertices, where q_i denotes the spin at the i^{th} vertex. The energy of a configuration σ is a function of the *Hamiltonian*

$$H(\sigma) = \sum_{(i,j)\in E(G)} J \cdot \delta(q_i, q_j),$$

where δ is the Kronecker-δ function that takes the value 1 if its arguments are equal and zero otherwise. When $J > 0$ the model corresponds to the *ferromagnetic* case where neighbors prefer the same color, while $J < 0$ corresponds to the *anti-ferromagnetic* case where neighbors prefer to be differently colored.

The state space Ω of the q-state ferromagnetic Potts model is the space of all q^n q-colorings of G. We will thus use colorings and configurations interchangeably. Define the inverse temperature $\beta = \frac{1}{kT}$, where k is Boltzmann's constant. The *Gibbs distribution* on configurations at inverse temperature β is given by

$$\pi_\beta(\sigma) = \frac{e^{\beta H(\sigma)}}{Z(\beta)},$$

where $Z(\beta)$ is the normalizing constant. Note that at $\beta = 0$, this is just the uniform distribution on all (not necessarily proper) q-colorings of G.

We consider the ferromagnetic *mean-field model* where G is the complete graph on n vertices and all pairs of particles influence each other. For the 3-state Potts model, $q = 3$. Let σ_1, σ_2, and σ_3 be the number of vertices assigned the first, second, and third colors. Letting $\widetilde{\beta} = \beta J/2$, we can rewrite the Gibbs distribution for the 3-state Potts model as

$$\pi_{\widetilde{\beta}}(\sigma) = \pi_{\widetilde{\beta}}(\sigma_1, \sigma_2, \sigma_3) = \frac{e^{\widetilde{\beta}(\sigma_1^2+\sigma_2^2+\sigma_3^2)}}{Z(\widetilde{\beta})},$$

where the linear terms in the exponent are cancelled by those in the denominator since $\sigma_1 + \sigma_2 + \sigma_3 = n$. We will use this formulation from now on, substituting β for $\widetilde{\beta}$ and denoting $\sigma_1^2 + \sigma_2^2 + \sigma_3^2$ by $H(\sigma)$.

2.2 Markov chains. To sample from a given distribution, a common approach is to design a Markov chain so that an appropriately defined random walk run for a sufficiently long time provides a good sample. We formalize how long "sufficiently long" must be, as well as when a sample is "good" as follows. Let $\mathcal{M}$ be an ergodic (i.e., irreducible and aperiodic), reversible Markov chain with finite state space Ω, transition probability matrix P, and stationary distribution π. Let $P^t(x,y)$ denote the t-step transition probability from x to y.

DEFINITION 2.1. *The* total variation distance *at time t is*

$$\|P^t, \pi\| = \max_{x\in\Omega} \frac{1}{2} \sum_{y\in\Omega} |P^t(x,y) - \pi(y)|.$$

DEFINITION 2.2. *Let $\varepsilon > 0$, then the* mixing time $\tau(\varepsilon)$ *is*

$$\tau(\varepsilon) = \min\{t : \|P^{t'}, \pi\| \le \varepsilon, \forall t' \ge t\}.$$

$\mathcal{M}$ is *rapidly mixing* if the mixing time is bounded above by a polynomial in n and $\log \frac{1}{\varepsilon}$, where n is the size of each configuration in the state space. When the mixing time is exponential in n, we say the chain is *torpidly mixing*.

2.2.1 The Metropolis algorithm. The Metropolis-Hastings algorithm is useful for sampling from non-uniform distributions [14]. Let π be the distribution to be sampled from. A graph G (the Markov kernel) is chosen so as to connect the state space, where vertices are configurations and edges are allowable 1-step transitions. The transition probabilities on G are defined as

$$P(x,y) = \frac{1}{2\Delta} \min\left(1, \frac{\pi(y)}{\pi(x)}\right),$$

for all x, y, neighbors in G, where Δ is the maximum degree of G. It is easy to verify that if the kernel is connected then π is the stationary distribution.

For the Potts model, a natural choice for the Markov kernel is to connect configurations at Hamming distance one. Unfortunately, for large values of β, the Metropolis algorithm converges exponentially slowly on the Potts model for this kernel [1, 2]. This is because the most probable states are largely monochromatic and to go from a predominantly red configuration to a predominantly blue one we would have to pass through states that are highly unlikely at low temperatures.

2.2.2 Simulated tempering. Simulated tempering attempts to overcome this bottleneck by introducing a temperature parameter that is varied during the simulation, effectively modifying the distribution being sampled from. Let $0 = \beta_0 < ... < \beta_M$ be a set of inverse temperatures. The state space of the tempering chain is $\widehat{\Omega} = \Omega \times \{0, \cdots, M\}$, which we can think of as the union of $M+1$ copies of the original state space Ω, each corresponding to a different inverse temperature. Our choice of $\beta_0 = 0$ corresponds to infinite temperature where the Metropolis algorithm converges rapidly to stationarity (on the uniform distribution), and β_M is the inverse temperature at which we wish to sample. We interpolate by setting $\beta_i = \beta_M \cdot \frac{i}{M}$, and let the i^{th} fixed temperature distribution π_i be

$$\pi_i = \pi_{\beta_i}, \ 0 \le i \le M.$$

The stationary distribution of the tempering chain $\widehat{\pi}$, is chosen to be uniform over temperatures, and the conditional distributions are the fixed temperature Gibbs distributions:

$$\widehat{\pi}(x,i) = \frac{1}{M+1}\pi_i(x), \quad x \in \Omega.$$

The tempering Markov chain consists of two types of moves: *level moves*, which update the configuration while keeping the temperature fixed, and *temperature moves,* which update the temperature while remaining at the same configuration.

• A **level move** connects (x,i) and (x',i), where x and x' are connected by one-step transitions of the Metropolis algorithm on Ω at inverse temperature β_i. The move $\widehat{P}((x,i),(x',i))$ is accepted with probability

$$\frac{1}{2(M+1)}P_i(x,x') = \frac{1}{2(M+1)} \min\left(1, \frac{\pi_i(x')}{\pi_i(x)}\right).$$

Here $P_i(x,x')$ is the Metropolis probability of going from x to x' according to the stationary probability π_i.

• A **temperature move** connects (x,i) to $(x, i \pm 1)$ and the move is accepted with probability

$$\begin{aligned}
&\widehat{P}(\,(x,i),(x,i\pm 1)) \\
&\quad = \frac{1}{2(M+1)} \min\left(1, \frac{\widehat{\pi}(x,i\pm 1)}{\widehat{\pi}(x,i)}\right) \\
&\quad = \frac{1}{2(M+1)} \min\left(1, \frac{Z(\beta_i)}{Z(\beta_{i\pm 1})} e^{(\beta_{i\pm 1}-\beta_i)H(x)}\right).
\end{aligned}$$

Notice that while the exponential factor is simple to calculate given σ and i, it is not easy to compute the ratio of partition functions since they are sums over exponentially many configurations at different temperatures. The swapping algorithm, also an aggregate chain using these temperatures, circumvents this difficulty in implementing temperature moves.

2.2.3 Swapping. The swapping algorithm of Geyer [5] is a variant of tempering. The state space is the product space $\widehat{\Omega} = \Omega^{(M+1)}$, the product of $M+1$ copies of the original state space, corresponding to inverse temperatures $\beta_0 < ... < \beta_M$. Let $\pi_M(x) = \pi(x)$ be the distribution from which we wish to sample and let $\pi_0(x) = \frac{1}{|\Omega|}$ (the uniform distribution), for $x \in \Omega$. A configuration in the swapping chain is an $(M+1)$-tuple $x = (x_0, ..., x_M) \in \widehat{\Omega}$, where each component represents a configuration chosen from the i^{th} distribution. The probability distribution $\widehat{\pi}$ is the product measure

$$\widehat{\pi}(x) = \prod_{i=0}^{M} \pi_i(x_i).$$

The swapping chain also consists of two types of moves:

• A **level move** connects $x = (x_0, ..., x_i, ..., x_M)$ and $x' = (x_0, ..., x_i', ..., x_M)$ if x and x' agree in all but the i^{th} components, and x_i and x_i' are connected by one-step transitions of the Metropolis algorithm on Ω. The move $\widehat{P}(x,x')$ is accepted with probability

$$\frac{1}{2(M+1)}P_i(x,x') = \frac{1}{2(M+1)} \min\left(1, \frac{\pi_i(x')}{\pi_i(x)}\right).$$

• A **swap move** connects $x = (x_0, ..., x_i, x_{i+1}, ..., x_M)$ to $x' = (x_0, ..., x_{i+1}, x_i, ..., x_M)$, i.e., it interchanges the i^{th} and $i+1^{st}$ components, with the appropriate Metropolis probabilities on $\widehat{\pi}$. In particular,

$$\begin{aligned}
\widehat{P}(x,x') &= \frac{1}{2(M+1)} \min\left(1, \frac{\widehat{\pi}(x')}{\widehat{\pi}(x)}\right) \\
&= \frac{1}{2(M+1)} \min\left(1, \frac{\pi_{i+1}(x_i)\pi_i(x_{i+1})}{\pi_i(x_i)\pi_{i+1}(x_{i+1})}\right) \\
&= \frac{1}{2(M+1)} \min\left(1, e^{(\beta_{i+1}-\beta_i)(H(x_i)-H(x_{i+1})}\right).
\end{aligned}$$

Notice that now the normalizing constants cancel out. Hence, implementing a move of the swapping chain is straightforward, unlike tempering where good approximations for the partition functions are required. Zheng proved that fast mixing of the swapping chain implies fast mixing of the tempering chain [17], although the converse is unknown.

For both tempering and swapping, we must be careful about how we choose the number of distributions $M+1$. It is important that successive distributions π_i and π_{i+1} have sufficiently small variation distance so that temperature moves are accepted with nontrivial probability. However, M must be small enough so that it does not blow up the running time of the algorithm. Following [10], we set $M = O(n)$. This ensures that for the values of β_M at which we wish to sample, the ratio of π_i and π_{i+1} is bounded from above and below by a constant.

3 Torpid Mixing of Tempering on the Potts model

We will show lower bounds on the mixing time of the tempering chain on the mean-field Potts model by bounding the *spectral gap* of the transition matrix of the chain. Let $\lambda_0, \lambda_1, \ldots, \lambda_{|\Omega|-1}$ be the eigenvalues of the transition matrix P, so that $1 = \lambda_0 > |\lambda_1| \geq |\lambda_i|$ for all $i \geq 2$. Let $Gap(P) = \lambda_0 - |\lambda_1|$.

The mixing time is related to the spectral gap of the chain by the following theorem (see [16]) :

THEOREM 3.1. *Let* $\pi_* = \min_{x\in\Omega} \pi(x)$. *For all* $\varepsilon > 0$,

(a) $\tau(\varepsilon) \leq \frac{1}{Gap(P)} \log(\frac{1}{\pi_*\varepsilon})$.

(b) $\tau(\varepsilon) \geq \frac{|\lambda_1|}{2Gap(P)} \log(\frac{1}{2\varepsilon})$.

The *conductance*, introduced by Jerrum and Sinclair, provides a good measure of the mixing rate of a chain [7]. For $S \subset \Omega$, let

$$\Phi_S = \frac{F_S}{C_S} = \frac{\sum_{x\in S, y\notin S} \pi(x)P(x,y)}{\pi(S)}.$$

Then, the conductance [1] is given by

$$\Phi = \min_{S:\pi(S)\leq 1/2} \Phi_S.$$

It has been shown by Jerrum and Sinclair [7] that, for any reversible chain, the spectral gap satisfies

THEOREM 3.2. *For any Markov chain with conductance* Φ *and eigenvalue gap* $Gap(P)$,

$$\frac{\Phi^2}{2} \leq Gap(P) \leq 2\Phi.$$

[1] It suffices to minimize over $\pi(S) \leq 1/p(n)$, for any polynomial p; this decreases the conductance by at most a polynomial factor (see [16]).

Thus, to lower bound the mixing time it is sufficient to show that the conductance is small.

If a chain converges rapidly to its stationary distribution it must have large conductance, indicating the absence of "bad cut," i.e., a set of edges of small capacity separating $S \subset \Omega$ from $\overline{S} = \Omega \setminus S$. The cut we will use to bound the conductance in the context of the Potts model comes from the *first-order phase transition*. This characterizes the following phenomenon. At high temperature (low β) we are in a *disordered* state and see roughly equal numbers of each color in a typical coloring, while at low temperature (high β) we are in an *ordered* state, where one color clearly dominates. The crucial concept is how we go from the disordered to the ordered state as we slowly lower the temperature. Rather than seeing a gradual change in the size of the largest color class, the change is *discontinuous* and we see an abrupt change around some critical value β_c. To show slow mixing of the tempering chain, we show that this discontinuity translates to a bad cut, even when we take the union of Metropolis chains at many temperatures.

3.1 Slow mixing.

Let $n = |V|$ be the size of the vertex set of the underlying graph being colored. Let $\Omega = 3^n$ be the set of (not necessarily proper) colorings of the graph. Consider a partition of Ω into sets Ω_σ so that $\sigma = (\sigma_1, \sigma_2, \sigma_3)$ and $\sigma_1 + \sigma_2 + \sigma_3 = n$ with $\sigma_1, \sigma_2, \sigma_3 \in \{0, \cdots, n\}$. Since there are exactly $\binom{n}{\sigma_1,\sigma_2,\sigma_3}$ colorings in Ω_σ, we have

$$\pi_i(\Omega_\sigma) = \binom{n}{\sigma_1, \sigma_2, \sigma_3} \frac{e^{\beta_i(\sigma_1^2+\sigma_2^2+\sigma_3^2)}}{Z(\beta_i)}.$$

Let $\Omega_{n/3}$ denote the set of configurations Ω_σ, where $\sigma = (\frac{n}{3}, \frac{n}{3}, \frac{n}{3})$; $\Omega_{2n/3}$, configurations where $\sigma = (\frac{2n}{3}, \frac{n}{6}, \frac{n}{6})$; and $\Omega_{n/2}$, configurations where $\sigma = (\frac{n}{2}, \frac{n}{4}, \frac{n}{4})$. The following lemmas will demonstrate that there is a critical temperature at which $\Omega_{n/3}$ and $\Omega_{2n/3}$ have very large weight although there is a region around $\Omega_{n/2}$ that has very small weight. This will allow us to bound the conductance. (For convenience, we assume throughout that $n = 12k$, for some integer k.)

LEMMA 3.1. *There exists* $0 < \beta_c < \infty$ *such that*

(i) $\pi_{\beta_c}(\Omega_{n/3}) = \pi_{\beta_c}(\Omega_{2n/3}) + o(1)$.

(ii) $\pi_{\beta_c}(\Omega_{n/3})$ *is exponentially larger than* $\pi_{\beta_c}(\Omega_{n/2})$.

Proof. (i) First we determine β_c using Stirling's equation. Let $\pi_{\beta_i}(\Omega_{n/3}) = \pi_{\beta_i}(\Omega_{2n/3})$. Then,

$$\binom{n}{\frac{2n}{3}, \frac{n}{6}, \frac{n}{6}} \frac{e^{\beta_i(\frac{4n^2}{9}+\frac{n^2}{18})}}{Z(\beta_i)} = \binom{n}{\frac{n}{3}, \frac{n}{3}, \frac{n}{3}} \frac{e^{\beta_i(n^2/3)}}{Z(\beta_i)}.$$

This implies

$$\begin{aligned} e^{\beta_i n^2(1/6)} &= \frac{\left(\frac{2n}{3}!\right)\left(\frac{n}{6}!\right)\left(\frac{n}{6}!\right)}{\left(\frac{n}{3}!\right)\left(\frac{n}{3}!\right)\left(\frac{n}{3}!\right)} \\ &= \frac{\left(\frac{2}{3}\right)^{\frac{2n}{3}}\left(\frac{1}{6}\right)^{\frac{n}{3}}}{\left(\frac{1}{3}\right)^n}\left(\frac{1}{\sqrt{2}}\right)\left(1+O(n^{-1})\right) \\ &= \frac{2^{\frac{n}{3}}}{\sqrt{2}}\left(1+O(n^{-1})\right), \end{aligned}$$

which occurs when

$$\beta_i \;=\; \frac{2\ln(2)}{n} + \frac{2}{\sqrt{2}n^2}\ln\left(1+O(n^{-1})\right).$$

Setting β_c to $\frac{2\ln(2)}{n}$ gives the desired result.

(ii) Let $\beta_c = \frac{2\ln(2)}{n}$. Then we have

$$\begin{aligned} \frac{\pi_{\beta_c}(\Omega_{n/2})}{\pi_{\beta_c}(\Omega_{n/3})} &= \frac{\binom{n}{\frac{n}{2},\frac{n}{4},\frac{n}{4}} e^{\beta_c(3n^2/8)}}{\binom{n}{\frac{n}{3},\frac{n}{3},\frac{n}{3}} e^{\beta_c(n^2/3)}} \\ &= \frac{\left(\frac{n}{3}!\right)^3}{\left(\frac{n}{2}!\right)\left(\frac{n}{4}!\right)^2} e^{\beta_c n^2/24} \\ &= \sqrt{\frac{27}{32}}\left(\frac{8}{9}\right)^{\frac{n}{2}} e^{\ln(2)n/12}\left(1+O(n^{-1})\right) \\ &= \sqrt{\frac{27}{32}}\, e^{-\frac{n}{12}\ln\left(\frac{3^{12}}{2^{13}}\right)}\left(1+O(n^{-1})\right). \quad \blacksquare \end{aligned}$$

The first part of the lemma verifies that at the critical temperature there are 3 ordered modes (one for each color, by symmetry) and 1 disordered mode. In the next lemmas, we show that the disordered mode is separated from the ordered modes by a region of exponentially low density. To do this, we use the second part Lemma 3.1 and show that $\pi_i(\Omega_{n/2})$ bounds the density of the separating region at each β_i.

Let $\overline{\pi}_i(x) = \left(\frac{n}{2}, xn, \frac{n}{2} - xn\right)$, for $x \in [0, \frac{1}{2}]$, be the continuous extension of the discrete function $\pi_i\left(\frac{n}{2}, xn, \frac{n}{2} - xn\right)$.

LEMMA 3.2. *For n sufficiently large, the real function $\overline{\pi}_i(x)$ is unimodal for $0 < x < \frac{1}{2}$ and attains its maximum at $x = \frac{1}{4}$ for all i such that $\beta_i \le \beta_c$.*

Proof. Examining $\overline{\pi}_i$ on this line, we find

$$\overline{\pi}_i(x) = \binom{n}{\frac{n}{2}, xn, \frac{n}{2} - xn} \frac{e^{\beta_i n^2\left(\left(\frac{1}{2}\right)^2 + x^2 + \left(\frac{1}{2}-x\right)^2\right)}}{Z(\beta_i)}.$$

Neglecting factors not dependent on x and simplifying using Stirling's formula, we need to check for the stationary points of the function

$$\frac{f(x)}{g(x)} = \frac{e^{\beta_i n(x^2+(\frac{1}{2}-x)^2)}}{(x(\frac{1}{2}-x))^{\frac{1}{2n}} x^x (\frac{1}{2}-x)^{(\frac{1}{2}-x)}}.$$

To test the sign of the derivative $\left(\frac{f(x)}{g(x)}\right)'$, we compare the quantities $\frac{f'}{f}$ and $\frac{g'}{g}$, where $\frac{f'}{f} = \beta_i n(4x-1)$ and $\frac{g'}{g} = \ln(\frac{x}{\frac{1}{2}-x}) + \frac{1}{2n}\frac{1-4x}{1-2x}$. At $x = \frac{1}{4}$ we have $\frac{f'}{f} = 0 = \frac{g'}{g}$, and $g \neq 0$, giving a stationary point. Let $\beta_c = 2\ln(2)/n$, and thus $\beta_i n \le 2\ln(2)$. For $n \ge 100$,

$$\begin{aligned} \beta_c n(4x-1) &> \frac{g'}{g},\ x \in \left(0, \frac{1}{4}\right), \\ \beta_c n(4x-1) &< \frac{g'}{g},\ x \in \left(\frac{1}{4}, \frac{1}{2}\right). \end{aligned}$$

As β_i is decreased, the slope of the line $\frac{f'}{f}$ decreases from the (positive) slope of the line $\beta_c n(4x-1)$. Thus, it is sufficient that the above inequalities hold at β_c to prove the lemma for $\beta_i < \beta_c$ since $\frac{g'}{g}$ is independent of β_i. ■

LEMMA 3.3. *For every inverse temperature $\beta_i \le \beta_c$, $\pi_i(\Omega_{n/2})$ is exponentially smaller than $\pi_i(\Omega_{n/3})$.*

Proof. Note that only the exponential term in $\frac{\pi_i(\Omega_{n/2})}{\pi_i(\Omega_{n/3})}$ varies with β_i. Thus, for some function $h(n)$, we have

$$\begin{aligned} \frac{\pi_i(\Omega_{n/2})}{\pi_i(\Omega_{n/3})} &= h(n)^{\beta_i n^2(H(\frac{1}{2},\frac{1}{4},\frac{1}{4}) - H(\frac{1}{3},\frac{1}{3},\frac{1}{3}))} \\ &= h(n)e^{\beta_i n^2(1/24)} \\ &\le h(n)e^{\beta_c n^2(1/24)} \;=\; \frac{\pi_{\beta_c}(\Omega_{n/2})}{\pi_{\beta_c}(\Omega_{n/3})}. \end{aligned}$$

The claim follows by the second part of Lemma 3.1. ■

We use these facts to bound the conductance of the tempering chain at the critical temperature β_c. Let $A \subset \Omega$ be the region bounded by and including the lines $\sigma_1 = \sigma_2 = \sigma_3 = \frac{n}{2}$. Let $S = \{(x,i) \mid x \in A,\ \beta_0 \le \beta_i \le \beta_c\}$. Let $B = \{x \in S \mid \exists\, x' \in \bar{S},\ p_{xx'} \neq 0\}$ be the boundary of S. The set S defines a bad cut in the state space of the tempering chain.

THEOREM 3.3. *For n sufficiently large, there exists $0 < \alpha < \infty$ such that $\Phi_S \le e^{-\alpha n + o(n)}$.*

Proof. Using the definition of conductance, we have

$$\Phi_S \;=\; \frac{F_S}{C_S} = \frac{\sum_\beta \sum_{x \in B} \pi_\beta(x) \sum_{x' \in \overline{A}} P(x,x')}{\sum_\beta \sum_{x \in A} \pi_\beta(x)}$$

$$
\begin{aligned}
&= \frac{\sum_{\beta} \sum_{x \in B} \pi_\beta(x)}{\sum_{\beta} \sum_{x \in A} \pi_\beta(x)} \\
&\le \frac{\pi_{\beta_0}(\Omega_{n/2}) + \cdots + \pi_{\beta_c}(\Omega_{n/2})}{\pi_{\beta_0}(\Omega_{n/3}) + \cdots + \pi_{\beta_c}(\Omega_{n/3})} O(n) \\
&\le \frac{\pi_{\beta_c}(\Omega_{n/2})}{\pi_{\beta_c}(\Omega_{n/3})} O(n).
\end{aligned}
$$

The first inequality follows from Lemma 3.2, and the second from Lemma 3.3. By Stirling's formula, we find

$$\Phi_S = e^{\frac{\beta_c n^2}{24} + \frac{n}{2} \ln\left(\frac{8}{9}\right) + \ln(O(n))} = e^{-\alpha n + o(n)},$$

where $-\alpha = \frac{\beta_c n}{24} + \frac{1}{2} \ln\left(\frac{8}{9}\right)$ at $\beta_c = \frac{2\ln(2)}{n}$. ■

It can be verified that C_S and $C_{\overline{S}}$ are within a linear factor of each other. By Theorem 3.2 the upper bound on Φ_S bounds the spectral gap of the tempering chain at the inverse temperature β_c. Applying Theorem 3.1, we find the tempering chain for the 3-state Potts model mixes slowly. As a consequence of Zheng's demonstrating that rapid mixing of the swapping chain implies fast mixing of the tempering chain [17], we also have established the slow mixing of the swapping chain for the mean-field Pott model.

4 Modifying the Swapping Algorithm for Rapid Mixing

We now reexamine the swapping chain on two classes of distributions: one is an asymmetric exponential distribution (generalizing a symmetric distribution studied by Madras and Zheng [10]), and the other a class of the mean-field models. First, we show that swapping and tempering are fast on the exponential distribution. The proofs suggest that a key idea behind designing fast sampling algorithms for models with first-order phase transitions is to define a new set of interpolants that do not preserve the bad cut. We start with a careful examination of the exponential distribution since the proofs easily generalize to the new swapping algorithm applied to bimodal mean-field models.

Example I: Let $C > 1$ be a real constant. Let N and N' be positive integers. The bimodal exponential distribution is defined as

$$\pi(x) = \pi_C(x) = \frac{C^{|x|}}{Z}, \qquad x \in [-N, N'],$$

where Z is the normalizing constant. Define the interpolating distributions for the swapping chain as

$$\pi_i(x) = \frac{C^{\frac{i}{M}|x|}}{Z_i}, \; 0 \le i \le M, \; x \in [-N, N']$$

where Z_i is a normalizing constant.

THEOREM 4.1. *The swapping chain with inverse temperatures $\beta_0, \cdots, \beta_M$, where $\beta_i = \beta^* \cdot \frac{i}{M}$ is rapidly mixing on the bimodal exponential distribution defined on $[-N, N']$ where $M = \max(N, N')$.*

We briefly state the comparison and decomposition theorems, which will be the main tools used to prove the results in this section.

Comparison: The comparison theorem of Diaconis and Saloff-Coste is useful in bounding the mixing time of a Markov chain when the mixing time of a related chain on the same state space is known.

Let $\mathcal{M}$ and $\widetilde{\mathcal{M}}$ be two Markov chains on Ω. Let P and π be the transition matrix and stationary distributions of $\mathcal{M}$ and let $\widetilde{P}$ and $\widetilde{\pi}$ be those of $\widetilde{\mathcal{M}}$. Let $E(P) = \{(x,y) : P(x,y) > 0\}$ and $E(\widetilde{P}) = \{(x,y) : \widetilde{P}(x,y) > 0\}$ be sets of directed edges. For $x, y \in \Omega$ such that $\widetilde{P}(x,y) > 0$, define a *path* γ_{xy}, a sequence of states $x = x_0, \cdots, x_k = y$ such that $P(x_i, x_{i+1}) > 0$. Let $\Gamma(z,w) = \left\{(x,y) \in E(\widetilde{P}) : (z,w) \in \gamma_{xy}\right\}$ denote the set of endpoints of paths that use the edge (z,w).

THEOREM 4.2. **(Diaconis and Saloff-Coste [3])**

$$Gap(P) \ge \frac{1}{A} \cdot Gap(\widetilde{P}),$$

where

$$A = \max_{(z,w) \in E(P)} \left\{ \frac{1}{\pi(z)P(z,w)} \sum_{\Gamma(z,w)} |\gamma_{xy}| \widetilde{\pi}(x) \widetilde{P}(x,y) \right\}.$$

Decomposition: Decomposition theorems are useful for breaking a complicated Markov chain into smaller pieces that are easier to analyze [9, 12]. Let $\Omega_1, \cdots, \Omega_m$ be a disjoint partition of Ω. For each $i \in [m]$, define the Markov chain $\mathcal{M}_i$ on Ω_i whose transition matrix $P_i = P[\Omega_i]$, the *restriction* of P to Ω_i is defined as

- $P_i(x,y) = P(x,y)$, if $x \ne y$ and $x, y \in \Omega_i$;
- $P_i(x,x) = 1 - \sum_{y \in \Omega_i, y \ne x} P_i(x,y), \quad \forall x \in \Omega_i$.

The stationary distribution of $\mathcal{M}_i$ is $\pi_i(A) = \frac{\pi(A \cap \Omega_i)}{\pi(\Omega_i)}$. Define the *projection* $\overline{P}$ to be the transition matrix on the state space $[m]$

$$\overline{P}(i,j) = \frac{1}{\pi(\Omega_i)} \sum_{x \in \Omega_i, y \in \Omega_j} \pi(x) P(x,y).$$

THEOREM 4.3. **(Martin and Randall [12])**

$$Gap(P) \ge \frac{1}{2} Gap(\overline{P}) \left(\min_{i \in [m]} Gap(P_i) \right).$$

4.1 Swapping on the exponential distribution.

We are now prepared to prove Theorem 4.1. The state space for the swapping chain applied to Example I is $\widehat{\Omega} = \{-N, N'\}^{M+1}$.

DEFINITION 4.1. *Let $x = (x_0, ..., x_M) \in \widehat{\Omega}$. The* trace *$Tr(x) = t = (t_0, ..., t_M) \in \{0,1\}^{M+1}$ where $t_i = 0$ if $x_i < 0$ and $t_i = 1$ if $x_i \geq 0$, $i = 0, \cdots, M$.*

The 2^{M+1} possible values of the trace characterize the partition we use. Letting $\widehat{\Omega}_t$ be the set of configurations with trace t, we have the decomposition

$$\widehat{\Omega} = \bigcup_{t \in \{0,1\}^{M+1}} \widehat{\Omega}_t.$$

This partition of Ω into sets of fixed trace sets the stage for the decomposition theorem. The restrictions $\widehat{P}_t$ simulate the swapping Markov chain P on regions of fixed trace. The projection $\overline{P}$ is the $M+1$-dimensional hypercube, representing the set of allowable traces t. The analysis of the restrictions follows precisely from [10], the symmetric case. Analyzing the projection, however, becomes more difficult, since in this case the stationary distribution over the hypercube is highly non-uniform. This reflects the fact that at "low temperatures," one side of the bimodal distribution becomes exponentially more favorable. We resolve this by appealing to the comparison theorem.

Bounding the mixing rate of the restricted chains:

If we temporarily ignore swap moves on the restrictions, the restricted chains move independently according to the Metropolis probabilities on each of the $M+1$ distributions. The following lemma reduces the analysis of the restricted chains to analyzing the moves of $\widehat{P}_t$ at each fixed temperature.

LEMMA 4.1. **(Diaconis and Saloff-Coste [3])** *For $i = 1, \ldots, m$, let P_i be a reversible Markov chain on a finite state space Ω_i. Consider the product Markov chain P on the product space $\Omega_0 \times \cdots \times \Omega_M$, defined by*

$$P = \frac{1}{M+1} \sum_{i=0}^{M} \underbrace{I \otimes \cdots \otimes I}_{i} \otimes P_i \otimes \underbrace{I \otimes \cdots \otimes I}_{M-i}.$$

Then $Gap(P) = \frac{1}{M+1} \min_{0 \leq i \leq M} \{Gap(P_i)\}$.

Now $\widehat{\Omega}_t$ restricted to each of the $M+1$ distributions is unimodal, suggesting that $\widehat{P}_t$ should be rapidly mixing at each temperature. Madras and Zheng formalize this in [10] and show that the Metropolis chain restricted to the positive or negative parts of Ω_i mixes quickly. Thus, from Lemma 4.1 and following the arguments in [10], we can conclude that each of the restricted Markov chains $\widehat{P}_t$ is rapidly mixing.

Bounding the mixing rate of the projection:

The graph underlying the Markov chain for the projection $\overline{P}$ is an $M+1$ dimensional hypercube. The stationary probabilities of the projection chain are given by

$$\overline{\pi}(t) = \sum_{x \in \widehat{\Omega} \,:\, Tr(x) = t} \widehat{\pi}(x).$$

Intuitively, we can think about a simpler random walk RW1 on the weighted hypercube as follows. Start at some $M+1$ bit vector, say $0, 0, ..., 0$. At each step we are allowed to transpose two neighboring bits, or we can flip just the lowest bit. Each of these moves is performed with the appropriate Metropolis probability. We will show that this chain is rapidly mixing for the weights that arise in the projection $\overline{P}$. This captures the idea that for the true projection chain, swap moves (transpositions) always have constant probability, and at the highest temperature there is high probability of changing sign. Of course there is a chance of flipping the bit at each higher temperature, but we will see that this is not even necessary for rapid mixing.

To analyze RW1, we can compare it to an even simpler walk, RW2, that chooses *any* bit at random and updates it to 0 or 1 with the correct stationary probabilities. It is easy to argue that RW2 converges very quickly and we use this to infer the fast mixing of RW1.

More precisely, let $\widetilde{P}$ be a new chain on the hypercube for the purpose of the comparison. At each step it picks $i \in_u \{0, ..., M\}$ and updates the i^{th} component t_i by choosing t'_i exactly according to the appropriate stationary distribution at β_i. In other words, the i^{th} component is at stationarity as soon as it is chosen. Using the coupon collector's theorem, we have

LEMMA 4.2. *The chain $\widetilde{P}$ on $\{0,1\}^{M+1}$ mixes in time $O\left(M \log(M + \varepsilon^{-1})\right)$ and $Gap(\widetilde{P})^{-1} = O(M \log M)$.*

We are now in a position to prove the following theorem.

THEOREM 4.4. *The projection $\overline{P}$ of the swapping Markov chain is rapidly mixing on $\{0,1\}^{M+1}$*

To apply the comparison theorem, we translate transitions in the chain $\widetilde{P}$, (whose mixing time we know) into a canonical path consisting of moves in the chain $\overline{P}$. Let (t, t') be a single transition in $\widetilde{P}$ from $t = (t_0, ..., t_i, ..., t_M)$ to $t' = (t_0, ..., 1 - t_i, ..., t_M)$ that flips the i^{th} bit.

The canonical path from t to t' is the concatenation of three paths $p_1 \circ p_2 \circ p_3$. In terms of tempering, p_1 is a heating phase and p_3 is a cooling phase.

- p_1 consists of i swap moves from t to $(t_i, t_0, ..., t_{i-1}, t_{i+1}, ..., t_M)$;
- p_2 consists of one step that flips the bit corresponding to the highest temperature to move to $(1 - t_i, t_0, ..., t_M)$;

- p_3 consists of i swaps until we reach $t' = (t_0, ..., 1 - t_i, ..., t_M)$.

To bound A in Theorem 4.2, we will establish that

$$\overline{\pi}(z)\,\overline{P}(z,z') \geq \overline{\pi}(t)\,\widetilde{P}(t,t'), \tag{4.1}$$

for any transition (z, z') in the canonical path. Second, we need to ensure that the number of paths using the transition (z, z'), $\Gamma_{z,z'}$, is at most a polynomial. These two conditions are sufficient to give a polynomial bound on the parameter A in the comparison theorem. For any (z, z') we have $|\Gamma(z, z')| \leq M^2$, so it remains to establish the condition in Equation 4.1.

Case 1: (Transitions along p_1)
Let $z = (t_0, ..., t_{j-1}, t_i, t_j, ..., t_{i-1}, t_{i+1}, ..., t_M)$ and $z' = (t_0, ..., t_i, t_{j-1}, ..., t_{i-1}, t_{i+1}, ..., t_M)$.

$$\begin{aligned} \overline{\pi}(z)\overline{P}(z,z') &= \frac{\overline{\pi}(z)}{2(M+1)} \min\left(1, \frac{\overline{\pi}(z')}{\overline{\pi}(z)}\right) \\ &= \frac{1}{2(M+1)} \min\left(\overline{\pi}(z), \overline{\pi}(z')\right). \end{aligned} \tag{4.2}$$

First we consider $\overline{\pi}(z)$.

$$\overline{\pi}(z) = \prod_{\ell=0}^{M} \sum_{Tr(x)_\ell = z_\ell} \pi_\ell(x) \triangleq \prod_{\ell=0}^{M} \pi_\ell(z_\ell).$$

Let us assume, without loss of generality, that $N \leq N'$. Then we have

$$\begin{aligned} \overline{\pi}(t)\widetilde{P}(t,t') &= \frac{\overline{\pi}(t)}{2(M+1)} \min\left(1, \frac{\overline{\pi}(t')}{\overline{\pi}(t)}\right) \\ &= \frac{1}{2(M+1)} \min\left(\overline{\pi}(t), \overline{\pi}(t')\right) \\ &= \frac{\overline{\pi}(t^*)}{2(M+1)}, \end{aligned}$$

where $t^* = (t_0, ..., t_{i-1}, 0, t_{i+1}, ..., t_M)$. We want to show that $\overline{\pi}(t^*) \leq \overline{\pi}(z)$. It is useful to partition t^* into blocks of bits t_ℓ that equal 1, separated by one or more zeros. Let $k < i$ be the largest value such that $t_k = 0$. Then it is easy to verify that

$$\prod_{\ell=k+1}^{i} \pi_\ell(z_\ell) \geq \prod_{\ell=k+1}^{i} \pi_\ell(t^*_\ell).$$

Similarly, considering the next block of t^* (i.e., the next set of bits such that $t_\ell = 1$) until the first index k' such that $t_{k'} = 0$,

$$\prod_{\ell=k'+1}^{k} \pi_\ell(z_\ell) \geq \prod_{\ell=k'+1}^{k} \pi_\ell(t^*_\ell).$$

Continuing in this way we find

$$\prod_{\ell=j}^{i} \pi_\ell(z_\ell) \geq \prod_{\ell=j}^{i} \pi_\ell(t^*_\ell),$$

and thus

$$\overline{\pi}(z) \geq \overline{\pi}(t^*).$$

Likewise, by taking one more term, we find that $\overline{\pi}(z') \geq \overline{\pi}(t^*)$. Together with equation 4.2 this implies

$$\overline{\pi}(z)\,\overline{P}(z,z') \geq \overline{\pi}(t)\,\widetilde{P}(t,t').$$

Case 2: (The transition along p_2) Consider the transition from $z = (t_i, t_0, ..., t_{i-1}, t_{i+1}, ..., t_M)$ to $z' = (1 - t_i, t_0, ..., t_M)$ that flips the first bit of z. Repeating the argument from Case 1, it follows that

$$\min\left(\overline{\pi}(z), \overline{\pi}(z')\right) \geq \overline{\pi}(t^*).$$

Therefore, again we find equation 4.1 is satisfied.

Case 3: (Transitions along p_3) This is similar to Case 1.

In all three cases, we find that if (z, z') is one step on the canonical path from t to t', equation 4.1 is satisfied. Therefore, it follows that

$$A = \max_{(z,z') \in E(\overline{P})} \left\{ \frac{\sum_{\Gamma(z,z')} |\gamma_{t,t'}| \overline{\pi}(t) \widetilde{P}(t,t')}{\overline{\pi}(z)\overline{P}(z,z')} \right\} \leq M^2.$$

By the comparison theorem we find that $Gap(\overline{P}) \geq M^{-1}$. We have now established all the results necessary to apply the decomposition theorem 4.3 and show Theorem 4.1.

4.2 Bimodal mean-field spin models.

We now look more closely at mean-field models to see how to modify the swapping algorithm. Consider the following very general class of mean-field models.

Example II: Bimodal mean-field spin models: Fix constants $\beta > 0$, $A_1, A_2, ..., A_k$, and let n be a large integer. The state space of the mean-field model consists of all spin configurations on the complete graph K_n, namely $\Omega = \{1, ..., q\}^n$. The probability distribution over these configurations is determined by β, inverse temperature, and $\{A_k\}$, the k-wise interactions between particles. The Hamiltonian is given by

$$H(x) = \sum_k \sum_{\{i_1,...,i_k\} \subset [n]} A_k\, \delta_{(x_{i_1},...,x_{i_k})},$$

where δ is the Kronecker-δ function that takes the value 1 if all of the arguments are equal and is 0 otherwise (when

$k = 1$ we set δ_{x_i}=1 iff $x_i = 1$). The Gibbs distribution is

$$\pi(x) = \pi_{(\beta, A_1, \dots, A_k)}(x) = \frac{e^{\beta H(x)}}{Z}, \text{ for } x \in \Omega,$$

where $Z = \sum_{y \in \Omega} e^{\beta H(y)}$ is the normalizing constant.

For any partition of n, $\sigma = (\sigma_1, ..., \sigma_q)$, $\sum_q \sigma_q = n$, define Ω_σ as the set of configurations with σ_i vertices assigned color i. Let us consider the *total spins distribution*:

$$S_\sigma = \pi(\Omega_\sigma) = \sum_{x \in \Omega_\sigma} \pi(x).$$

We consider here the cases when S_σ is a bimodal function in $\mathbb{Z}^{q-1}$ (i.e., when there are exactly two local optima).

An important special case of Example II is the mean-field Ising model in the presence of an external field. This model is defined by parameters $q = 2$, $\beta > 0$, the inverse temperature, and $J > 0$, the external magnetic field. The Gibbs distribution over configurations $x \in \Omega$ is

$$\pi(x) = \pi_{(\beta, J)}(x) = \frac{e^{\beta\left(\sum_{i,j} \delta_{x_i = x_j} + J \sum_i \delta_{x_i = 1}\right)}}{Z(\beta, J)},$$

where $Z(\beta, J)$ is the normalizing constant. This can be described by the model in Example II by taking $k = 2, A_1 = J$ and $A_2 = 2$. It can be shown that this distribution is bimodal for all values of β and J.

A second important special case included in Example II is the q-state Potts model where we restrict to the part of the state space $\Omega_{\text{ord}} \subset \Omega$ such that $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_q$. Note that $\pi(\Omega_{\text{ord}}) = \pi(\Omega)/n!$. Consequently, sampling from Ω_{ord} is sufficient since we can randomly permute the colors once we obtain a sample and get a random configuration of the Potts model on the nonrestricted state space Ω. Here we take $k = 2$, $A_1 = 0$ and $A_2 = J$, and the Gibbs distribution becomes

$$\pi(x) = \pi_{(\beta, J)}(x) = \frac{e^{\beta J \sum_{i,j} \delta_{(x_i, x_j)}}}{Z(\beta, J)}.$$

Restricting to Ω_{ord} provides a bimodal distribution, which is required for the arguments that follow.

Our results from the previous section indicate that swapping is not always fast on models defined in Example II; it is easy to see that the arguments directly apply to Potts model restricted to Ω_{ord}. In contrast, the new swapping algorithm we define next is can be shown to be rapidly mixing for the entire class of models defined in Example II.

4.3 A new swapping algorithm. In the traditional swapping algorithm, the interpolating distributions are defined as

$$\pi_i(x) = \pi_{(\beta_i, A_1, \dots, A_k)}(x) = \frac{e^{\beta_i H(x)}}{Z_i}, \ 0 \leq i \leq M,$$

where $\beta_i = \beta_M \frac{i}{M}$ and Z_i normalizes the distribution. Our new algorithm stems from the observation that this is a poor choice of interpolants because they preserve the first-order phase transition. We can do much better by exploring a wider class of interpolating distributions.

To see the flexibility we have in defining the set of distributions, define

$$\rho_i(x) = \frac{\pi_i(x) f_i(x)}{Z_i'},$$

where $Z_i' = \sum_{x \in \Omega} \pi_i(x) f_i(x)$ is another normalizing constant. When $f_i(x)$ is taken to be the constant function, then we obtain the distributions of the usual swapping algorithm.

The Flat-Swap Algorithm:

For our variant, the *Flat-Swap* algorithm, let us consider

$$f_i(x) = \binom{n}{\sigma_1, \dots, \sigma_q}^{\frac{i-M}{M}}.$$

We shall see that this gradually flattens out the total spins distributions uniformly, thus eliminating the bad cut that can occur when we take $f_i(x)$ constant. The function $f_i(x)$ effectively dampens the entropy (multinomial) just as the change in temperature dampens the energy term coming from the Hamiltonian. We have the following theorem.

THEOREM 4.5. *The Flat-Swap algorithm is rapidly mixing for any bimodal mean-field model.*

To prove Theorem 4.5, we follow the strategy set forth for Theorem 4.1, using decomposition and comparison in a similar manner. For simplicity, we concentrate our exposition here on the Ising model in an external field. The advantage of this special case is that the total spins configurations form a one-parameter family (i.e., the number of vertices assigned +1), much like in Example I. The proofs for the general class of models, including the Potts model on Ω_{ord}, are analogous. We sketch the proof of Theorem 4.5.

For the Ising model, we have

$$f_i(x) = \binom{n}{k}^{\frac{i-M}{M}},$$

where k vertices are assigned $+1$ and $n - k$ are assigned -1. Again we take $\beta_i = \beta^* \cdot \frac{i}{M}$. Note that $f_i(x)$ is easy to compute given x. A simple calculation reveals that, for $x \in \Omega_{(k, n-k)}$,

$$\rho_i(\Omega_{(k,n-k)}) = \binom{n}{k} \rho_i(x) = \frac{1}{Z_i'} \left(\rho_M(\Omega_{(k,n-k)})\right)^{\frac{i}{M}}.$$

Thus, all the total spins distributions have the same relative shape, but get flatter as i is decreased. This no longer preserves the non-analytic nature of the phase transition seen for the usual swap algorithm. It is this property that makes this choice of distributions useful.

The total spins distribution for the Ising model is known to be bimodal, even in the presence of an external field. With our choice of interpolants, it now follows that all $M + 1$ distributions are bimodal as well. Moreover, the minima of the distributions occur at the same location for all $M + 1$ distributions. Let $t_{\min}$ be the place at which these minima occur.

In order to show that this swapping chain is rapidly mixing we use decomposition. Let $\widehat{\Omega} = \Omega^{M+1}$ be the state space of the swapping chain on the Ising model, where $\Omega = \{+1, -1\}^n$. Define the trace $\mathrm{Tr}(x) = t \in \{0, 1\}^{M+1}$, where $t_i = 0$ if the number of $+1$s in x_i is less than $t_{\min}$ and let $t_i = 1$ if the number of $+1$s in x_i is at least $t_{\min}$.

The analysis of the restricted chains given in [10] in the context of the Ising model without an external field can be readily adapted to show the restrictions $\widehat{\Omega}_t$ are also rapidly mixing. The analysis of the projection is analogous to the arguments used to bound the mixing rate of the projection for Example I. Hence, we can conclude that the swapping algorithm is rapidly mixing for the mean-field Ising model at any temperature, with any external field. We leave the details, including the extension to the Potts model, for the full version of the paper.

5 Conclusions

Swapping, tempering and annealing provide a means, experimentally, for overcoming bottlenecks controlling the slow convergence of Markov chains. However, our results offer rigorous evidence that heuristics based on these methods might be incorrect if samples are taken after only a polynomial number of steps. In recent work, we have extended the arguments presented here to show an even more surprising result; tempering can actually be slower than the fixed temperature Metropolis algorithm by an exponential multiplicative factor.

Many other future directions present themselves. It would be worthwhile to continue understanding examples when the standard (temperature based) interpolants fail to lead to efficient algorithms, but nonetheless variants of the swapping algorithm, such as presented in Section 4.3, succeed. The difficulty in extending our methods to more interesting examples, such as the Ising and Potts models on lattices, is that it is not clear how to define the interpolants. We would want a way to slowly modify the the entropy term in addition to the temperature, as we did in the mean-field case, to avoid the bad cut arising from the phase transition. It would be worthwhile to explore whether it is possible to determine a good set of interpolants algorithmically by bootstrapping, rather than analytically, as was done here, to define a more robust family of tempering-like algorithms.

Acknowledgments

The authors thank Christian Borgs, Jennifer Chayes, Claire Kenyon, and Elchanan Mossel for useful discussions.

References

[1] C. Borgs, J.T. Chayes, A. Frieze, J.H. Kim, P. Tetali, E. Vigoda, and V.H. Vu. Torpid mixing of some MCMC algorithms in statistical physics. *Proc. 40th IEEE Symposium on Foundations of Computer Science*, 218–229, 1999.

[2] C. Cooper, M.E. Dyer, A.M. Frieze, and R. Rue. Mixing Properties of the Swendsen-Wang Process on the Complete Graph and Narrow Grids. *J. Math. Phys.* **41**: 1499–1527: 2000.

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

[4] C.J. Geyer. Markov Chain Monte Carlo Maximum Likelihood. *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface* (E.M. Keramidas, ed.), 156-163. Interface Foundation, Fairfax Sta tion, 1991.

[5] C.J. Geyer and E.A. Thompson. Annealing Markov Chain Monte Carlo with Applications to Ancestral Inference. *J. Amer. Statist. Assoc.* **90**: 909–920, 1995.

[6] V.K. Gore and M.R. Jerrum. The Swendsen-Wang Process Does Not Always Mix Rapidly. *J. Statist. Phys.* **97**: 67–86, 1995.

[7] M.R. Jerrum and A.J. Sinclair. Approximate counting, uniform generation and rapidly mixing Markov chains. *Information and Computation*. **82**: 93–133, 1989.

[8] S. Kirkpatrick, L. Gellatt Jr., and M. Vecchi. Optimization by simulated annealing. *Science*. **220**: 498–516, 1983.

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

[10] N. Madras and Z. Zheng. On the swapping algorithm. *Random Structures and Algorithms*. **22**: 66–97, 2003.

[11] E. Marinari and G. Parisi. Simulated tempering: a new Monte Carlo scheme. *Europhys. Lett.* **19**: 451–458, 1992.

[12] R.A. Martin and D. Randall. Sampling adsorbing staircase walks using a new Markov chain decomposition method. *Proc. 41st Symposium on the Foundations of Computer Science (FOCS 2000)*, 492–502, 2000.

[13] R.A. Martin and D. Randall. Disjoint decomposition with applications to sampling circuits in some Cayley graphs. Preprint, 2003.

[14] N. Metropolis, A. W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**: 1087–1092, 1953.

[15] R.B. Potts. Some Generalized Order-disorder Transformations *Proceedings of the Cambridge Philosophical Society*, **48**: 106–109, 1952.

[16] A.J. Sinclair. *Algorithms for random generation & counting: a Markov chain approach.* Birkhäuser, 1993.

[17] Z. Zheng. Analysis of Swapping and Tempering Monte Carlo Algorithms. Dissertation, York Univ., 1999.