Mirko Ebbers

Swapping, Tempering and Equi-Energy sampling on a selection of models in statistical mechanics

2010

Mathematik

Swapping, Tempering and Equi-Energy sampling on a selection of models in statistical mechanics

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften im Fachbereich
Mathematik und Informatik
der Mathematisch-Naturwissenschaftlichen Fakultät
der Westfälischen Wilhelms-Universität Münster

vorgelegt von
Mirko Ebbers
aus Düsseldorf

2010

Dekan	Prof. Dr. Matthias Löwe
Erstgutachter	Prof. Dr. Matthias Löwe
Zweitgutachter	Prof. Dr. Gerold Alsmeyer
Tag der mündlichen Prüfung	27.01.2011
Tag der Promotion	27.01.2011

Contents

Introduction

With Newton's Laws of Motion, the world seemed an understandable and predictable place to be in. It is easily explained why an apple falls downwards, eventually hitting the ground when being separated from the tree's limb. Why horses have a hard time pulling a carriage up to a castle's hilltop while they do not seem to mind pulling the same carriage on a path alongside a river seems an easy question to answer. Knowing these answers, it seems obvious to say why a steam-driven locomotive takes so much more coal and water when riding uphill, compared to riding on a flat plane. But then again, going a fixed distance with the train has the crankshaft rotate a defined amount of times. This does not depend on the steepness of the tracks. The expansion chamber has a fixed size. So why should the water and coal consumption increase, just because the train is tilted slightly?

Thinking about this more thoroughly it seems that Newton could give an answer to this too. Newton's third law, the Action-Reaction Law, says it needs more force to rotate the crankshaft if the train increases its altitude. A combination of the laws now demands that there must either be more molecules in the expansion chamber that hit the piston in a given time interval, the molecules need to have a higher momentum, thereby increasing the force acting on the piston on impact, or a combination of both. In order to answer the question of water and coal usage all we have to do is to solve the dynamics of this simple looking model. We have to take the molecules and start solving the equations given by Newton.

As it turns out, this leads to a vast system of $\sim 10^{23}$ coupled differential equations, which seems impossible to explicitly solve with the worlds current processing power. To address this problem, we come to the idea of classical thermodynamics. This theory states that the observable state of a system does not depend on the mechanics of a single molecule, but on quantities of vastly reduced information: temperature, pressure and volume of the system. This theory coupled with the knowledge of the classical mechanics introduced by Newton gives us just what we experience in real life. Riding uphill takes more pressure on the piston, which takes a higher pressure in the expansion chamber, which means, the steam needs to be *thicker* and thereby implicitly hotter. So more water needs to be brought to a higher temperature

in the pressure chamber, which takes more energy. This excess energy can only be brought in by using more coal.

Even though the classical theory of thermodynamics is very successful in explaining real world behavior, it lacks a foundation in a more global theory like for instance the theory of mechanics. A connection between mechanics and thermodynamics has been introduced through the works of Maxwell and Boltzmann by inventing the theory of statistical mechanics. The fundamental difference of this approach to the approach stated earlier involving solving these coupled differential equations is that we do not need to know the exact behavior of every single molecule, but it suffices to know the mean behavior of the molecules under consideration in order to understand the macroscopic properties of the system. The huge amount of molecules guarantees that a small atypically behaving mass of molecules cannot change the observable behavior. The real world behavior is solemnly determined by the broad masses. This argument is founded on the *equal a priori probability* postulate, which is the fundamental assumption of statistical mechanics.

As the name suggests, statistical mechanics involves some ideas of probability theory. A fundamental concept being used is the idea of a probability measure on all possible microscopic states of a thermodynamical system. A realistic system is – in theory – drawn out of all possible systems with regard to this probability measure. Now, even though it is easy to write down this measure, it turns out to be difficult to actually calculate probabilities with this definition.

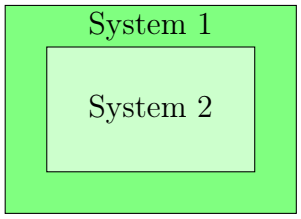
As these probability measures cannot be rigorously calculated in acceptable time, one has to settle for approximations of these quantities. There are several ways to gain such approximations, some of them involving methods of numerical mathematics, some involving the use of stochastic techniques. This thesis deals with two closely related methods, called *Swapping* and *Simulated Tempering* and with a derivative thereof called Equi-Energy sampling.

We will first introduce statistical mechanics and give definitions for the models which are of interest for this thesis. In Chapter 2 the *Markov-Chain-Monte-Carlo* method for sampling from certain distributions is being introduced and afterwards a short example is given, showing that this technique does not always yield favorable results. In Chapter 3 the previously named methods of Swapping and Simulated Tempering are defined and its usefulness is being justified by an overview of what is known in literature so far. Chapters 4 through 6 deal with the behavior of these Markov chains on some of the models introduced in Chapter 1. Chapter 7 introduces the Equi-Energy sampler and gives a lower bound for the speed of convergence of this algorithm for the Potts model.

CHAPTER 1

Statistical Mechanics

The statistical mechanics in physics is an approximation of the real world through probabilistic methods. The idea is to look at a real world piece of matter (1 liter of air, a 1 kg iron bar, ...) as a set of N individual pieces (atoms, molecules,...) and to realize that it suffices to know the probabilities of certain states one single atom can be in to understand the macroscopic behavior of the model. This is granted by the usually high number of atoms to be considered in a real world probe, which makes solving the classical mechanics or quantum mechanics coupled differential equations practically impossible in the first place. Consider a system which is in the physical state x. The Hamilton function $H(x)$ has proven to be the correct function to look at in classical mechanics in order to describe the evolution of the model over time. By reading $H(x)$ one can think of a certain energy-level the system is in while being in state x. Understanding this much, we can think of two disjoint systems which touch each other on the bordering line. Looking at Figure 1, each system's dynamics is given



$$H_1 = H_1(q_1^{(1)}, ..., p_{f_1}^{(1)})$$
$$H_2 = H_2(q_1^{(2)}, ..., p_{f_2}^{(2)})$$

FIGURE 1. Two disjoint but touching systems. The Hamilton functions of each system are given by H_1 and H_2.

by independent Hamiltonians. Now, looking at the dynamics of both systems considered as an union, it is physically reasonable to write the Hamiltonian as

$$(1) \quad H = H_1\left(q_1^{(1)}, ..., p_{f_1}^{(1)}\right) + H_2\left(q_1^{(2)}, ..., p_{f_2}^{(2)}\right) + H_{1,2}\left(q_1^{(1)}, ..., p_{f_1}^{(1)}, q_1^{(2)}, ..., p_{f_2}^{(2)}\right)$$

with $H_{1,2}$ describing the interaction between the two systems. Looking at the picture, the interaction area is small compared to the mass of

the system, which physically speaking leads to very small interaction energies $H_{1,2} \ll H_1 + H_2$ and therefore to

$$H = H_1 + H_2.$$

Now consider Figure 1 from the probabilist's point of view. The probability of finding System 1 in a state x should only depend on x through the Hamilton function, thus it should read $\pi^{H_1}(x)$. Due to the small interaction area, we can again argue that the two systems are evolving, up to a small degree of inaccuracy, interdependent of each other, thus yielding

$$\pi^H = \pi^{H_1}\pi^{H_2}.$$

Combining these two points, we gain

$$\pi^{H_1+H_2} = \pi^H = \pi^{H_1}\pi^{H_2}$$

which gives us the Boltzmann distribution

$$\pi = \frac{e^{\beta H}}{Z(\beta)} \tag{2}$$

as the probability measure to consider. The parameter $\beta = \frac{1}{T}$ is called the inverse temperature and describes how agitated the system is. Low temperature T, thus high β, is a cold system with little thermal energy, while high temperature, thus small β, describes a system with high thermal energy. The constant $Z(\beta)$ is the normalization factor, which makes π be a probability measure. For this, the assumption is that

$$Z(\beta) := \int e^{\beta H}\, d\mu < \infty$$

for a reference measure μ. This will most usually be the counting measure or the Lebesgue measure.

Most realistic models are, mathematically speaking, difficult to study therefore, this thesis will deal with the following, compared to real world physical models, simplified models.

1. The Curie-Weiss model

This is one of the simplest models of statistical mechanics which still yields some interesting behavior for varying β. Define the set $\Lambda = \{1, ..., N\}$ for some $N \in \mathbb{N}$ and think of it as a set of N atoms. Each such atom x_i can either have spin up ($\sigma_i = 1$) or spin down ($\sigma_i = -1$). Thus the state space we consider is

$$\Omega = \{-1, 1\}^\Lambda. \tag{3}$$

To keep every calculation as simple as possible the Curie-Weiss model is the mean field model derived from the Ising model (see **[3]**). "Mean

field" means that every atom interacts in the same way and in the same strength with any other atom. The Hamiltonian is therefore given by

$$H(\sigma) = \frac{1}{2N}\left(\sum_{i=1}^{N}\sigma_i\right)^2 = \frac{1}{2N}\sum_{i,j=1}^{N}\sigma_i\sigma_j \tag{4}$$

This model is of great theoretical interest as, despite being mathematically simple, the model undergoes a phase transition as β passes through $\beta_c = 1$. Above the critical temperature the system has just one macrostate which has zero total magnetization

$$m(\sigma) := \sum_{i=1}^{N}\sigma_i. \tag{5}$$

Below the critical temperature the system has two distinct macrostates, one of which has positive total magnetization while the other has negative total magnetization. See **[3]** for an extensive analysis of the model.

1.1. The Generalized-Curie-Weiss model. The Generalized-Curie-Weiss model is an extension of the classical Curie-Weiss model considered by Eisele and Ellis **[13]**. Looking at (4) we can write

$$H(\sigma) = \frac{1}{2N}\left(\sum_{i=1}^{N}\sigma_i\right)^2 = \frac{1}{2N}\big(m(\sigma)\big)^2$$

thus in the Curie-Weiss model the energy of the current state is only dependent on the reduced information of the total magnetization of the current state, given by $m(\sigma)$. Eisele and Ellis suggest to look at the model induced through the Hamiltonian

$$H(\sigma) := N\ g\left(\frac{m(\sigma)}{N}\right) \tag{6}$$

on the state space $\Omega = A^\Lambda \subseteq [-L, L]^\Lambda$ and for an arbitrary function g on $[-L, L]$. In order to give a reasonable physical model, the following restrictions are applied to this construction:

(1) g is an even, real analytic function on $\mathbb{R}$ and is strictly increasing on $[0, L]$ with $g(0) = 0$.
(2) ρ is a symmetric non-degenerate (i.e., $\rho \neq \delta_0$) Borel measure on $\mathbb{R}$.
Note: We will henceforth take ρ to be the normalized counting measure on a finite symmetric subset $A \subseteq [-L, L]$ in order to make Ω technically a finite state space. The reason we only allow this single measure is that only in this case has equation (7) the desired Boltzmann-shape of having the probability of a state only determined by the product of the Hamiltonian H and β, as only in this case the factor

$$\exp\big(\sum \log(\rho(\sigma_i))\big)$$

cancels out.

(3) There exists a symmetric, non-constant and convex function h on $[-L, L]$ such that

$$g(x) \leq h(x) \quad \text{for } x \in [-L, L]$$

and

$$\int_{[-L,L]} e^{\beta h(x)} \rho(dx) < \infty \quad \text{for all } \beta > 0.$$

Define the sequence of probability measures for the Generalized-Curie-Weiss model by

$$P_{N,\beta}(\sigma) = \frac{e^{N\beta g\left(\frac{m(\sigma)}{N}\right)} \prod_{i=1}^{N} \rho(\sigma_i)}{Z_N(\beta)} = \frac{e^{N\beta g\left(\frac{m(\sigma)}{N}\right) + \sum_{i=1}^{N} \log(\rho(\sigma_i))}}{Z_N(\beta)}. \tag{7}$$

Eisele and Ellis give a detailed analysis of the phase behavior encountered in the model, depending on the choice of g, in **[13]**. The last restriction guarantees a finite normalization constant of the model for arbitrary ρ. In our setting of a finite set A this is not needed, thus restriction (3) is only given for a thorough definition of the Generalized-Curie-Weiss model.

2. The Potts model

The model we will consider in Chapter 7 is the mean field Potts model. Just as the Generalized-Curie-Weiss model, the Potts model is a generalization of the Curie-Weiss model. This time the difference does not lie in the Hamilton function but in the state space. While the Curie-Weiss model and the Generalized-Curie-Weiss model both have two types of *spin* values, positive and negative, the Potts model generalizes this to an arbitrary number q of *colors*. So here $\Omega = E^N$ where $E = \{1, \ldots, q\}$ with $q \geq 3$. The energy function H of the mean field Potts model is then defined as

$$H_N(\sigma) = \frac{1}{2N} \sum_{i,j=1}^{N} \delta_{\sigma_i = \sigma_j}, \qquad \sigma \in \Omega \tag{8}$$

and the Gibbs measures are defined accordingly. Obviously, H_N can be written as a function of the vector

$$\mathfrak{m}_N(\sigma) = \left(\frac{1}{N} \sum_{i=1}^{N} \delta_{\sigma_i = 1}, \frac{1}{N} \sum_{i=1}^{N} \delta_{\sigma_i = 2}, \ldots, \frac{1}{N} \sum_{i=1}^{N} \delta_{\sigma_i = q} \right),$$

the order parameter of the Potts model.

As the Curie-Weiss model the mean field Potts model undergoes a phase transition. There is a critical temperature β_c, such that in the high temperature phase $\beta < \beta_c$ the distribution of $\mathfrak{m}_N$ converges to the

Dirac measure in $(\frac{1}{q},\dots,\frac{1}{q})$. For $\beta \geq \beta_c$ there is $1 > m^*(\beta) > \frac{1}{q}$ such that at β_c the distribution of $\mathfrak{m}_N$ converges to

$$\frac{1}{q+1}\left(\delta_{(\frac{1}{q},\dots,\frac{1}{q})} + \sum_{i=1}^{q} \delta_{v_i(\beta_c)}\right)$$

where $v_i(\beta)$ is the vector that has $m^*(\beta)$ in its i'th component and all other components equal such that they sum up to one. For $\beta > \beta_c$ the distribution of $\mathfrak{m}_N$ converges to

$$\frac{1}{q}\sum_{i=1}^{q} \delta_{v_i(\beta)}.$$

This phase transition is of first order, since $m^*(\beta_c) > 1/q$, i.e. the jump is discontinuous. Moreover the vector $(\frac{1}{q},\dots,\frac{1}{q})$ remains to be a local maximum of the distribution of $\mathfrak{m}_N$ for all temperatures. This fact will be of utmost importance for our calculations. All these results can be found in the article by Ellis and Wang [**16**].

3. The BEG model

The mean field Blume-Emery-Griffiths (BEG) model is also closely related to the Curie-Weiss model. For a given $K > 0$ the Hamilton function on $\Omega = \{-1, 0, 1\}^N$ is given by

$$H(\sigma) = H_K(\sigma) := -\sum_{j=1}^{N} \sigma_j^2 + \frac{K}{N}\left(\sum_{j=1}^{N} \sigma_j\right)^2 \tag{9}$$

for $\sigma \in \Omega$. A third spin direction is added in this model, which has a neutral role to the spins $+1$ and -1. The first summand puts weight into those configurations in which many spins have value 0, while the second summand puts emphasis on those states which have a non-zero total magnetization. Here, the parameter K gives a reference of how close this model should be to a Curie-Weiss model. Large K will give the second summand large weight while the first summand only plays a minor role. Choosing K small will give a totally different model, namely one whose macrostate has zero total magnetization for any $\beta \geq 0$. To gain a thorough understanding of the macroscopic behavior of the mean field BEG model, see [**14**].

4. Spin glasses

So far, we considered very simple models in which neighboring spins interact in a deterministic way. More realistic models would model interaction strengths depending not only on the geometry (we so far only considered mean field models, thus models with no geometry at all) but also on chance. Thinking of quantum mechanics it is inevitable

to think of interaction strength between two atoms as being random variables satisfying some sanity properties.

The following models are inspired by the so–called Sherrington-Kirkpatrick model (SK model) of a spin glass. In that model the Hamiltonian is given by

$$H_N(\sigma) = \frac{1}{\sqrt{2N}} \sum_{1 \leq i < j \leq N} \sigma_i \sigma_j J_{ij}.$$

Here σ is an element of the hypercube $\{0,1\}^N$ and the J_{ij} are independent $\mathscr{N}(0,1)$ Gaussian random variables. Hence for fixed σ, $H_N(\sigma)$ also is a Gaussian random variable with expectation 0. For $\sigma, \sigma' \in \{0,1\}^N$ we can compute the covariance

$$\operatorname{cov}(H_N(\sigma), H_N(\sigma')) = N R_N(\sigma, \sigma')$$

where $R_N(\sigma, \sigma') := \frac{1}{N} \sum_{i=1}^N \sigma_i \sigma_j$ is the so–called overlap between the configurations σ and σ'. Moreover, it is conjectured that the energy function of the SK model has an ultrametric structure (see e.g. [**30**]) which has been one of the key inspirations for the Parisi solution. These observations led Derrida to the introduction of the following toy models of a spin glass (see [**7**], [**8**], also see [**3**]).

4.1. The Random Energy Model. The Random Energy Model (REM) is defined on the state space $\Omega = \Omega_N = \{-1,1\}^N$. From the SK model one merely wants to keep the properties that for each σ the Hamiltonian is a Gaussian random variable with expectation 0 and variance N. This is easily accomplished by choosing

$$H_N(\sigma) = -Y_\sigma \tag{10}$$

with $Y_\sigma \sim \mathscr{N}(0, N)$ i.i.d.. Choosing H_N such that $\mathbb{V}(H_N(\sigma)) = N$ we obtain, as seen in Theorem 1.2, a model with a free energy of order N. An alternative notation we could use is taking

$$H_N(\sigma) = -\sqrt{N} X_\sigma$$

for our Hamiltonian, with 2^N many i.i.d. standard normal random variables X_σ. In favor of a standardized notation we will prefer the second choice. This implies that the Gibbs measure of the REM, which we want to simulate, is given by

$$\pi(\sigma) = \frac{e^{\beta H(\sigma)}}{Z(\beta)} = \frac{e^{-\beta\sqrt{N} X_\sigma}}{Z(\beta)}. \tag{11}$$

Here, of course,

$$Z(\beta) = \sum_{\sigma' \in \{0,1\}^N} e^{\beta H(\sigma')}$$

is the partition function that makes π a probability measure. For results on the REM the reader is referred to [**3**] or [**6**]. Here, among others, one can find the following estimate on the maximum energy in the REM:

LEMMA 1.1 ([**3**] Lemma 9.1.1). *The family X_σ of random variables satisfies*

$$\lim_{N\to\infty} \max_{\sigma\in\Omega_N} N^{-\frac{1}{2}} X_\sigma = \sqrt{2\ln(2)} \tag{12}$$

both almost surely and in mean.

As a consequence one obtains that the REM has a third order phase transition:

THEOREM 1.2 ([**3**] Theorem 9.1.2). *In the REM, with $\mathbb{P}$-probability 1*

$$\lim_{N\to\infty}\left[\frac{1}{N}\ln\left(2^{-N}\sum_\sigma e^{\beta H(\sigma)}\right)\right] = \begin{cases} \frac{\beta^2}{2} & \text{for } \beta \le \beta_c \\ \frac{\beta_c^2}{2} + (\beta-\beta_c)\beta_c & \text{for } \beta \ge \beta_c \end{cases}$$

holds with $\beta_c = \sqrt{2\ln 2}$.

4.2. The Generalized Random Energy Model. Of course, we cannot really hope to find a good approximation to the SK model by just keeping one of its characteristics. The idea of the Generalized Random Energy Model (GREM) (also invented by Derrida [**8**]) is to alter the REM in such a way that the energy of a configuration σ bears at least some information about the energy of neighboring configurations σ'. The way this is implemented is inspired by the ultrametricity conjecture in the SK model.

So again our state space is $\Omega := \Omega_N := \{-1,1\}^N$. To measure distance between two states we introduce

$$d_N(\sigma,\sigma') = \frac{1}{N}\left(\min\{i|\sigma_i \neq \sigma_i'\} - 1\right) \tag{13}$$

and note that $\exp(d_N(\sigma,\sigma'))$ is an ultra-metric on Ω. As was noted by Bovier [**3**] this choice of the distance function is basically the only difference between the GREM and the much challenging SK model.

We want the covariance between two states to be given by a non-decreasing function of d_N while assuming that $\mathbb{E}X_\sigma = 0$ for all $\sigma \in \Omega$. We therefore set

$$\operatorname{cov}(X_\sigma, X_{\sigma'}) = \mathbb{E}X_\sigma X_{\sigma'} = A(d_N(\sigma,\sigma')) \tag{14}$$

with $A(x)$ being a probability distribution function on $[0,1]$.

For the standard GREM let A denote the distribution function of a measure μ_A supported in $n \in \mathbb{N}$ many points $(x_i)_{0\le i\le n}$ with $x_0 = 0$, $x_n = 1$ and $x_i \in (0,1)$. Further assume that

$$x_i < x_j \quad \text{for } i < j.$$

Set $a_i := \mu_A(x_i)$ and assume that $a_i > 0$ for all $i \in \{1,...,n-1\}$. Then we define

$$\alpha_i := 2^{x_i - x_{i-1}} \text{ for } i = 1,...,n \tag{15}$$

so we get $\sum a_i = 1$ and $\prod \alpha_i = 2$.

Now we can decompose Ω as an n-fold product

$$\Omega = \Omega_N = \{-1,1\}^N = \bigotimes_{i=1}^{n} \{-1,1\}^{N\frac{\ln \alpha_i}{\ln 2}} = \bigotimes_{i=1}^{n} \Omega_{N\frac{\ln \alpha_i}{\ln 2}}. \tag{16}$$

Hence each $\sigma \in \Omega$ can be written as an n-tuple

$$\sigma = \sigma_1\sigma_2...\sigma_n \quad \text{ with } \sigma_i \in \Omega_{N\frac{\ln \alpha_i}{\ln 2}}.$$

Now the (Gaussian) Hamiltonian – which is indeed a Gaussian process X_σ in σ – can be constructed such that there is a contribution from each of the factors that build Ω. To this end, take $\alpha_1^N + (\alpha_1\alpha_2)^N + ... + (\alpha_1...\alpha_n)^N$ many independent $\mathscr{N}(0,1)$-distributed random variables enumerated in the following way: The first α_1^N many X_{σ_1} are indexed by $\sigma_1 \in \Omega_{N\frac{\ln \alpha_1}{\ln 2}}$. The next $(\alpha_1\alpha_2)^N$ many $X_{\sigma_1\sigma_2}$ are indexed by $\sigma_1 \in \Omega_{N\frac{\ln \alpha_1}{\ln 2}}$ and $\sigma_2 \in \Omega_{N\frac{\ln \alpha_2}{\ln 2}}$, and so forth, so that (with $\sigma = \sigma_1\sigma_2...\sigma_n$)

$$X_\sigma := \sqrt{a_1}X_{\sigma_1} + \sqrt{a_2}X_{\sigma_1\sigma_2} + ... + \sqrt{a_n}X_{\sigma_1...\sigma_n} \tag{17}$$

is a $\mathscr{N}(0,1)$-distributed random variable as well and obviously $\operatorname{cov}(X_\sigma, X_{\sigma'})$ depends on d_N in the desired way. Analogously to the REM we define

$$H(\sigma) := -\sqrt{N}X_\sigma \tag{18}$$

which gives us

$$\pi(\sigma) = \frac{e^{\beta H(\sigma)}}{Z(\beta)} = \frac{e^{-\beta\sqrt{N}X_\sigma}}{Z(\beta)} \tag{19}$$

as our Gibbs measure. Here again

$$Z(\beta) = \sum_{\sigma' \in \{0,1\}^N} e^{\beta H(\sigma')}$$

is the partition function of the model. For further results on the GREM the reader is referred to Bovier's book [**3**].

CHAPTER 2

Markov-Chain-Monte-Carlo Method

Having introduced the ideas of statistical mechanics in Chapter 1 we can now come to introducing the probability part of statistical mechanics. Given the Boltzmann distribution π for a given finite model at a given inverse temperature β, a physically realistic realization of the model can be obtained by drawing out of all realizations according to π. Even though this might sound easy, most models require an extensive amount of calculations in order to do so. We need to know the probability of any possible state in order to simulate this distribution with a uniform-random-number-generator on $[0, 1]$. Here, the idea of the Markov-Chain-Monte-Carlo Method (MCMC) comes to mind in a rather natural way. Thinking of the dynamics of a thermodynamic system, we could imagine that at some point in time, a single atom is being selected and its spin value is being changed in some way. At some later point in time, another atom is selected and its spin is changed as well. Waiting long enough the system will equilibrate, while this dynamics keeps on going, keeping the system in equilibrium henceforth. This idea carries over to Markov chains as we will see in Section 3. We will first introduce the MCMC-Method in a more general setting.

1. Definition of the MCMC Method

We first give two definitions:

DEFINITION 2.1. *Let $\mathcal{S} = \{s_1, ..., s_k\}$ be a finite state space and let P be a $(k \times k)$ matrix. A stochastic process $(X_i)_{i\in\mathbb{N}}$ on Ω is called a homogeneous Markov chain with the transition matrix P, if for all n and all $i_0, i_1, ..., i_{n+1} \in \{1, ..., k\}$*

$$\begin{aligned}\mathbb{P}(X_{n+1} = s_{i_{n+1}} | X_0 = s_{i_0}, ..., X_n = s_{i_n}) &= \mathbb{P}(X_{n+1} = s_{i_{n+1}} | X_n = s_{i_n}) \\ &= P_{i_n, i_{n+1}}\end{aligned}$$

holds.

DEFINITION 2.2. *Let $(X_0, X_1, ...)$ be a Markov chain with finite state space $\mathcal{S} = \{s_1, ..., s_k\}$ and transition matrix P. A probability measure π on $\mathcal{S}$ is said to be a stationary distribution for the Markov chain, if it satisfies*

$$\sum_{i=1}^{k} \pi_i P_{i,j} = \pi_j \text{ for } j = 1, ..., k.$$

DEFINITION 2.3. *Let $\mathcal{A}$ be a sigma-field on a set Ω. The total variation distance between two probability measures π and τ on $(\Omega, \mathcal{A})$ is defined by*

$$d(\pi, \tau)_{TV} := \sup\big\{|\pi(A) - \tau(A)|\big|A \in \mathcal{A}\big\}.$$

The fundamental result for all that follows is

THEOREM 2.4 (Ergodic Theorem for Markov chains). *Let $(X_0, X_1, X_2, ...)$ be an irreducible aperiodic Markov chain with state space $\mathcal{S} = \{s_1, ..., s_k\}$, transition matrix P and arbitrary initial distribution $\mu^{(0)}$. Then there exists a unique distribution π which is stationary for the transition matrix P. If $\mu^{(n)}$ denotes the distribution of X_n then*

$$\mu^{(n)} \xrightarrow{TV} \pi.$$

See [**22**] for the proof, to gain an elementary understanding of Markov chains and to find all definitions needed for formulating Theorem 2.4.

In general, the definition of stationarity proves complicated to construct or to verify for a given transition matrix P or for a given probability distribution π. There is the tighter concept of reversibility which, in most cases, is much easier to construct.

DEFINITION 2.5. *Let $(X_0, X_1, ...)$ be a Markov chain with state space $\mathcal{S} = \{s_1, ..., s_k\}$ and transition matrix P. A probability distribution π on $\mathcal{S}$ is said to be reversible for the chain if for all $i, j \in \{1, ..., k\}$ we have*

$$\pi_i P_{i,j} = \pi_j P_{j,i}.$$

The Markov chain is said to be reversible if there exists a reversible distribution for it.

It is common knowledge that a reversible distribution π to a transition matrix P is also stationary to P.

2. Technical preparation: Gap and Conductance

The key question for all kind of MCMC algorithms is how fast they mix, i.e. how rapidly they converge to the desired invariant measure. So in general, let X_n be a homogeneous, irreducible and aperiodic Markov chain on a finite state space Ω, reversible with respect to a probability measure π (on Ω, that necessarily charges every point). The speed of convergence is determined in terms of

$$\tau(\varepsilon) = \min\{n : \mathrm{d_{TV}}(\mathbb{P}^{X_n}, \pi) \leq \varepsilon\}.$$

Here, of course, $\mathbb{P}^{X_n}$ is the distribution at time n of the Markov chain corresponding to the algorithm and $\mathrm{d_{TV}}(\mathbb{P}^{X_n}, \pi)$ is the total variation distance between this distribution at time n and the invariant measure π of the chain. Rapid convergence of such a MCMC algorithm means that one can bound $\tau(\varepsilon)$ by a polynomial in ε^{-1} and the problem size.

There is an intrinsic relationship between $\tau(\varepsilon)$ and the spectral gap of the chain defined by

$$\mathrm{Gap}((X_n)) := \mathrm{Gap}(P) := 1 - \max\{|\lambda_i|, \lambda_i \neq 1\} =: 1 - |\lambda_1|,$$

where we write λ_i for the eigenvalues of the transition matrix $P = (P(i,j))_{i,j}$ of the chain (X_n) and have λ_1 denote the second largest eigenvalue. As a matter of fact, for an irreducible and aperiodic chain the following estimates hold true (see e.g. **[31]**): Let $\underline{\pi} := \min_x \pi(x)$ (which is non-zero by the ergodic theorem for Markov chains), then

$$\tau(\varepsilon) \leq \frac{1}{\mathrm{Gap}(\mathbb{P}^{X_n})} \log(\frac{1}{\underline{\pi}\varepsilon})$$

as well as

$$\tau(\varepsilon) \geq \frac{|\lambda_1|}{2\mathrm{Gap}(\mathbb{P}^{X_n})} \log(\frac{1}{2\varepsilon}).$$

We can thus control the speed of convergence of the Markov chain (or the MCMC algorithm, respectively), if we control the size of the spectral gap of P. There is, of course, a variety of methods to obtain such a control. In this chapter we will only need

THEOREM 2.6 (Jerrum and Sinclair **[23]**). *Let P be a Markov chain on a finite set Ω reversible with respect to π. For all $\mathcal{S} \subset \Omega$, define*

$$\Phi_{\mathcal{S}} = \frac{\sum_{x\in\mathcal{S}, y\notin\mathcal{S}} \pi(x)P(x,y)}{\pi(\mathcal{S})},$$

to have the conductance Φ given by

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

Then the following holds true:

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

3. Metropolis-Hastings Algorithm

With these ingredients we can catch up on the idea stated earlier in order to gain a realization of a model. We have got the Boltzmann distribution π we want to sample from. All that has to be done is to find an appropriate Markov chain which has stationary distribution π, start this chain at an arbitrary distribution $\mu^{(0)}$ (which could be a Dirac measure) and wait long enough to get a sample which is drawn according to a good approximation of π.

Remembering the definition of the Boltzmann-Distribution given in (2) we note that even though it is easy to calculate the enumerator, in general, calculating the denominator takes exponentially many steps. So in order to give an algorithm for sampling from the Boltzmann distribution efficiently we need to find a Markov chain with a transition kernel which does not depend on the normalization constant $Z(\beta)$.

Let Ω be a finite state space and π be a probability distribution on Ω with $\pi(x) > 0$ for all $x \in \Omega$. Further let K be the transition matrix of a symmetric and irreducible Markov chain on Ω. Defining

$$T(x,y) := \begin{cases} K(x,y) & \text{if } \pi(y) \geq \pi(x) \text{ and } x \neq y \\ K(x,y)\frac{\pi(y)}{\pi(x)} & \text{if } \pi(y) < \pi(x) \\ 1 - \sum_{z \neq x} T(x,z) & \text{if } x = y \end{cases}$$

yields an irreducible Markov transition matrix T on Ω which is reversible with respect to π. The constructed Markov chain is called Metropolis-Hastings chain. Note that there are several different Metropolis—Hasting chains for the same distribution π, as T depends heavily on the choice of the proposal chain K. If $K(x,x) > 0$ for all $x \in \Omega$ so is $T(x,x) > 0$ for all $x \in \Omega$. Thus T is aperiodic in this case. Also note that $K(x,x) \geq \frac{1}{2}$ implies $T(x,x) \geq \frac{1}{2}$, which will be of interest in conjunction with Lemma 3.3.

Coming back to the Boltzmann distribution

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

and the need for an algorithm that gives good samples from π we will use our previously stated physical intuition in order to construct a Metropolis-Hastings chain for a statistical mechanics model. Say we have N atoms and any atom can be in one of finitely many spins. Consider as proposal chain a Markov chain, that suggests to select one atom and to change this atom's spin to one of the possible spins. We can do both according to the uniform distribution which in the models given in Chapter 1 is an easy task (linearly in the model size N and in the amount of spins an atom can be in). This fulfills $K(\sigma,\sigma) > 0$ for any state $\sigma \in \Omega$ and thus the constructed Metropolis chain satisfies the Ergodic Theorem for Markov chains with stationary distribution π. It is important that T does not depend on the constant $Z(\beta)$, as in the fraction

$$\frac{\pi(\tau)}{\pi(\sigma)} = \frac{e^{\beta H(\sigma)}}{e^{\beta H(\tau)}}$$

the normalization constant $Z(\beta)$ cancels out, the comparisons can be done without knowing $Z(\beta)$ and everything else is independent of π as long as K does not depend on π. Now there is good hope that using the Metropolis algorithm can reduce computational cost compared to drawing from π directly. This hope however proves to be not necessarily true as can be seen in the easy case of the Curie-Weiss model in Section 4 of this chapter.

4. Torpid mixing of Metropolis in the Curie-Weiss-Model

Consider the Curie-Weiss model given in Section 1 of Chapter 1. The Hamiltonian is given through

$$H(\sigma) = \frac{1}{2N} m(\sigma)^2$$

for $\sigma \in \Omega = \{-1, 1\}^\Lambda$ with $\Lambda = \{1, ..., N\}$ and $m(\sigma) = \sum_{i=1}^N \sigma_i$. The distribution of interest is

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

which we want to give an intuitive proposal chain K for. Define

$$K(\sigma, \tau) := \begin{cases} \frac{1}{2N} & \text{if } \|\sigma - \tau\|_1 = 2 \\ 1 - \sum_{\tau' \neq \sigma} K(\sigma, \tau') & \text{if } \sigma = \tau \\ 0 & \text{otherwise} \end{cases}$$

to be the transition matrix of the Markov chain which inverts a randomly selected coordinate, thus changing a 1 to a -1 or the other way around. With probability $\frac{1}{2}$ the chain does not change its state, which guarantees $K(\sigma, \sigma) \geq \frac{1}{2}$ for every $\sigma \in \Omega$, thus K is a positive operator. Defining

$$T(\sigma, \tau) := \begin{cases} K(\sigma, \tau) & \text{if } \pi_\beta(\tau) \geq \pi_\beta(\sigma) \text{ and } \sigma \neq \tau \\ K(\sigma, \tau) \frac{\pi_\beta(\tau)}{\pi_\beta(\sigma)} & \text{if } \pi_\beta(\tau) < \pi_\beta(\sigma) \\ 1 - \sum_{\tau' \neq \sigma} T(\sigma, \tau') & \text{if } \sigma = \tau \end{cases}$$

gives an irreducible, aperiodic and, with respect to π_β, reversible Markov chain with transition matrix T.

We are going to see that this Metropolis chain is actually torpidly mixing for any $\beta > \beta_c$. The main idea here is as follows: It is well known that the Curie-Weiss model exhibits a phase transition at $\beta_c = 1$. For $\beta < \beta_c$ the system has only one macrostate, while for $\beta > \beta_c$ the system has two distinct macrostates. At finite N this can be expressed in the term of modes. For $\beta < \beta_c$ the system consists of only one mode, while for $\beta > \beta_c$ the system has two distinct modes. One mode has all states with negative total magnetization, while the other mode consists of the states with positive total magnetization. A Metropolis chain started in one of these modes will actually explore this mode very fast and will therefore equilibrate rapidly in this mode. Due to the symmetry of the model both modes have the same probability, thus the chain should be seeing both modes equally often in order for the chain to rapidly mix on the whole state space. Now going from one mode to the other requires the chain to pass through a region of the state space in which the total magnetization is close to zero. For $\beta > \beta_c$ this region has only exponential little mass and the chain started in one mode will take exponentially long to pass this region. In order to give

a short proof which does not rely too much on tedious calculations this intuition is adapted a little bit in order to show

THEOREM 2.7. *For every $\beta > \beta_c = 1$ there exists a $c > 0$ such that*

$$\mathrm{Gap}(T) \leq e^{-cN}.$$

Therefore the Metropolis chain induced by the proposal chain K for the Curie-Weiss model in the two mode region is torpidly mixing.

PROOF. Assume N to be even. The idea of the proof will also work for N odd, but this will rid us of non-instructive case differentiations. Define $A_i := \{\sigma \in \Omega | m(\sigma) = i\} \subseteq \Omega$ to contain all the states with the given total magnetization i. Obviously $\pi_\beta(A_i) = 0$ for i odd. So let henceforth only consider i as being even. The Hamilton function does not differentiate between different states in one of these sets. Using Stirling's approximation and $a_1 := \frac{i+N}{2N}$ as the relative amount of $+1$ spins we calculate

$$\begin{aligned}\pi_\beta(A_i) &= \binom{N}{Na_1} Z(\beta)^{-1} e^{2\beta N(2a_1-1)^2} \\ &= Z(\beta)^{-1} N^{-\frac{1}{2}} e^{N\left(\frac{\beta}{2}(2a_1-1)^2 - a_1\log(a_1) - (1-a_1)\log(1-a_1)\right) + \Delta(a_1)} \qquad (20)\end{aligned}$$

with $\Delta(a_i) = O(1)$ if there exists $\varepsilon > 0$ with $\varepsilon < a_1 < 1 - \varepsilon$. Consider

$$f(a_1) := \frac{\beta}{2}(2a_1 - 1)^2 - a_1 \log(a_1) - (1 - a_1)\log(1 - a_1))$$

as a smooth function $f : (0, 1) \to \mathbb{R}$. For $\beta > 1$ it is easy to verify, that f has a local minimum at $a_1 = \frac{1}{2}$. We further see that f is symmetric to the $\frac{1}{2}$-axis and

$$\lim_{a_1 \to 1} f(a_1) = \lim_{a_1 \to 0} f(a_1) = 0.$$

Therefore we can find $\frac{1}{2} < a_m < 1$ with $f(a_m) \geq f(a_1)$ for all $a_1 \geq \frac{1}{2}$. Note that it is possible to find $\varepsilon > 0$ such that f is strictly concave on $(a_m - 2\varepsilon, a_m + 2\varepsilon)$.

Define the sets

$$\mathcal{N} := \left\{\sigma \in \Omega \middle| \left|\frac{m(\sigma)}{N} - a_m\right| < 2\varepsilon\right\}$$

and

$$\mathcal{N}_{\text{edge}} := \mathcal{N} \setminus \left\{\sigma \in \Omega \middle| \left|\frac{m(\sigma)}{N} - a_m\right| < \varepsilon\right\}$$

and choose N at least big enough, such that both sets are nonempty. Obviously there are only N many non empty sets A_i. Due to the exponential structure of (20) in N we therefore gain

$$\frac{\pi_\beta(\mathcal{N}_{\text{edge}})}{\pi_\beta(\mathcal{N})} \leq e^{-cN}$$

for a constant $c > 0$.

We have now everything set up for proofing that $\mathcal{N}_{\text{edge}}$ constitutes a bad cut in the state space by using a conductance argument. As $\pi_\beta(\mathcal{N}) \leq \frac{1}{2}$ using Theorem 2.6 with

$$\begin{aligned}\Phi_{\mathcal{N}} &= \frac{\sum_{\sigma\in\mathcal{N},\tau\notin\mathcal{N}} \pi(\sigma)T(\sigma,\tau)}{\pi(\mathcal{N})} \\ &= \frac{\sum_{\sigma\in\mathcal{N}_{\text{edge}}} \pi(\sigma)\sum_{\tau\notin\mathcal{N}} T(\sigma,\tau)}{\pi(\mathcal{N})} \\ &\leq \frac{\sum_{\sigma\in\mathcal{N}_{\text{edge}}} \pi(\sigma)}{\pi(\mathcal{N})} \\ &\leq e^{-cN}.\end{aligned}$$

concludes the proof. □

CHAPTER 3

Simulated Tempering and Swapping

We introduce two variants of the Metropolis–Hastings Algorithm in this section. These algorithms include an additional change of temperature with the idea to speed up the Metropolis chain when it is slow. They are specifically tailored for situations where the invariant measure is a Gibbs measure with respect to some energy function and the Metropolis Algorithm mixes slowly at low temperatures but quickly at high temperatures. We start with the Simulated Tempering Algorithm proposed by Geyer and Thompson [**20**].

1. Simulated Tempering

From now on and for the rest of the chapter let us assume that the target distribution is a Gibbs measure on a finite set Ω. Let $H(\cdot)$ be the corresponding energy function or Hamiltonian of the system. For every inverse temperature $\beta > 0$ a probability function on Ω is given by

$$\pi_\beta(\sigma) := \frac{e^{\beta H(\sigma)}}{\sum_{\sigma' \in \Omega} e^{\beta H(\sigma')}} = \frac{e^{\beta H(\sigma)}}{Z(\beta)} \tag{21}$$

We have seen that, despite being natural, the metropolis algorithm is sometimes slow in natural situations, e.g. when sampling from the low temperature distribution of the Curie-Weiss model. To speed up its convergence, we consider $\Omega \times \{0, 1, ..., M\}$ for some $M \in \mathbb{N}$ as state space, which is typically chosen as $M := c_1 N$ for some constant $c_1 > 0$. The second component of the new state space refers to the current temperature of the model (or the chain, resp.). Define $\beta_i := \frac{i}{M}\beta$ and the probability measures $\pi_i := \pi_{\beta_i}$. On $\Omega \times \{0, ..., M\}$ we take $\pi(x) = \pi((\sigma, i)) = \frac{1}{M+1}\pi_i(\sigma)$ as probability measure. We construct a Markov chain that starts in $(\sigma, i) \in \Omega \times \{0, 1, ..., M\}$ and chooses a new state (σ', i) according to the metropolis chain T_{β_i} introduced in Section 3 of Chapter 2. In a second step the temperature is changed according to a similar Metropolis chain. The idea is that in case of the chain being in an energy-valley, it can increase its temperature (reduce β) and thereby reduce the cost of switching to another energy-valley. Explicitly, this works as follows:

In the first step let $i \in \{0, ..., M\}$ be fixed. Then a transition from (σ, i) to (σ', i) has probability $P_{st}((\sigma, i), (\sigma', i)) := T_{\beta_i}(\sigma, \sigma')$. In the

second step let $\sigma \in \Omega$ be fixed. Then the chain moves from (σ, i) to (σ, j) according to the transition probabilities

$$Q((\sigma,i),(\sigma,j)) := \begin{cases} K_{\mathrm{tm}}(i,j) & \text{if } \pi_j(\sigma) \geq \pi_i(\sigma) \\ & \text{and } i \neq j \\ K_{\mathrm{tm}}(i,j)\frac{\pi_j(\sigma)}{\pi_i(\sigma)} & \text{if } \pi_j(\sigma) < \pi_i(\sigma) \\ 1 - \sum_{k \neq i} Q((\sigma,i),(\sigma,k)) & \text{if } i = j \end{cases}$$

with

$$K_{\mathrm{tm}}(i,j) := \begin{cases} \frac{1}{2(M+1)} & \text{if } j = i \pm 1 \text{ and } j \in \{0,\ldots,M\} \\ 0 & \text{if } |i-j| > 1 \\ 1 - \sum_{k \neq i} K_{\mathrm{tm}}(i,k) & \text{if } i = j. \end{cases}$$

The actual Simulated Tempering algorithm now consists of any reasonable combination of these two chains. Usually one first applies a temperature move Q, then a Metropolis move at the present temperature (the transition matrix of which is denoted by P_{st}), and finally another temperature move. The precedence for this combination is because it can easily be verified that this combination yields a Markov chain which is reversible with respect to π if Q and P_{st} themselves are. Hence, in terms of transition matrices the Simulated Tempering algorithm is given by $QP_{st}Q$.

Notice that the computation of $\frac{\pi_j(\sigma)}{\pi_i(\sigma)}$ in the matrix Q needs knowledge of the normalizing constants $Z(\beta_i)$ and $Z(\beta_j)$ which in most cases is hard to get by. This is the reason for introducing the now following Swapping Algorithm.

2. Swapping

The so called Swapping Algorithm was suggested by Geyer in [**19**]. The basic idea of changing the temperature is maintained. As state space for the Swapping chain we choose:

$$\Omega^{\mathrm{sw}} := \Omega^{M+1}$$

A natural choice for a probability measure on Ω^{sw} is:

$$\pi(x) := \prod_{i=0}^{M} \pi_i(x_i) = \frac{\prod_{i=0}^{M} e^{\frac{i\beta}{M}H(x_i)}}{\prod_{i=0}^{M} Z(\beta_i)} \tag{22}$$

with $x = (x_0, ..., x_M) \in \Omega^{\mathrm{sw}}$. As in the Simulated Tempering Algorithm the Swapping Algorithm consists of two steps. In the first step, we choose an $i \in \{0, ..., M\}$ uniformly and update the i-th component of the current state $x = (x_0, ..., x_M)$ according to the usual Metropolis chain T_{β_i} at inverse temperature β_i. In the second step, we choose an

$i \in \{0, ..., M-1\}$ uniformly at random and swap the components x_i and x_{i+1} of x with probability

$$\min\left(1, \frac{\pi(x_0, ..., x_{i+1}, x_i, ..., x_M)}{\pi(x_0, ..., x_i, x_{i+1}, ..., x_M)}\right).$$

So explicitly the first step works as follows: The transition probabilities from

$$x = (x_0, ..., x_{i-1}, x_i, x_{i+1}, ..., x_M) \in \Omega^{\mathrm{sw}}$$

to

$$x' = (x_0, ..., x_{i-1}, x'_i, x_{i+1}, ..., x_M) \in \Omega^{\mathrm{sw}}$$

are $T_i(x, x') := T_{\beta_i}(x_i, x'_i)$. Note that the product chain

(23)

$$P(x,y) = \frac{1}{2}\delta(x,y) + \frac{1}{2(M+1)} \sum_{i=0}^{M} \delta(x_0, y_0) \cdot ... \cdot \delta(x_{i-1}, y_{i-1}) T_i(x_i, y_i) \times \delta(x_{i+1}, y_{i+1}) \cdot ... \cdot \delta(x_M, y_M)$$

gives us a Markov chain on Ω^{sw}. Also note that we never change more than one component at a time. The second step is the temperature swap. Here the transition probabilities from $x = (x_0, ..., x_i, x_{i+1}, ..., x_M)$ to $x' = (x_0, ..., x_{i+1}, x_i, ..., x_M)$ are

$$Q(x,x') := \begin{cases} K_{\mathrm{sw}}(x,x') & \text{if } \pi(x') \geq \pi(x) \text{ and } x \neq x' \\ K_{\mathrm{sw}}(x,x')\frac{\pi(x')}{\pi(x)} & \text{if } \pi(x') < \pi(x) \\ 1 - \sum_{z \neq x} Q(x,z) & \text{if } x = x'. \end{cases}$$

K_{sw} is defined by

$$K_{\mathrm{sw}}(x,x') := \begin{cases} \frac{1}{2M} & \text{if } \exists i \text{ with } x_j = x'_j \ \forall j \notin \{i, i+1\} \\ & \text{and } x_i = x'_{i+1}, x_{i+1} = x'_i, x \neq x' \\ 0 & \text{if } \nexists i \text{ with } x_j = x'_j \ \forall j \notin \{i, i+1\} \\ & \text{and } x_i = x'_{i+1}, x_{i+1} = x'_i, x \neq x' \\ 1 - \sum_{k \neq i} K_{\mathrm{sw}}(i,k) & \text{if } x = x'. \end{cases}$$

Note that the factor $\frac{1}{2}$ in the definition of K_{sw} and P guarantees that both transition kernels, P and Q, are aperiodic and that the corresponding operators are positive as seen in Lemma 3.3. Notice that all the normalizing constants in Q and P cancel out, such that the transition probabilities can be effectively computed.

The Swapping algorithm is now any reasonable combinations of P and Q, usually one takes QPQ as it is reversible with respect to π if Q and P are reversible (which in our situation is the case). The following sections will give an idea of what is known about these algorithms so far.

3. Known results

3.1. Madras and Zheng. In their paper [**29**] Madras and Zheng are the first to consider the swapping algorithm for a model of statistical mechanics. As we have seen in Section 4 of Chapter 2 a natural realization of the Metropolis algorithm for the Curie-Weiss model with no external field is torpidly mixing. By using techniques invented by Madras, Piccioni [**27**] and Madras, Randall [**28**] Madras and Zheng were able to show that the swapping algorithm with the same natural Metropolis sampler as an underlying updating chain is rapidly mixing. That way they were able to show that at least in some cases the swapping algorithm is provably better than the standard metropolis sampler. This has actually previously been claimed by some physicists who were only able to give simulations as evidence.

The proof Madras and Zheng give relies heavily on the decomposition theorem, in this thesis Theorem 3.9. Consider a state with total magnetization $m(\sigma)$. By inverting each spin the new state $-\sigma$ has total magnetization $m(-\sigma) = -m(\sigma)$, thus there is a one to one correspondence of the set of all states with positive magnetization and the set with states of negative total magnetization. Using a smart aggregation of states they restrict the chain to only switch sign in $m(\sigma)$ at inverse temperature $\beta = 0$. They can then use the fact that the sets described previously fulfill a unimodality condition in some sense, thus using the well known Poincaré inequality given in Theorem 3.4 gives rapid mixing on each subset in any temperature. Putting this together guarantees rapid convergence of the chain after some tedious calculations.

Apart from looking at the Curie-Weiss model the paper also contains the proof of rapid convergence to equilibrium of the exponential valley distribution. This paper seems to have come out of Zheng's PhD-Thesis in which Zheng also considers the relationship between the two algorithms of simulated tempering and swapping introduced in section 1 and section 2. With

THEOREM 3.1 (Zheng [**35**]). *If there exists a $\delta > 0$ such that*

$$\sum_{x \in \Omega} \min\{\pi_i(x), \pi_{i+1}(x)\} \geq \delta \quad \textit{for all } 1 \leq i \leq M$$

holds then if the Swapping algorithm converges in polynomial time, so does the Simulated Tempering algorithm.

he was able to show (under moderate regularity conditions) that rapid convergence to equilibrium of the swapping algorithm implies rapid convergence of the tempering algorithm. This is readily used to show rapid convergence to equilibrium of the tempering chain for the Curie-Weiss model.

3.2. Bhatnagar and Randall. Bhatnagar and Randall picked up on the idea of tempering and swapping in [**2**]. A natural extension of

the Curie-Weiss model would be to consider the Potts model. This model has $\{1, ..., q\}^N$ as its state space, thus every atom can have one of q many spin values, which, in this model are usually called *colors*. The Hamiltonian is then given by

$$H(\sigma) = \frac{1}{N} \sum_{i,j=1}^{N} \delta_{\sigma_i, \sigma_j}$$

which makes the Potts model an extension of the Curie-Weiss model in the sense that it has more than two colors. Looking at the idea of Theorem 2.7 it is not difficult to think of a proof which grands torpid mixing of the metropolis algorithm which would be the natural adaption of the one used for the Curie-Weiss model. The maybe surprising result is that this slight modification of the Curie-Weiss model already warrants for torpid mixing of the tempering and thus swapping algorithm as well. Bhatnagar and Randall use the fact that even though below the critical temperature the Potts model with three colors has three macrostates, in this respect extending the Curie-Weiss model in the anticipated way, it actually has a fourth, local, mode in the center of the state space which does not yield a macrostate. This persists at all temperatures below the critical one. At $\beta = 0$ the Potts model has only one mode which yields the macrostate of all colors appearing equally often. A tempering chain started at this temperature in one of the center states will stay in the center of the state space, as this area is separated from any other mode by an exponentially deep energy barrier. Even though the position of the barrier is shifted by changing the inverse temperature β, the property itself exists at any temperature. The proof relies on a conductance argument involving the bad cut described and Lemma 2.6.

In the second part of their paper Bhatnagar and Randall suggest a slightly different technique compared to the one introduced by Madras and Zheng in order to show rapid convergence of the swapping algorithm to equilibrium. It is a minor modification which might make proving rapid convergence to equilibrium of the swapping chain in many interesting situations easier. Chapter 5 of this Thesis gives a detailed use of their technique, extending it in those parts where Bhatnagar and Randall only gave a slightly fragmentary idea.

3.3. Huber, Schmidler and Woodard. Huber, Schmidler and Woodard extend the techniques given by Madras, Zheng and Bhatnagar, Randall to a slightly more general setting.

3.3.1. *Rapid mixing results.* In **[32]** they give general properties for which a model satisfying these will have a rapidly mixing swapping chain based on a reasonable underlying metropolis chain. The conditions they give are seemingly the outer most possible conditions which

still make the general proof given by Madras and Zheng for the Curie-Weiss model go through unharmed. Using their result they show rapid convergence to equilibrium of the swapping algorithm for two models. The first one is the Curie-Weiss model. Even though the proof given fits on only one page, it relies heavily on calculations done by Madras and Zheng in their much longer paper. The second model is the Symmetric Mixture of Normals in $\mathbb{R}^M$.

We will give a brief summary of their result as it will be used in the case of the Generalized-Curie-Weiss model in Chapter 4. Let Ω be a finite state space with the desired probability distribution π_β. Further let T_β denote the transition kernel of a Metropolis chain corresponding to π_β. Let $\mathcal{A} = \{A_j, j = 1, ..., J\}$ be a partition of Ω. (For further details on the construction see Theorem 3.9.) Let $\overline{T_0}$ denote the aggregated transition chain of T_0 with respect to the partition $\mathcal{A}$. Further define the *overlap* of $\{\pi_k : k = 0, ..., M\}$ with respect to $\mathcal{A}$ by

$$\delta(\mathcal{A}) = \min_{\substack{|k-l|=1 \\ j\in\{1,...,J\}}} \left[\frac{\sum_{x\in A_j} \min\{\pi_{\beta_k}(x), \pi_{\beta_l}(x)\}}{\pi_{\beta_k}(A_j)} \right] \tag{24}$$

and have

$$\gamma(\mathcal{A}) = \min_{j\in\{1,...,J\}} \prod_{k=1}^{M} \min\left\{1, \frac{\pi_{\beta_{k-1}}(A_j)}{\pi_{\beta_k}(A_j)}\right\} \tag{25}$$

denote a quantity that measures the relative loss of mass in one of the A_j. With these definitions in place Huber, Schmidler and Woodard show

THEOREM 3.2 (Theorem 3.1 in **[32]**). *For any partition $\mathcal{A} = \{A_j, j = 1, ..., J\}$ of Ω let P_{sc} denote the swapping chain induced by the Metropolis chain T_β. The spectral gap of P_{sc} satisfies*

$$\mathrm{Gap}(P_{sc}) \geq \left(\frac{\gamma(\mathcal{A})^{J+3}\delta(\mathcal{A})^2}{2^{11}(N+1)^4 J^3} \right) \mathrm{Gap}(\overline{T_0}) \cdot \min_{k,j} \mathrm{Gap}(T_k|_{A_j}).$$

3.3.2. *Torpid mixing results.* In **[33]**, Huber, Schmidler and Woodard once again extend the proof given by Bhatnagar and Randall to the outer most possible conditions a general model needs to satisfy in order for the proof given by Bhatnagar and Randall for the Potts model to work unscratched. They are able to apply their result to the Curie-Weiss model's tempering chain if one fixes the amount of temperatures in the algorithm. The generalization of the model "Symmetric Mixture of Normals in $\mathbb{R}^M$" to a Mixture of Normals with Unequal Variance in $\mathbb{R}^M$ yields a torpidly mixing tempering chain. The third model they are able to consider is again the Potts model with $q \geq 3$ colors.

4. Technical preparations

For the proofs of the results in the following chapters, we need several well known results on Markov chains. The proofs of these results can be found in the citation given. A slightly longer introduction with some simple proofs can also be found in [**12**].

LEMMA 3.3 (Lemma 3 of [**29**]). *Let P be a Markov chain that is reversible with respect to a probability measure π on the finite state space $\mathcal{S}$. Also assume that $P(x,x) \geq \frac{1}{2}$ for every $x \in \mathcal{S}$. Then P is a positive operator.*

LEMMA 3.4 (Poincaré inequality, Proposition 1' of [**11**]). *Let P be an irreducible and reversible Markov chain on a finite state space $\mathcal{S}$. We associate to P the graph with vertex set $\mathcal{S}$ and edges $\langle x,y\rangle$ if and only if $P(x,y) > 0$. For each pair of distinct points $x, y \in \mathcal{S}$, we choose a path γ_{xy} from x to y, such that a given edge appears at most once in a given path. Then the second largest eigenvalue λ_1 of P satisfies*

$$\lambda_1 = 1 - \mathrm{Gap}(P) \leq 1 - \frac{1}{A}$$

where

$$A := \max_{\langle x,y\rangle} \frac{1}{\pi(x)P(x,y)} \sum_{\gamma_{z_1z_2} \ni \langle x,y\rangle} |\gamma_{z_1z_2}|\pi(z_1)\pi(z_2)$$

and $|\gamma_{z_1z_2}|$ denotes the number of edges in the path $\gamma_{z_1z_2}$.

LEMMA 3.5 (Comparison of Dirichlet forms, Theorem 2.1 of [**9**]). *Let P, π and $\tilde{P}, \tilde{\pi}$ be reversible Markov chains on a finite state space $\mathcal{S}$, with respective Dirichlet forms $\mathcal{E}$ and $\tilde{\mathcal{E}}$. For each pair $x \neq y$, with $\tilde{P}(x,y) > 0$, we fix a path $\gamma_{xy} = (x_0 = x, x_1, x_2, \ldots, x_k = y)$, such that $P(x_i, x_{i+1}) > 0$, of length $|\gamma_{xy}| = k$. Set $E = \{(x,y) : P(x,y) > 0\}$, $\tilde{E} = \{(x,y) : \tilde{P}(x,y) > 0\}$ and $\tilde{E}(e) = \{(x,y) \in \tilde{E} : e \in \gamma_{xy}\}$, where $e \in E$. Then*

$$\tilde{\mathcal{E}} \leq A\mathcal{E},$$

where

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

LEMMA 3.6 (Lemma 5 of [**29**]). *Let (K,π) and (K',π') be two Markov chains on the same finite state space $\mathcal{S}$, with respective Dirichlet forms $\mathcal{E}$ and $\mathcal{E}'$. Assume that there exists constants $A, a > 0$ such that*

$$\mathcal{E}' \leq A\mathcal{E} \quad \textit{and} \quad a\pi \leq \pi'.$$

Then

$$\mathrm{Gap}(K') \leq \frac{A}{a}\,\mathrm{Gap}(K).$$

Remark A sufficient condition for $\mathcal{E}' \leq A\mathcal{E}$ is that

$$\pi'(x)K'(x,y) \leq A\pi(x)K(x,y) \quad \text{for all } x, y \in \mathcal{S} \text{ such that } x \neq y.$$

LEMMA 3.7 (Lemma 6 of [**29**]). *For any reversible finite Markov chain P,*

$$\mathrm{Gap}(P) \geq \frac{1}{m}\,\mathrm{Gap}(P^m) \quad \forall m \in \mathbb{N}^*.$$

LEMMA 3.8 (Lemma 7 of [**29**]). *Let A and B be Markov kernels. The following holds for A and B and also for A substituted by A's positive square root $A^{\frac{1}{2}}$:*

$$\mathrm{Gap}(ABA) \geq \mathrm{Gap}(B).$$

THEOREM 3.9 (Caracciolo-Pelissetto-Sokal [**29**]). *Let μ be a probability distribution on a finite state space $\mathcal{S}$, and let $\mathcal{P}$ be a transition matrix reversible with respect to μ. Suppose that we partition the set $\mathcal{S}$ as*

$$\mathcal{S} = \bigcup_{i=1}^{m} \mathcal{S}_i, \text{ with } \mathcal{S}_i \cap \mathcal{S}_j = \emptyset, \text{ if } i \neq j.$$

For each $i = 1, \ldots, m$, let $\mathcal{P}_i$ be the restriction of $\mathcal{P}$ to $\mathcal{S}_i$. Let $\mathcal{Q}$ be a positive operator, that is also reversible with respect to μ, and $\overline{\mathcal{Q}}$ the aggregated chain associated to the partition $(\mathcal{S}_i)_{i=1,\ldots,m}$: more precisely, for $i, j = 1, \ldots, m$,

$$\overline{\mathcal{Q}}(i,j) = \frac{1}{\mu(\mathcal{S}_i)} \sum_{x \in \mathcal{S}_i} \sum_{y \in \mathcal{S}_j} \mu(x)\mathcal{Q}(x,y).$$

Let $\mathcal{Q}^{\frac{1}{2}}$ be the positive square root of $\mathcal{Q}$. Then

$$\mathrm{Gap}(\mathcal{Q}^{\frac{1}{2}}\mathcal{P}\mathcal{Q}^{\frac{1}{2}}) \geq \mathrm{Gap}(\overline{\mathcal{Q}}) \cdot \min_{1 \leq i \leq m} \mathrm{Gap}(\mathcal{P}_i). \tag{26}$$

THEOREM 3.10 (Diaconis and Saloff-Coste [**9**]). *For $i = 1, ..., M$, 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 $\Omega_0 \times ... \times \Omega_M$, defined by*

$$P = \frac{1}{M+1} \sum_{i=0}^{M} I \otimes ... \otimes I \otimes P_i \otimes I \otimes ... \otimes I, \tag{27}$$

where (in a slight abuse of notation) I denotes the identity on the space it is defined. Then $\mathrm{Gap}(P) = \frac{1}{M+1} \min_{i \in \{0,\ldots,M\}} \{\mathrm{Gap}(P_i)\}$.

CHAPTER 4

The Generalized-Curie-Weiss model

This chapter will deal with the swapping algorithm on the Generalized—Curie—Weiss model. As we have seen in Section 2.4 the Curie-Weiss model's canonical Metropolis chain exhibits a bad cut in the state space at temperatures in the two macrostate region. Thus in general we can only expect torpid mixing of the Metropolis algorithm in a Generalized-Curie-Weiss setting. This chapter is organized as follows: In Section 1 we will define the Metropolis chain for the Generalized-Curie-Weiss setting. After this it is straightforward to define the swapping chain. Section 3 contains the result of this chapter while the proof is given in Section 4.

1. Defining the Metropolis chain

We will stay very close the the definition of the Metropolis chain in the Curie-Weiss setting given in 2.4. Note that we regain the Curie-Weiss model by taking $A := \{-1, 1\}$,

$$\rho = \frac{1}{2}(\delta_{-1} + \delta_{-1})$$

as probability measure on $[-1, 1]$ and

$$\begin{array}{rccc} g: & [-1,1] & \to & [0,1] \\ & x & \mapsto & x^2 \end{array}$$

as symmetric function which is strictly increasing on $[0, 1]$ and satisfies $g(0) = 0$. For the proposal chain we can take the almost identical proposal chain as in the Curie-Weiss case. The only difference we have to pay attention to is that the size $|A|$ of the set A is not necessarily 2. Define the proposal chain

$$K(\sigma, \tau) := \begin{cases} \frac{1}{2|A|N} & \text{if } \|\sigma - \tau\|_1 = 1 \\ 1 - \sum_{\tau' \neq \sigma} K(\sigma, \tau') & \text{if } \sigma = \tau \\ 0 & \text{otherwise} \end{cases}$$

and verify that K is positive, irreducible and aperiodic on $\Omega = A^\Lambda$. Let T then denote the transition matrix of the Metropolis chain for the desired Gibbs measure

$$\pi_\beta(\sigma) = \frac{e^{\beta N g\left(\frac{m(\sigma)}{N}\right)}}{\sum_\tau e^{\beta N g\left(\frac{m(\tau)}{N}\right)}}$$

of the Generalized-Curie-Weiss model.

2. Preparations

In this section, we will give a rapidly mixing Markov chain $(X_i)_i$ which has the uniform distribution of all states with a given total magnetization as its stationary distribution. This will be of use as we intend to compare the Metropolis algorithm on A_i (see equation (38)) of the Generalized-Curie-Weiss model with this chain in order to show rapid mixing.

Let $\Lambda = \{1, ..., N\}$,

$$A := \left\{ i\alpha \middle| i \in \{-n, .., n\} \right\} \quad \text{for an } \alpha > 0 \text{ and some } n \in \mathbb{N}$$

and define the state space to be $\Omega = A^\Lambda$. Now let

$$\mathcal{C} = \left\{ x \in \Omega \middle| m(x) = j \right\} \tag{28}$$

be the set of all states with total magnetization j and let ν be the uniform distribution on $\mathcal{C}$. Our aim is to give a Markov chain $(X_i)_i$ which compares well to the chain we consider later in Section 4.3 and which also samples efficiently from ν.

2.1. Rapid mixing of (X_i). Fix $\mathcal{C}$ as in (28). Consider the Markov chain (X_i) on $\mathcal{C}$ with the following transition kernel. Take $(R_1(i))_{i\in\mathbb{N}}$ and $(R_2(i))_{i\in\mathbb{N}}$ independent and uniformly distributed on $\{1, ..., N\}$ and $(U(i))_{i\in\mathbb{N}}$ independent and uniformly distributed on

$$\{1\alpha, \ldots, (2n+1)\alpha\}$$

such that $(R_1(i))$, $(R_2(i))$ and $(U(i))$ are independent. Define

$$\text{Move} : \mathcal{C} \times \{1, ..., N\}^2 \times \{1\alpha, ..., 2n\alpha\} \to \mathcal{C}$$

by

$$\text{Move}(x, r_1, r_2, u) := \begin{cases} x & \text{if } x_{r_1} - u < -n \\ & \text{ or } x_{r_2} + u > n \\ (x_1, ..., x_{r_1} - u, & \text{if } r_1 \le r_2 \\ ..., x_{r_2} + u, ..., x_N) & \text{ and not first case} \\ (x_1, ..., x_{r_2} + u, & \text{if } r_1 \ge r_2 \\ ..., x_{r_1} - u, ..., x_N) & \text{ and not first case} \end{cases} \tag{29}$$

and using this define

$$X_1 := X \in \mathcal{C}$$

$$X_{i+1} := \begin{cases} X_i & R_1(i) = R_2(i) \\ \text{Move}(X_i, R_1(i), R_2(i), U(i)) & R_1(i) \neq R_2(i) \end{cases} \tag{30}$$

(where X is any admissible starting point). Verify that (X_i) is irreducible, aperiodic and has reversible distribution ν on $\mathcal{C}$. We will use a

coupling argument in order to show rapid convergence to equilibrium of (X_i). To this end define

$$X_1' := X' \in \mathcal{C} \tag{31}$$

with X' drawn according to ν and

$$X_{i+1}' := \begin{cases} X_i' & R_1(i) = R_2(i) \\ X_i' & X_i(R_1(i)) \neq X_i'(R_1(i)) \wedge \\ & \quad X_i(R_2(i)) \neq X_i'(R_2(i)) \\ \mathrm{Move}(X_i', R_1(i), R_2(i), U(i)) & \text{else.} \end{cases}$$

Again verify that (X_i') is an irreducible and aperiodic Markov chain which is reversible with respect to ν on $\mathcal{C}$. Thus (X_i') is in equilibrium in every step.

LEMMA 4.1. *The expected coupling time $T_{\mathcal{C}}$ of the Markov chains (X_i) and (X_i') is bounded from above by*

$$\mathbb{E}T_{\mathcal{C}} \leq (2n+1)N^3.$$

PROOF. Define

$$\mathcal{C}(i) := \big\{k \in \{1, ..., N\} \big| X_i(k) \neq X_i'(k)\big\} \tag{32}$$

such that $\Psi(i) := |\mathcal{C}(i)|$ denotes the amount of components which have not coupled so far. Once $\Psi(i) = 0$, the two chains have coupled. Due to the construction, Ψ is monotonically decreasing. We now argue why

$$P\big(\Psi(i+1) \leq j-1 \big| \Psi(i) = j > 0\big) \geq \frac{1}{(2n+1)N^2} \tag{33}$$

holds.

First, assume there are exactly two components k_1 and k_2 satisfying $X_i(k_1) \neq X_i'(k_1)$ and $X_i(k_2) \neq X_i'(k_2)$. Without loss of generality assume further that $X_i(k_1) > X_i'(k_1)$ and $X_i(k_2) < X_i'(k_2)$. As $X_i, X_i' \in \mathcal{C}$ there exists a $u \in \{1\alpha, ..., (2n+1)\alpha\}$ such that

$$X_i(k_1) - u = X_i'(k_1) \text{ and } X_i(k_2) + u = X_i'(k_2).$$

All that needs to happen is drawing

$$\mathbb{P}\big(R_1(i) = k_1, R_2(i) = k_2, U(i) = u\big) = \frac{1}{(2n+1)N^2}.$$

Second, assume there are three or more components which have not coupled yet. There are either components k_1 and k_2 satisfying

$$X_i(k_1) - u = X_i'(k_1),\ X_i'(k_2) \neq X_i(k_2) \text{ and } X_i(k_2) + u \leq 2n\alpha$$

or satisfying

$$X_i(k_2) + u = X_i'(k_2),\ X_i'(k_1) \neq X_i(k_1) \text{ and } X_i(k_1) - u \geq -2n\alpha$$

for a suitable $u \in \{1\alpha, ..., 2n\alpha\}$. Again we know

$$\mathbb{P}\big(R_1(i) = k_1, R_2(i) = k_2, U(i) = u\big) = \frac{1}{(2n+1)N^2}$$

which proofs (33).

Using [**1**, Chapter 4-3, Lemma 1] we get an upper bound of

$$\mathbb{E}T_{\mathcal{C}} \leq \sum_{i=1}^{N}(2n+1)N^2 = (2n+1)N^3$$

for the coupling time. □

3. Result

Note the restrictions to g and ρ given in the definition for the Generalized-Curie-Weiss model in Section 1.1. Further assume that the finite set A is of the form

$$A \cap (0, \infty) = \big\{i\alpha \big| i \in \{1, .., n\}\big\} \quad \text{for an } \alpha > 0 \text{ and some } n \in \mathbb{N}. \tag{34}$$

Define

$$A_i := \{\sigma \in A_+ | m(\sigma) = i\alpha\} \quad \text{for } i \in \{0, ..., 2nN\} \tag{35}$$

to be the set of all states with total magnetization $i\alpha$ respectively. Here A_+ as defined in (36) consists of all the states we want to call positive. The main result of this chapter is

THEOREM 4.2. *Let g satisfy*

$$\pi_{\beta'}(A_i) \text{ is unimodal in } i \in \{0, ..., 2nN\}$$

for any $\beta' > 0$. Then, the swapping chain with transition kernel QPQ is rapidly mixing for the Generalized-Curie-Weiss model with parameters g and the normalized counting measure ρ on the finite and symmetric set $A \subset [-L, L]$ satisfying (34).

Remark This implies rapid convergence of the simulated tempering chain to equilibrium as the calculations given in Section 4.1 imply the condition necessary for using Theorem 3.1.

4. Proof

Fix $\beta > 0$. We will be using Theorem 3.2 as a means of showing rapid convergence to equilibrium in the situation of Theorem 4.2. We

start by partitioning

$$\begin{aligned}
\Omega &= A_+ \cup A_- \\
A_+ &= \Big\{\sigma \in \Omega \Big| \sum_{i=1}^N \sigma_i > 0\Big\} \cup \big\{\sigma \in \Omega \Big| \sum_{i=1}^N \sigma_i = 0, \sigma_{\text{argmin}_k\{\sigma_k \neq 0\}} > 0\Big\} \\
&\quad \cup \Big(\big\{(0,...,0)\big\} \cap A^\Lambda\Big) \\
A_- &= \Big\{\sigma \in \Omega \Big| \sum_{i=1}^N \sigma_i < 0\Big\} \cup \big\{\sigma \in \Omega \Big| \sum_{i=1}^N \sigma_i = 0, \sigma_{\text{argmin}_k\{\sigma_k \neq 0\}} < 0\Big\}
\end{aligned} \tag{36}$$

such that $A_+ \cap A_- = \emptyset$. Thus take $\mathcal{A} := \{A_+, A_-\}$. It is obviously clear, that for any $\beta' \in [0, \beta]$ there exists $c_1 > 0$ only dependent on β not on β' such that $Z_{\beta'}(N) \geq e^{c_1 N}$. This leads to

$$\pi_{\beta'}\big((0,...,0)\big) \leq e^{-c_1 N}$$

uniformly in β' such that the symmetry of g and ρ implies

$$\pi_\beta(A_i) \to \frac{1}{2} \quad \text{uniformly in } \beta' \in [0, \beta] \text{ and for } i \in \{+, -\}.$$

This implies $\gamma(\mathcal{A})$ to be bounded by a constant which can be chosen arbitrarily close to 1.

4.1. Bounding the overlap $\delta(\mathcal{A})$. Remember the definition of $\delta(\mathcal{A})$ given through equation (24). We first calculate

$$\begin{aligned}
\frac{Z_k(N)}{Z_{k+1}(N)} &= \frac{\sum_{\tau\in\Omega} e^{\beta N \frac{k}{M} g\left(\frac{m(\tau)}{N}\right)}}{\sum_{\tau\in\Omega} e^{\beta N \frac{k+1}{M} g\left(\frac{m(\tau)}{N}\right)}} \\
&= \frac{\sum_{\tau\in\Omega} e^{\beta N \frac{k}{M} g\left(\frac{m(\tau)}{M}\right)}}{\sum_{\tau\in\Omega} e^{\beta N \frac{k}{M} g\left(\frac{m(\tau)}{N}\right)} e^{\beta \frac{N}{M} g\left(\frac{m(\tau)}{N}\right)}} \\
&\leq \frac{Z_k(N)}{Z_k(N)} \\
&= 1
\end{aligned}$$

and

$$\begin{aligned}\frac{Z_k(N)}{Z_{k+1}(N)} &= \frac{\sum_{\tau\in\Omega} e^{\beta N \frac{k}{M} g\left(\frac{m(\tau)}{N}\right)}}{\sum_{\tau\in\Omega} e^{\beta N \frac{k+1}{M} g\left(\frac{m(\tau)}{N}\right)}} \\ &= \frac{\sum_{\tau\in\Omega} e^{\beta N \frac{k}{M} g\left(\frac{m(\tau)}{M}\right)}}{\sum_{\tau\in\Omega} e^{\beta N \frac{k}{M} g\left(\frac{m(\tau)}{N}\right)} e^{\beta g\left(\frac{m(\tau)}{N}\right)}} \\ &\geq \frac{\sum_{\tau\in\Omega} e^{\beta N \frac{k}{M} g\left(\frac{m(\tau)}{M}\right)}}{\sum_{\tau\in\Omega} e^{\beta N \frac{k}{M} g\left(\frac{m(\tau)}{N}\right)}} e^{-\beta \frac{N}{M}\|g\|_\infty} \\ &= e^{-\beta \frac{N}{M}\|g\|_\infty}.\end{aligned}$$

We can use this to gain a two-sided bound on

$$\begin{aligned}\frac{\pi_{k+1}(\sigma)}{\pi_k(\sigma)} &= e^{\beta N\left(\frac{k+1}{M}-\frac{k}{M}\right) g\left(\frac{m(\sigma)}{N}\right)} \frac{Z_k(N)}{Z_{k+1}(N)} \\ &= e^{\beta N \frac{N}{M} g\left(\frac{m(\sigma)}{N}\right)} \frac{Z_k(N)}{Z_{k+1}(N)} \\ &\in [1, e^{\beta \frac{N}{M}\|g\|_\infty}] \frac{Z_k(N)}{Z_{k+1}(N)} \\ &\subseteq [e^{-\beta \frac{N}{M}\|g\|_\infty}, e^{\beta \frac{N}{M}\|g\|_\infty}].\end{aligned}$$

This leads to $\delta(\mathcal{A})$ being upper and lower bounded by constants. Note that we only need the lower bound on $\delta(\mathcal{A})$ for using Theorem 3.2.

4.2. Bounding $\mathrm{Gap}(\overline{T_0})$**.** Again take $A_i = \{\sigma \in A_+ | m(\sigma) = i\alpha\}$ to be the set of all states with given total magnetization $i\alpha$ in Ω. We want to show that a transition from A_{-1} (or $A_0 \cap A_-$) to A_1 is probable. To achieve this, we first need to know that being in a state in A_{-1} is probable. To this end define

$$c_2(\sigma) := \min\left\{k' \Big| \sum_{k=1}^{k'} \sigma_k < 0\right\}$$

to be the first component where the partial sum of all prior spins turns negative. Note that c_2 is well defined for any $x \in A_-$. Define

$$h(\sigma) = \sigma'$$

to be the function which increases this component by 1α, such that $\sigma_{c_1(\sigma)} + 1\alpha = \sigma'_{\sigma_1(\sigma)}$ and $\sigma'_k = \sigma_k$ for all other components. Note that $h(A_i) \subseteq A_{i+1}$ is an injective mapping for all $i < 0$. This implies

$$\text{either} \quad \pi_0(A_1) \geq \frac{1}{(2n+1)N} \quad \text{or} \quad \pi_0(A_0 \cap A_-) \geq \frac{1}{(2n+1)N}. \tag{37}$$

PROPOSITION 4.3. *$\overline{T_0}$ is rapidly mixing.*

PROOF. Take B to denote either A_{-1} or $A_0 \cap A_-$ depending of which one fulfills the inequality in (37). Each $\sigma \in B$ has at least one component of which the spin value can be increased far enough, such that the resulting σ' has $m(\sigma') > 0$. Thus $\sigma' \in A_+$. Using this short notation and the definition of $\overline{T_0}$ we calculate

$$\begin{aligned}
\overline{T_0}(-,+) &= \frac{1}{\pi_0(A_-)} \sum_{\sigma \in A_-} \sum_{\tau \in A_+} \pi_0(\sigma) T_0(\sigma,\tau) \\
&\geq 2 \sum_{\sigma \in B} \sum_{\tau \in A_+} \pi_0(\sigma) T_0(\sigma,\tau) \\
&\geq 2 \sum_{\sigma \in B} \pi_0(\sigma) T_0(\sigma,\sigma') \\
&\geq \frac{1}{|A|N} \pi_0(B) \\
&\geq \frac{1}{2|A|(n+1)N^2} \\
&= \frac{1}{(2n+1)^2 N^2}
\end{aligned}$$

where we use that $T_0(\sigma,\sigma') = K(\sigma,\sigma') = (2|A|N)^{-1}$. Note that $\overline{T_0}(+,-)$ can be bounded by the same technique and therefore by the same bound. Comparing this with the Markov chain which tosses a coin independently in every step, taking the displayed value, by using Lemma 3.6 yields the desired result. □

4.3. Bounding $\min_{k,j} \operatorname{Gap}(T_k|_{A_j})$. We will focus on

$$\min_k \operatorname{Gap}(T_k|_{A_+})$$

as π_β is symmetric and A_+ and A_- differ only insignificantly. All bounds shown work for both cases. Fix an arbitrary $k \in \{1,...,M\}$. We will see that $T_k|_{A_+}$ is rapidly mixing, thus giving a polynomial lower bound for $\min_k \operatorname{Gap}(T_k|_{A_+})$. This bound is independent of β_k. For an easier notation denote

$$P := T_k|_{A_+}.$$

Partition

$$A_+ = \bigcup_{i=0}^{2nN} A_i \tag{38}$$

with $A_i = \{\sigma \in A_+ | m(\sigma) = i\alpha\}$ according to the total magnetization. Using Lemma 3.7 on P we gain

$$\operatorname{Gap}(P) \geq \frac{1}{3} \operatorname{Gap}(P^3) = \frac{1}{3} \operatorname{Gap}(P^{\frac{1}{2}} P^2 P^{\frac{1}{2}}).$$

This, on first sight counterproductive, step makes sure that the chains induced on the decomposition state space are nontrivial. Only considering $P|_{A_i}$ would lead to a constant chain, as the chain cannot leave any state. The chain $P^2|_{A_i}$ does not have this disadvantage as P^2 can increase/decrease the total magnetization by some amount in the first step and reverse this change in magnetization by the second step. Using Theorem 3.9 we get

$$\mathrm{Gap}(P^3) = \mathrm{Gap}(P^{\frac{1}{2}}P^2P^{\frac{1}{2}}) \geq \mathrm{Gap}(\overline{P}) \min_{1\leq i\leq N|A_+|} \mathrm{Gap}(P_i^2)$$

with $P_i^2 := P^2|_{A_i}$.

PROPOSITION 4.4. *$\overline{P}$ is rapidly mixing.*

PROOF. Note the condition in Theorem 4.2 of $\pi_{\beta'}(A_i)$ being unimodal in i for any $\beta' > 0$. We will be using this in conjunction with the Poincaré-Inequality of Lemma 3.4 in order to see that $\mathrm{Gap}(\overline{P})$ can be bounded below by the inverse of a polynomial. First note for any $i \in \{0, .., nN-1\}$ we get

$$\begin{aligned}
\overline{P}(i, i+1) &= \frac{1}{\pi_{\beta'}(A_i)} \sum_{\sigma\in A_i} \sum_{\tau\in A_{i+1}} \pi_{\beta'}(\sigma)P(\sigma,\tau) \\
&\geq \frac{1}{\pi_{\beta'}(A_i)} \sum_{\sigma\in A_i} \pi_{\beta'}(\sigma) \frac{1}{2|A|N} \\
&= \frac{1}{2|A|N}
\end{aligned}$$

and using reversibility in (39)

$$\begin{aligned}
\overline{P}(i+1, i) &= \frac{1}{\pi_{\beta'}(A_{i+1})} \sum_{\sigma\in A_{i+1}} \sum_{\tau\in A_i} \pi_{\beta'}(\sigma)P(\sigma,\tau) \\
&= \frac{1}{\pi_{\beta'}(A_{i+1})} \sum_{\sigma\in A_{i+1}} \sum_{\tau\in A_i} \pi_{\beta'}(\tau)P(\tau,\sigma) \qquad (39) \\
&\geq \frac{1}{\pi_{\beta'}(A_{i+1})} \sum_{\tau\in A_i} \pi_{\beta'}(\tau) \frac{1}{2|A|N} \\
&= \frac{1}{2|A|N} \frac{\pi_{\beta'}(A_i)}{\pi_{\beta'}(A_{i+1})}.
\end{aligned}$$

Assume the worst case scenario of $\frac{\pi_{\beta'}(A_i)}{\pi_{\beta'}(A_{i+1})} < 1$. We will use this for bounding K in Lemma 3.4 by

$$\begin{aligned}
K &= \max_{\substack{i,j\in\{0,\dots,nN\}:\\ |i-j|=1}} \frac{1}{\pi_{\beta'}(A_i)\overline{P}(i,j)} \sum_{\gamma_{z_1,z_2}\ni\langle i,j\rangle} |\gamma_{z_1,z_2}|\pi_{\beta'}(z_1)\pi_{\beta'}(z_2) \\
&\leq nN \max_{\substack{i,j\in\{0,\dots,nN\}:\\ |i-j|=1}} \sum_{\gamma_{z_1,z_2}\ni\langle i,j\rangle} \frac{\pi_{\beta'}(z_1)\pi_{\beta'}(z_2)}{\pi_{\beta'}(A_i)\overline{P}(i,j)} \\
&\leq nN \max_{i\in\{0,\dots,2n-1\}} \sum_{\gamma_{z_1,z_2}\ni\langle i+1,i\rangle} \frac{\pi_{\beta'}(z_1)\pi_{\beta'}(z_2)}{\pi_{\beta'}(A_{i+1})\overline{P}(i+1,i)} \\
&\leq nN \max_{i\in\{0,\dots,2n-1\}} \sum_{\gamma_{z_1,z_2}\ni\langle i+1,i\rangle} \frac{\pi_{\beta'}(z_1)}{\pi_{\beta'}(A_{i+1})}\frac{\pi_{\beta'}(z_2)}{\pi_{\beta'}(A_i)}\pi_{\beta'}(A_{i+1}) \\
&\leq (nN)^3
\end{aligned}$$

as every factor in the last product is bounded by 1 due to the unimodality condition. □

PROPOSITION 4.5. *There exists a polynomial $p(N)$ such that*

$$\mathrm{Gap}(P_i^2) \geq p(N)^{-1}$$

thus P_i^2 is rapidly mixing for any $i \in \{1,...,nN\}$.

Remark Note that the technique used in the proof of Lemma 4.1 also works for the case of restricting the state space to $A_0 \cap A_+$. We refrain from showing this explicitly to prevent unnecessarily complicated notation.

PROOF. It is necessary to consider two cases. The first is P_{nN}^2 in which A_{nN} only contains one state. Thus P_{nN}^2 is the constant chain and therefore rapidly mixing.

The second case is the case of P_i^2 with $i < nN$. We will compare P_i^2 with the Markov chain given in Section 2 which has been shown to be rapidly mixing in Lemma 4.1. Consider two states $\sigma \neq \tau$ with $m(\sigma) = m(\tau) = i\alpha$ and

$$\mathbb{P}(X_{j+1} = \tau | X_j = \sigma) = 0.$$

This directly implies $P_i^2(\sigma,\tau) = 0$, as all P_i^2 can do is select two components and transfer magnetization between these two, just as the Markov chain (X_j) does too. Now consider the other case of two states $\sigma \neq \tau$ with $m(\sigma) = m(\tau) = i\alpha$ and

$$\mathbb{P}(X_{j+1} = \tau | X_j = \sigma) = \frac{1}{2nN^2} = \frac{1}{|A|N^2} > 0.$$

Think of the transition of X_i to X_{i+1} as consisting of two steps. First increasing one component's amount, thus reaching state τ' with $m(\tau') >$

$i\alpha$ and then decreasing another component and therefore reaching τ. We know

$$P_i^2(\sigma,\tau) \geq P(\sigma,\tau')P(\tau',\tau)$$

and due to the definition of the Metropolis sampler it is either $P(\sigma,\tau) = \frac{1}{2|A|N}$ or $P(\tau,\sigma) = \frac{1}{2|A|N}$. Assume the first equality to hold true. Then we gain

$$\begin{aligned} P(\tau',\tau) &= \frac{1}{2|A|N} \min\left\{1, e^{\beta_k N\left(g\left(\frac{m(\tau)}{N}\right)-g\left(\frac{m(\tau')}{N}\right)\right)}\right\} \\ &= \frac{1}{2|A|N} \min\left\{1, e^{\beta \frac{N}{M}\left(g\left(\frac{m(\tau)}{N}\right)-g\left(\frac{m(\tau')}{N}\right)\right)}\right\} \\ &\geq \frac{1}{2|A|N} e^{-\beta \frac{N}{M}\|g\|}. \end{aligned}$$

This argument would work for bounding $P(\sigma,\tau')$ just as well, so altogether we get

$$P_i^2(\sigma,\tau) \geq \left(\frac{1}{2|A|N}\right)^2 e^{-\beta \frac{N}{M}\|g\|}. \tag{40}$$

From Lemma 4.1 we can deduce $\mathrm{Gap}((X_i)) \geq \frac{1}{6|A|N^3}$ such that Lemma 3.6 yields the claim. Note that this bound does not depend on i. □

CHAPTER 5

The Blume-Emery-Grifiths model

In this chapter we will analyze the Swapping algorithm on the mean-field version of the Blume-Emery-Griffiths model which is given in Chapter 1 Section 3. This model has two parameters and depending on their choice, the model exhibits either a first or a second order phase transition. In agreement with a conjecture by Bhatnagar and Randall we find that the Swapping algorithm mixes rapidly in presence of a second order phase transition, while becoming slow when the phase transition is first order.

As mentioned in Chapter 3 Section 3.3, Woodard, Schmidler and Huber are able to give the first known result on rapid mixing of the Swapping algorithm in a general, non model-specific, setting. It is noticeable that their result cannot be used in the case of rapid mixing in the BEG model. The technique used by Woodard, Schmidler and Huber relies heavily on a static, non temperature-dependent, partitioning of the state space. The underlying Metropolis chain needs to mix rapidly in each part and for any temperature in order for their technique to work. Furthermore, the probability of each part must not get too small, as the temperature is decreased. In the rapid mixing case of the BEG model, this partitioning cannot be achieved. Our proof relies on a dynamic, temperature dependent, partitioning in which one part gets very unlikely as the temperature is decreased.

This chapter is organized as follows: Section 1 contains some technical results which we need in order to prove our theorems. More precisely, we will propose a way to rewrite the BEG model and we will prove a result on the speed of convergence of a coloring algorithm on a graph. This will be useful in the proofs of our results in Section 3. Section 2 is devoted to our results – a characterization of the parameter regimes where the Swapping algorithm converges rapidly or slowly, respectively. These results will be proofed in Section 3. Next to the mentioned free energy bounds, the proofs use methods to bound the spectral gaps of Markov chains such as coupling methods or Poincaré inequalities. In the appendix a collection of results on the free energy in the BEG model is given, which contains refinements of some results given in [**14**].

1. Technical preparations

We will first do some system specific preparations, in order to get more familiar with the model and as evidence why swapping and tempering should be considered for this model.

1.1. BEG-specific preparations. In favor of an easy notation define the functions

$$S_N(\sigma) = \sum_{i=1}^{N} \sigma_i \tag{41}$$

$$R_N(\sigma) = \sum_{i=1}^{N} \sigma_i^2 \tag{42}$$

where S_N gives the total magnetization, and R_N the total amount of non-zero spins of the state σ. Using this, it is easy to define

$$\mathcal{A}_{s,r} := \big\{\sigma \in \Omega \big| S_N(\sigma) = s, R_N(\sigma) = r\big\} \tag{43}$$

as the set of states with fixed amount of 0s and fixed magnetization. As we consider the mean field BEG model, all states in $\mathcal{A}_{s,r}$ are basically indistinguishable in the system. We will later (Theorem 5.12) see, that the Metropolis chain T^2 restricted to $\mathcal{A}_{s,r}$ mixes rapidly for any combination of s and r.

In order to be able to better address non-negligible differences in the state space consider

$$\Upsilon = \Upsilon_N := \big\{\mathbf{a} = (a_{-1}, a_0, a_1) \in \mathbb{R}^3 \big| a_i \geq 0 \ \forall i, \sum_i a_i = 1, Na_i \in \mathbb{N} \ \forall i = -1, 0, 1\big\} \tag{44}$$

such that

$$\Omega = \bigcup_{\mathbf{a}\in\Upsilon} \Big\{\sigma \in \Omega \Big| \sum_{j=1}^{N} \delta_{\sigma_j, i} = Na_i \ \forall i \in \{-1, 0, 1\}\Big\} \tag{45}$$

is a disjoint union. This is inspired by Gore's and Jerrum's work on the Potts Model **[21]** as the following calculation makes the state space easier to handle.

Considering

$$\begin{aligned}\pi_\beta(\sigma \text{ has type } N\mathbf{a}) &= \binom{N}{Na_{-1}, Na_0, Na_1} Z(\beta)^{-1} \\ &\quad \times e^{-\beta\left(Na_{-1} + Na_1 - \frac{K}{N}(Na_1 - Na_{-1})^2\right)} \\ &= \binom{N}{Na_{-1}, Na_0, Na_1} Z(\beta)^{-1} \\ &\quad \times e^{-N\beta\left(a_{-1} + a_1 - K(a_1 - a_{-1})^2\right)}\end{aligned} \tag{46}$$

and using Stirling's approximation one obtains

$$\begin{aligned}\pi_\beta(\sigma \text{ has type } N\mathbf{a}) &= Z(\beta)^{-1}N^{-1}e^{-N\left(\sum_i a_i\log a_i\right)+\Delta(\mathbf{a})}\\ &\quad\times e^{-N\beta\left(a_{-1}+a_1-K(a_1-a_{-1})^2\right)}\\ &= Z(\beta)^{-1}N^{-1}\\ &\quad\times e^{N\left(\beta\left(-a_{-1}-a_1+K(a_1-a_{-1})^2\right)-\sum_i a_i\log a_i\right)+\Delta(\mathbf{a})} \qquad (47)\end{aligned}$$

with $|\Delta(\mathbf{a})| = O(1)$ if there exists an $\varepsilon > 0$ with $a_i \geq \varepsilon$ for all $i \in \{-1, 0, 1\}$. So understanding

$$f_\beta(\mathbf{a}) := \beta\big(-a_{-1}-a_1+K(a_1-a_{-1})^2\big) - \sum_i a_i\log a_i \qquad (48)$$

will give us better intelligence of how the BEG model behaves depending on β. See the appendix for the details. The rough description (that is, of course, in agreement with the findings of Ellis et al. in [**14**]) is, that for small K and small $\beta > 0$ the free energy is unimodal, while for small enough K and large β there are three minima. For large enough K, f_β is bimodal.

1.1.1. *Why Metropolis is torpidly mixing for BEG.* We know, as seen in Section 1.1, that $\pi_\beta(\sigma$ has type $N\mathbf{a})$ has exponential structure for any $\mathbf{a}$ which is a local maximum. We also know, due to the analysis sketched above, that f_β has three local modes for suitable K and sufficiently (depending on K) large β. Take $\mathbf{a}$ to represent the lowest local maximum point. This leads to $B_\varepsilon(\mathbf{a})$ having exponential little conductance, therefore representing a bad cut in the state space. Here and in the following $B_\varepsilon(\mathbf{a})$ will always denote a ball of radius ε centered in $\mathbf{a}$ in the appropriate metric space. For more details see Section 3.4 where this technique is used in the more complicated setup of swapping.

1.2. Random 3-coloring of the complete graph. In this subsection, we will give a rapidly mixing Markov chain $(X_i)_i$ which has the uniform distribution on the set of of all 3-Colorings with given amounts of vertices of a certain color as its stationary distribution. This will be of use, as we intend to compare the Metropolis algorithm on $\mathcal{A}_{s,r}$ (see (43)) of the BEG model with this chain in order to show rapid mixing.

Let $\Lambda = \{1, ..., N\}$ and define $\Omega = \{-1, 0, 1\}^\Lambda$ to be the set of all possible 3-colorings of Λ. Note, that we do not restrict ourselves to 3-colorings in the graph theoretic sense, where adjacent vertices are required to have different colors. Further consider a tuple $(a_1, a_2, a_3) \in \Upsilon$. Na_i represents the amount of vertexes, which have color i. Now let

$$\mathcal{C} = \Big\{\sigma \in \Omega \Big| \frac{1}{N}\sum_j \delta_{i,\sigma_j} = a_i\Big\} \qquad (49)$$

be the set of appropriate 3-colorings and ρ the uniform distribution on $\mathcal{C}$. Our aim is to give a Markov chain $(X_i)_i$ which compares well to the

chain we will consider later in Section 3.3.2 for the BEG model and which also samples efficiently from ρ.

1.2.1. *Rapid mixing of* (X_i). Fix $\mathcal{C}$ as in (49). Consider the Markov chain (X_i) on $\mathcal{C}$ with the following transition kernel. Take $(R_1(i))_{i\in\mathbb{N}}$ and $(R_2(i))_{i\in\mathbb{N}}$ independently and uniformly distributed on $\{1,...,N\}$. Define

$$X_1 := X \in \mathcal{C}$$

$$X_{i+1} := \begin{cases} X_i & R_1(i) = R_2(i) \\ \big(R_1(i), R_2(i)\big)\big(X_i\big) & R_1(i) \neq R_2(i) \end{cases} \tag{50}$$

(where X is any admissible starting point and for a vector $x := (x_1, \ldots, x_N)$ and $i \neq j \in \{1,\ldots,N\}$ we write $(i,j)(x_1,\ldots x_N)$ for the vector x with the components i and j interchanged) and verify, that (X_i) has reversible distribution π on $\mathcal{C}$. We will use a coupling argument in order to show rapid convergence to equilibrium of (X_i). To this end define

$$X_1' := X' \in \mathcal{C} \tag{51}$$

with X' drawn according to ρ and iteratively

$$\mathcal{C}(i) := \big\{ j \in \{1,...,N\} \big| X_i(j) \neq X_i'(j) \big\} \tag{52}$$

with

$$X_{i+1}' := \begin{cases} X_i' & R_1(i) = R_2(i) \\ \big(R_1(i), R_2(i)\big)\big(X_i'\big) & X_i(R_1(i)) = X_i'(R_1(i)) \wedge X_i(R_2(i)) \\ & \neq X_i'(R_2(i)) \\ \big(R_1(i), R_2(i)\big)\big(X_i'\big) & X_i(R_1(i)) \neq X_i'(R_1(i)) \wedge X_i(R_2(i)) \\ & = X_i'(R_2(i)) \\ \big(R_1(i), R_2(i)\big)\big(X_i'\big) & X_i(R_1(i)) = X_i'(R_1(i)) \wedge X_i(R_2(i)) \\ & = X_i'(R_2(i)) \\ \big(R_1(i), R_3(i)\big)\big(X_i'\big) & \text{otherwise} \end{cases}$$

and R_3 being uniformly drawn out of $\mathcal{C}(i)$ and independent of $(R_1(i))$ and $(R_2(i))$. Again verify that (X_i') is a Markov chain which is reversible with respect to ρ on $\mathcal{C}$. Thus (X_i') is in equilibrium in every step.

LEMMA 5.1. *The expected coupling time $T_{\mathcal{C}}$ of the Markov chains (X_i) and (X_i') is bounded from above by*

$$\mathbb{E}T_{\mathcal{C}} \leq N^4.$$

PROOF. Define $\Psi(i) := |\mathcal{C}(i)|$. Once $\Psi(i) = 0$ the two chains have coupled. Due to the construction Ψ is monotonically decreasing. Indeed, if $X_i(k) = X_i'(k)$ holds for one i and a $k \in \{1,\ldots,N\}$, we will

have $X_j = X'_j$ for the position k is permuted to. We further know

$$\mathbb{P}\big(\Psi(i+1) \leq j-1 \big| \Psi(i) = j > 0\big) \geq \frac{1}{N^3}$$

as all that needs to happen is, find two components k_1 and k_2 with

$$X_i(k_1) = X'_i(k_2) \neq X_i(k_2) = X'_i(k_1)$$

and choose these with R_1 and R_2 which happens with probability $\frac{1}{N^2}$. In this case R_3 would be drawn out of all components in which X_i and X'_i differ. There are at most N of those. Using [**1**, Chapter 4-3, Lemma 1] we get an upper bound of

$$\mathbb{E}T_{\mathcal{C}} \leq \sum_{i=1}^{N} N^3 = N^4$$

for the coupling time. □

2. Results

In this section we will summarize our results which will be proved in Section 3. We will first define what chain exactly we want to look at.

The Simulated Tempering algorithm and the Swapping algorithm are defined in Chapter 3 Section 1 and 2 respectively. In the two cases, for the BEG model, the corresponding Metropolis-Hastings chain for the measure π_β, defined through (9), is given by (3), with the proposal chain

$$K_{gen}(x,y) = \frac{1}{4N},$$

if $x, y \in \{-1,0,1\}^N$ and differ in exactly one spin $x_i \neq y_i$, for some $i \in \{1,...,N\}$, and $K_{gen}(x,x) = \frac{1}{2}$. In all other cases define

$$K_{gen}(x,y) = 0.$$

The BEG Model, as Ellis et al. [**14**] show, exhibits different phase behavior depending on K. For small $K < K_{\text{low}}$ there is, for every temperature, only one macro state, which implies that there is no phase transition. Ellis et al. conjecture K_{low} to be 1, but they do not give a proof for this.

The first regime we want to look at is $K_{\text{low}} < K < K_c$ with $K_c = K(\log 4)$ as in [**14**, Eq. (3.19)]. The model exhibits a discontinuous phase transition at a $\beta_c(K)$ depending on K. We will use this discontinuity in the phase to show

THEOREM 5.2. *Consider the BEG model with $K_{low} < K < K_c$. The Simulated Tempering algorithm is torpidly mixing, since*

$$\text{Gap}(QP_{st}Q) \leq e^{-cN}$$

holds for $c > 0$ as constructed in Theorem 5.13.

COROLLARY 5.3. *This implies torpid mixing of the Swapping algorithm in this regime.*

For $K > K_c$ the model shows a continuous phase transition at $\beta_c(K)$ which will lead to a Swapping chain which behaves very much like a Curie-Weiß model's Swapping chain which Madras and Zheng already considered in [**29**]. This leads to

THEOREM 5.4. *For $K > K_c$ the Swapping chain with its transition kernel QPQ for the BEG model satisfies*

$$\mathrm{Gap}(QPQ) \geq \frac{1}{p(N)}$$

for some polynomial p of N.

Remark Giving an explicit bound would need a longer argument in the end of the proof of Theorem 5.10 which does not give a better insight of the situation. As we do not believe our technique to give a sharp bound anyway, we refrain from doing this extra step and do not give a suitable polynomial explicitly.

COROLLARY 5.5. *This implies rapid mixing of the Simulated Tempering chain $QP_{st}Q$ in this regime.*

3. Proofs

In this section we will first prove Theorem 5.4 to conclude this paper by showing Theorem 5.2.

3.1. General partitioning of the state space in the case of $K > K_c$. We will begin to show Theorem 5.4 by partitioning the state space

$$\Omega = \Omega_+ \cup \Omega_- \tag{53}$$

into two disjoint almost equally large parts

$$\Omega_+ = \Big\{x \in \Omega \Big| \sum_i x(i) > 0\Big\} \cup \{(0,...,0)\}$$
$$\cup \Big\{x \neq (0,...,0) \Big| \sum_i x_i = 0, \text{ with the first non-zero coordinate } = +1\Big\}$$
$$\Omega_- = \{x \in \Omega | \sum x(i) < 0\}$$
$$\cup \Big\{x \neq (0,...,0) \Big| \sum_i x_i = 0, \text{ with the first non-zero coordinate } = \text{-}1\Big\}.$$

Using this partitioning we will decompose $\Omega^{\mathrm{sw}} = \Omega^{M+1}$ in the same way as Madras and Zheng in [**29**, Section 4, Step two].

Let $\widetilde{\Omega^{\mathrm{sw}}} := \{+,-\}^M$ and take $x \in \Omega^{\mathrm{sw}}$. Define the signature of x by

$$\begin{array}{rccc} \mathrm{sgn} & : & \Omega^{\mathrm{sw}} & \to & \widetilde{\Omega^{\mathrm{sw}}} \\ & & x & \mapsto & v \end{array} \tag{54}$$

with

$$v_i = \begin{cases} + & \text{if } x_{i+1} \in \Omega_+ \\ - & \text{if } x_{i+1} \in \Omega_-, \end{cases} \tag{55}$$

such that $\mathrm{sgn}(x)$ contains the sign, of the total magnetization of each component of x except of the component for $\beta = 0$. The first component of x will have a special role, which will become apparent within the next paragraphs.

We will decompose the state space using the amount of $+$-signs in $\mathrm{sgn}(x)$. For fixed $k \in \{0, ..., M\}$ define

$$\widetilde{\Omega}_k := \{v \in \widetilde{\Omega^{\mathrm{sw}}} | v \text{ has exactly } k \ +\text{-signs}\}. \tag{56}$$

and note, that

$$\Omega^{\mathrm{sw}} = \bigcup_{k=0}^{M} \Omega_k$$

is a disjoint union of

$$\Omega_k := \{x \in \Omega^{\mathrm{sw}} | \, \mathrm{sgn}(x) \in \widetilde{\Omega}_k\}. \tag{57}$$

Define $\overline{Q}$ to be the aggregated transition matrix as described in Theorem 3.9 for this decomposition. Using Lemma 3.8 and Theorem 3.9 we get

$$\mathrm{Gap}(QPQ) \geq \mathrm{Gap}(Q^{\frac{1}{2}}(QPQ)Q^{\frac{1}{2}}) \tag{58}$$

$$\geq \mathrm{Gap}(\overline{Q}) \cdot \min_{k \in \{0,...,M\}} \mathrm{Gap}((QPQ)|_{\Omega_k}). \tag{59}$$

Citing [**29**, Sec. 4, step three], we can do all displayed calculations in our setting as well, which eventually leads to

$$\mathrm{Gap}(Q^{\frac{1}{2}}(QPQ)Q^{\frac{1}{2}}) \geq \frac{1}{8} \mathrm{Gap}(\overline{Q}) \cdot \min_{k \in \{0,...,M\}} \mathrm{Gap}((Q_k P_k Q_k)) \tag{60}$$

with P_k and Q_k being the restrictions of P and Q to Ω_k, respectively.

The transition kernel $\overline{Q}$ is, in this setting, responsible for changing the amount of components in $x \in \Omega^{\mathrm{sw}}$ which are in Ω_+ and Ω_- respectively. $\overline{Q}$ is essentially a one dimensional nearest neighbor random walk on $\{0, ..., M\}$ whose spectral gap is well understood. Due to the symmetry in the model it does not (noticeably) matter for the chain, whether we restrict a given component k of x to be in Ω_+ or Ω_-. This leads to

$$\mathrm{Gap}((Q_k P_k Q_k)) \approx \mathrm{Gap}((Q_{k'} P_{k'} Q_{k'})) \quad \forall k, k' \in \{0, ..., M\} \tag{61}$$

where $\approx$ means that both spectral gaps are of the same (polynomial or exponential) order. This in turn implies

$$\min_{k\in\{0,\dots,M\}} \mathrm{Gap}((Q_k P_k Q_k)) \approx \mathrm{Gap}((Q_M P_M Q_M)).$$

We will, by abuse of notation, write this as

$$\min_{k\in\{0,\dots,M\}} \mathrm{Gap}((Q_k P_k Q_k)) \approx \mathrm{Gap}((Q_M P_M Q_M)) = \mathrm{Gap}((Q P_+ Q)) \tag{62}$$

and note, that all arguments of the proof work in exactly the same way for any $k \in \{0, ..., M\}$. The only difference is, at what part of the state space we look at, for a given temperature β_i. $\mathrm{Gap}(\overline{Q})$ and $\mathrm{Gap}(QP_+Q)$ will be bounded below in the following subsections.

3.2. Speed of convergence of $\overline{Q}$. Following in principle the proof given in [**29**, Section 5] (also see [**34**, Section 2.5] for more details) we gain

LEMMA 5.6. *The spectral gap of the aggregated chain $\overline{Q}$ satisfies*

$$\mathrm{Gap}(\overline{Q}) \geq \frac{1}{4M^2} e^{-\beta(K+1)\frac{N}{M}}.$$

Remark Note the notation: The notation for the amount of spins N and the amount of temperatures M considered are interchanged between this paper and the reference given above. On the other hand, the notation now agrees with the standard notation in statistical mechanics.

PROOF. We first verify that the probability for an accepted swapping move is bounded below by a constant. Using the notation given in [**29**] let us define

$$\rho_{i,i+1} := \min\left(1, \frac{\pi_i(x_{i+1})\pi_{i+1}(x_i)}{\pi_i(x_i)\pi_{i+1}(x_{i+1})}\right).$$

Then

$$\begin{aligned}
\rho_{i,i+1} &= \min\left(1, \frac{e^{\beta_{i+1}H(x_i)}e^{\beta_i H(x_{i+1})}}{e^{\beta_i H(x_i)}e^{\beta_{i+1}H(x_{i+1})}}\right) \\
&= \min\left(1, e^{\beta_{i+1}H(x_i)+\beta_i H(x_{i+1})-\beta_i H(x_i)-\beta_{i+1}H(x_{i+1})}\right) \\
&= \min\left(1, e^{\beta\frac{i+1}{M}H(x_i)+\beta\frac{i}{M}H(x_{i+1})-\beta\frac{i}{M}H(x_i)-\beta\frac{i+1}{M}H(x_{i+1})}\right) \\
&= \min\left(1, e^{\beta\frac{H(x_i)}{M}-\beta\frac{H(x_{i+1})}{M}}\right) \\
&\geq e^{-\beta\frac{H(x_{i+1})}{M}} \\
&\geq e^{-\beta\frac{N(K+1)}{M}} \qquad (63)
\end{aligned}$$

as $H \leq (K+1)N$ implies (63) to be true.

Due to the definition of Ω_+ and Ω_- it is clear, that $\pi_\beta(\Omega_+) = \frac{1}{2}(1 + 1/Z_\beta)$ for any $\beta \geq 0$. Recalling equations (46) and (48) and

Theorem .9 given in the appendix it is possible to find for any $\beta > 0$ constants $0 < c_1 < c_2$ such that $Z_{\beta'} \in [e^{c_1 N}, e^{c_2 N}]$ for all $\beta' \in [0, \beta]$. Using

$$1 \leq \left(1 + e^{-cN}\right)^M \leq e^{e^{-cN} M} \longrightarrow 1 \tag{64}$$

as $N \to \infty$, we gain a constant $a > 1$ such that for all sufficiently large N and any $\sigma \in \{-, +\}^M$

$$\pi(\Omega \times \Omega_{\sigma_1} \times \cdots \times \Omega_{\sigma_M}) \in 2^{-M}[a^{-1}, a^1]$$

holds. Recalling the definition of Ω_k in (57) we conclude

$$\pi(\Omega_k) = \sum_{\sigma \in \widetilde{\Omega}_k} \pi(\Omega \times \Omega_{\sigma_1} \times \cdots \times \Omega_{\sigma_M}) \in \binom{M}{k} \left(\frac{1}{2}\right)^M [a^{-1}, a]. \tag{65}$$

As we want to use Lemma 3.6 later on, in order to compare $\overline{Q}$ to an easier Markov chain, it is of interest to study the quantity

$$\pi(\Omega_i)\overline{Q}(i, i+1). \tag{66}$$

Consider an $x \in \Omega_i$ and $y \in \Omega_j$. In case $|j - i| > 1$ it is obviously impossible for the pure swapping chain Q to accept a step from x to y, thus:

$$Q(x, y) = 0, \text{ if } x \in \Omega_i, y \in \Omega_j \text{ with } |i - j| > 1.$$

Hence,

$$\overline{Q}(i, j) = 0, \text{ if } |i - j| > 1.$$

The only way that i can change is by interchanging the first two coordinates x_0 and x_1 of x. For $0 \leq i < N$, we obtain

$$\begin{aligned}
\pi(\Omega_i)\overline{Q}(i,i+1) &= \sum_{x\in\Omega_i}\sum_{y\in\Omega_{i+1}} \pi(x)Q(x,y) \\
&= \sum_{x_0\in\Omega_+}\sum_{x_1\in\Omega_-}\sum_{\substack{x'\in\Omega_i \\ x'_0=x_0,x'_1=x_1}} \pi(x')Q(x',(0,1)x') \\
&= \sum_{x_0\in\Omega_+}\sum_{x_1\in\Omega_-}\sum_{\substack{x'\in\Omega_i \\ x'_0=x_0,x'_1=x_1}} \pi(x')\frac{1}{2M}\rho_{0,1}(x_0,x_1) \\
&= \frac{1}{2M}\sum_{x_0\in\Omega_+}\sum_{x_1\in\Omega_-} \pi_0(x_0)\pi_1(x_1)\rho_{0,1}(x_0,x_1) \\
&\quad \times \sum_{\substack{x'\in\Omega_i \\ x'_0=x_0,x'_1=x_1}} \prod_{j=2}^{M}\pi_j(x'_j) \\
&\in \frac{1}{2M}\sum_{x_0\in\Omega_+}\sum_{x_1\in\Omega_-} \pi_0(x_0)\pi_1(x_1) \\
&\quad \times \sum_{\substack{x'\in\Omega_i \\ x'_0=x_0,x'_1=x_1}} \prod_{j=2}^{M}\pi_j(x'_j)\left[e^{-\beta\frac{N(K+1)}{M}},1\right] \\
&\subseteq \frac{1}{2M}\binom{M-1}{i}\frac{1}{2^{M+1}}\left[e^{-\beta\frac{N(K+1)}{M}}a^{-1},a\right]
\end{aligned}$$

with the natural definitions of the sets in the last tow lines.

We will now give another, much simpler, Markov chain whose spectral gap has been intensively studied. Consider the symmetric random walk S on $\{0,...,M\}$, i.e.

$$\begin{aligned}
S(0,1) = S(0,0) &= S(M,M-1) = S(M,M) \\
&= S(i,i-1) = S(i,i+1) = \frac{1}{2} \text{ for } 0 < i < N.
\end{aligned}$$

Let $r(i) = \binom{M}{i}2^{-M}$ be the binomial distribution on $\{0,...,M\}$, and let R denote the Metropolis chain with proposal chain S and reversible distribution $r(i)$. As has been shown by Diaconis and Saloff-Coste [**10**, pp 698 and 719] R satisfies

$$\frac{1}{M} \leq \text{Gap}(R) \leq \frac{2}{M}. \tag{67}$$

In order to use Lemma 3.6 first note that

$$\pi(\Omega_i) \in \frac{1}{2^M}\binom{M}{i}[a^{-1},a] = r(i)[a^{-1},a] \tag{68}$$

implies $r(i) \geq \frac{1}{a}\pi(\Omega_i)$ for all $0 \leq i \leq M$. Second we conclude for $0 \leq i \leq N$,

$$\begin{aligned} r(i)R(i,i+1) &= r(i)S(i,i+1)\min\left\{1, \frac{r(i+1)}{r(i)}\right\} \\ &= r(i)\frac{1}{2}\min\left\{1, \frac{\binom{M}{i+1}}{\binom{M}{i}}\right\} \\ &= \begin{cases} r(i)\frac{1}{2}\cdot\frac{M-i}{i+1} & \text{if } i \geq \frac{M-1}{2} \\ r(i)\frac{1}{2} & \text{otherwise} \end{cases} \\ &= \begin{cases} \frac{1}{2^{M+1}}\binom{M}{i}\cdot\frac{M-i}{i+1} & \text{if } i \geq \frac{M-1}{2} \\ \frac{1}{2^{M+1}}\binom{M}{i} & \text{otherwise} \end{cases} \end{aligned}$$

Fixing $A := 4aMe^{\beta\frac{N(K+1)}{M}}$ it is now straightforward to check that

$$r(i)R(i,i+1) \leq A\pi(\Omega_i)\overline{Q}(i,i+1) \tag{69}$$

holds, for any i. It is now possible to use Lemma 3.6 which yields the desired inequality

$$\frac{1}{4M^2}e^{-\beta\frac{N(K+1)}{M}} = \frac{a}{A}\cdot\frac{1}{M} \leq \frac{a}{A}\operatorname{Gap}(R) \leq \operatorname{Gap}(\overline{Q}). \tag{70}$$

□

3.3. The Case $K > K_c$. Ellis et al. **[14]** show a continuous phase transition in the state space for these values of K. All but exponential little mass is located around

$$a_{\max}(0) := \left(\frac{e^{-\beta}}{1+2e^{-\beta}}, \frac{1}{1+2e^{-\beta}}, \frac{e^{-\beta}}{1+2e^{-\beta}}\right) \in \Upsilon_\infty \tag{71}$$

for $\beta < \beta_c(K)$ and for $\beta > \beta_c$ all but exponential little mass is located around the points

$$a_{\max}(-1) := \left(\frac{e^{2\beta K z_\alpha - \beta}}{C(\beta,K)}, \frac{1}{C(\beta,K)}, \frac{e^{-2\beta K z_\alpha - \beta}}{C(\beta,K)}\right) \in \Upsilon_\infty \tag{72}$$

$$a_{\max}(1) := \left(\frac{e^{-2\beta K z_\alpha - \beta}}{C(\beta,K)}, \frac{1}{C(\beta,K)}, \frac{e^{2\beta K z_\alpha - \beta}}{C(\beta,K)}\right) \in \Upsilon_\infty \tag{73}$$

with $C(\beta,K) = 1 + e^{-2\beta K z_\alpha - \beta} + e^{2\beta K z_\alpha - \beta}$ being the normalization constant and $z_\alpha(\beta,K) \geq 0$ as constructed but not computed in **[14]**, also see the appendix for an insight in the technical problems one faces. The standard Metropolis chain would get stuck in either of the regions around $a_{\max}(1)$ or $a_{\max}(-1)$ as it is exponentially unlikely for the chain to leave either of these local states. The swapping chain circumvents this bottleneck by swapping a component located close to $a_{\max}(-1)$ up to $\beta < \beta_c$ at which temperature the Metropolis chain is rapidly mixing on the whole state space. It will find a state close to $a_{\max}(0)$ and, if suggested to increase β, it will choose either of the two paths

leading to $a_{\max}(-1)$ or $a_{\max}(1)$ with equal probability. The bottleneck encountered in the intermediate regime $K_{\text{low}} < K < K_c$, which is described and used in Section 3.4, will not pose a problem, as

$$\beta \mapsto \begin{cases} a_{\max}(0) & \text{if } \beta \leq \beta_c \\ a_{\max}(1) & \text{if } \beta > \beta_c \end{cases} \tag{74}$$

is continuous.

To formalize this, a technique introduced by Bhatnagar and Randall [**2**, Sec. 4.1] will prove to be a powerful tool for showing rapid mixing of QP_+Q. We need to recall the notation of $\mathcal{A}_{s,r}$ introduced in (43). Assume β is big enough, such that the function f_β introduced in (48) on the field $\mathcal{A} = (A_{s,r})_{s,r}$ has two local maxima, such that it has two local modes. Inspired by (47) we define a probability measure $\mathbb{P}_{f_\beta}$ on $\mathcal{B} := \{(a_{-1}, a_1) \in [0,1]^2 | a_{-1} + a_1 \leq 1 \text{ and } a_{-1} \leq a_1\}$ by

$$\frac{dP_{f_\beta,N}}{d\lambda}(a_{-1}, a_1) := \frac{1}{Z_{f_\beta}(N)} e^{N f_\beta(a_{-1}, 1-a_{-1}-a_1, a_1)} \tag{75}$$

where λ denotes the Lebesgue-Measure restricted to the subset $\mathcal{B}$. $Z_{f_\beta}(N)$ denotes the normalization constant. Let $a_g(\beta_{i_c})$ denote the unique local maximum point of $f_{\beta_{i_c}}$ on $\mathcal{B}$ at the next to critical temperature

$$i_c := \max\{i | \beta_i \leq \beta_c(K)\}.$$

Further define the set

$$\mathcal{V} := \{a_{\max}(1) | \beta \geq \beta_c\}$$

which defines a continuous path from $a_{\max}(0)(\beta_c)$ to $(0,0,1)$ in $\mathcal{B}$. Take $\mathcal{V}$ to be an ordered set with the previously implied ordering. The path $\mathcal{V}$ separates $\mathcal{B}$ into two disjoint parts $\mathcal{B}_g \cup \mathcal{B}_l = \mathcal{B}$ with $\mathcal{V} \subseteq \mathcal{B}_g$. Obviously

$$P_{\beta_c,N}(\mathcal{B}_g) = 1 - P_{\beta_c,N}(\mathcal{B}_l) \to c \in (0,1)$$

for some K-specific constant c as $N \to \infty$. Remembering the models phase behavior we will define $\mathcal{B}_g$ and $\mathcal{B}_l$ by $(\frac{1}{2}, \frac{1}{2}) \in \mathcal{B}_g$ while $(0,0) \in \mathcal{B}_l$ as this notation reflects where the global and local maxima appear. With the definition of

$$\mathcal{A}_g(\beta_{i_c}) := \left(\mathcal{B}_g \cap \Upsilon\right) \text{ and } \mathcal{A}_l(\beta_{i_c}) := \left(\mathcal{B}_l \cap \Upsilon\right)$$

we know by continuity of π_β in β that $\pi_{\beta_{i_c}}(\mathcal{A}_g(\beta_{i_c})) \longrightarrow c$ and consequentially $\pi_{\beta_{i_c}}(\mathcal{A}_l(\beta_{i_c})) \longrightarrow 1 - c$. For any $i \in \{i_c + 1, ..., M\}$ there exist two local maxima, the global one denoted by $a_g(\beta_i)$ and the local

(non-global) one denoted by $a_l(\beta_i)$. We define $\mathcal{A}_g(\beta_i)$ and $\mathcal{A}_l(\beta_i)$ by

(76)

$$\text{there is no nondecreasing path from } a \text{ to } a_l \implies a \in \mathcal{A}_g(\beta_i)$$

(77)

$$\text{there is no nondecreasing path from } a \text{ to } a_g \implies a \in \mathcal{A}_l(\beta_i)$$

(78)

$$\text{there exist nondecreasing paths from } a \text{ to } a_g \text{ and from } a \text{ to } a_l \implies \begin{cases} a \in \mathcal{A}_g(\beta_i) \\ \text{ if } a \in \mathcal{A}_g(\beta_{i-1}) \\ a \in \mathcal{A}_l(\beta_i) \\ \text{ if } a \in \mathcal{A}_l(\beta_{i-1}) \end{cases}$$

Note that for each i the sets $\mathcal{A}_g(\beta_i)$ and $\mathcal{A}_l(\beta_i)$ form a partition of $\mathcal{B}$, since otherwise f_β would need to have more than two maxima on $\mathcal{B}$, in contradiction to Theorem A.2. It will prove convenient to have

LEMMA 5.7. *$(\pi_i(\mathcal{A}_g(\beta_i))_{i\in\{i_c,\ldots,M\}}$ is monotonically increasing, while $(\pi_i(\mathcal{A}_l(\beta_i))_{i\in\{i_c,\ldots,M\}}$ is monotonically decreasing.*

PROOF. This proof consists of multiple parts. We will first establish that for $\beta \geq \beta_c$

(79) $f_\beta(a_{\max}(0))$ is monotonically decreasing, while

(80) $f_\beta(a_{\max}(1))$ is monotonically increasing.

This is a straightforward calculation. Inserting $a_{\max}(0)(\beta)$ into f_β yields

$$\frac{df_\beta(a_{\max}(0))}{d\beta} = -\frac{2e^{-\beta}}{1+2e^{-\beta}} < 0$$

thus (79). Defining the canonical free energy of a thermodynamical system by

(81)

$$\phi(\beta) := \lim_{N\to\infty} \frac{1}{N} \log\left(Z_\beta(N)\right)$$

it follows from (47) that in the interesting phase of $\beta \geq \beta_c$

(82)

$$\phi(\beta) = f_\beta(a_{\max}(1)),$$

as

$$\begin{aligned}\phi(\beta) &= \lim_{N\to\infty}\frac{1}{N}\log\big(Z_\beta(N)\big)\\ &= \lim_{N\to\infty}\frac{1}{N}\log\big(\sum_{\mathbf{a}\in\Upsilon_N} e^{Nf_\beta(\mathbf{a})}\big)\\ &\leq \lim_{N\to\infty}\frac{1}{N}\log\big(N^2 e^{Nf_\beta(a_{\max}(1))}\big)\\ &= \lim_{N\to\infty}\big(\frac{2}{N}\log(N)+f_\beta(a_{\max}(1))\big)\\ &= f_\beta(a_{\max}(1))\end{aligned}$$

$$\begin{aligned}\phi(\beta) &= \lim_{N\to\infty}\frac{1}{N}\log\big(Z_\beta(N)\big)\\ &= \lim_{N\to\infty}\frac{1}{N}\log\big(\sum_{\mathbf{a}\in\Upsilon_N} e^{Nf_\beta(\mathbf{a})}\big)\\ &\geq \lim_{N\to\infty}\frac{1}{N}\log\big(e^{Nf_\beta(a_{\max}(1))}\big)\\ &= f_\beta(a_{\max}(1)).\end{aligned}$$

Differentiating for a fixed state $x = (x_{-1}, x_0, x_{+1})$ in the domain of f_β gives us

$$\frac{df_\beta}{d\beta}(x) = x_0 - 1 + K(x_1 - x_{-1})^2 \tag{83}$$

which implies

$$\frac{df_\beta}{d\beta}(x) \xrightarrow{x\to(0,0,1)} K - 1 > 0.$$

This guarantees $f_\beta(a_{\max}(1))$ to be strictly increasing for sufficiently large β. Together with the general fact (see for instance **[17]** or, for a non-rigorous overview, **[15]**) that $\phi(\beta)$ is concave for $\beta > \beta_c$ we gain (80).

In the second step we will confirm, that there is no point-movement from $\mathcal{A}_g$ to $\mathcal{A}_l$ by going from β_i to β_{i+1} for all $i_c \leq i \leq M-1$. For this, first note, that any point x, which has a nondecreasing path to any point $y \in \mathcal{V}$ also has a nondecreasing path to a_g. Assume, this to be wrong:

First note, that f_0 is monotonically decreasing on $\mathcal{V}$. Assume it would not be, then there are two points, $z_1, z_2 \in \mathcal{V}$ with $f_0(z_1) = f_0(z_2)$. As $a_{\max}(1)$ is continuously moving from $a_{\max}(0)(\beta_c)$ to $(0,0,1)$ there needs to be a $\beta' > \beta_c$ such that $f_{\beta'}(z_1) > f_{\beta'}(z_2)$. Of course, there also needs to be a $\beta'' > \beta'$ such that $f_{\beta''}(z_1) < f_{\beta''}(z_2)$. This contradicts (83).

Coming back to the original contradiction argument: By assumption, there exists a $\beta > \beta_c$ such that f_β, if restricted to $\mathcal{V}$, has at least

two modes – where, without loss of generality, the highest one is in the one containing $a_{\max}(0)(\beta_c)$. Take $z \in \mathcal{V}$ to be a local minimum. The points z' just further away from $a_{\max}(0)(\beta_c)$ than z must thus satisfy

$$\frac{df_\beta}{d\beta}(z) < \frac{df_\beta}{d\beta}(z'),$$

as f_0 is monotonically decreasing on $\mathcal{V}$ and the derivative of f_β with respect to β does not depend on β. This warrants for $f_\beta(z) < f_{\beta'}(z')$ for all $\beta' > \beta$ (again for the same reason), which in turn implies either $a_{\max}(1)$ stays left of z for all β or that $a_{\max}(1)$ exhibits a discontinuous behavior close to z. Both contradict a combination of Theorem .9 and the continuity of $a_{\max}(1)$.

This directly implies, that every point $x \in \mathcal{A}_g(\beta_{i_c})$ stays in $\mathcal{A}_g$ for all i, as any (nondecreasing) path leading from x to $a_l(\beta_i)$ will need to cross the set $\mathcal{V}$. A point $x \in \mathcal{A}_g(\beta_i)$ which does not lie in $\mathcal{A}_g(\beta_{i_c})$ must have been forced to switch from A_l to A_g at some index $i_c < j \leq i$. This means x is being separated from a_l by some path. Due to an argument close to the one given before, this path will block the way from x to a_l for any $i \geq j$, such that again, $x \in \mathcal{A}_g(\beta_{i+1})$.

Now, for any $\beta > \beta_c$ it follows from a similar calculations as for equation (82), that

$$\lim_{N\to\infty} \frac{1}{N} \log\left(\pi_{\beta_i}(\mathcal{A}_g)\right) = 0 \tag{84}$$

$$\lim_{N\to\infty} \frac{1}{N} \log\left(\pi_{\beta_i}(\mathcal{A}_l)\right) = f_{\beta_i}(a_{\max}(0)) - f_{\beta_i}(a_{\max}(1)) \tag{85}$$

which together with the first and second argument yields the claim. □

In preparation to use the decomposition theorem later on we need the following partitioning of the state space.

DEFINITION 5.8 (Definition 4.1 of [**2**]). *For $x \in {\Omega_+}^M$ define the trace*

$$\mathrm{Tr}(x) = t \in \{0,1\}^M$$

with $t_i = 0 \iff x_i \in \mathcal{A}_l$ and $t_i = 1 \iff x_i \in \mathcal{A}_g$ to indicate which part of the state space which component is in.

The 2^{M-i_c+1} possible values of $\mathrm{Tr}(x)$ characterize the partitioning

$${\Omega_+}^M = \bigcup_{t\in\{0,1\}^M} \Omega_{+t} \tag{86}$$

we will use. First using Lemma 3.7 for (87), Lemma 3.8 for (88) and afterwards Theorem 3.9 we gain

$$\text{Gap}(QP_+Q) \geq \frac{1}{3}\,\text{Gap}(QP_+QQP_+QQP_+Q) \tag{87}$$

$$\geq \frac{1}{3}\,\text{Gap}((QP_+Q)^{\frac{1}{2}}QP_+Q(QP_+Q)^{\frac{1}{2}}) \tag{88}$$

$$\geq \frac{1}{3}\,\text{Gap}(\widehat{Q}) \cdot \min_t \left\{ \text{Gap}\left((QP_+Q)|_{\text{Tr}^{-1}(t)}\right)\right\} \tag{89}$$

where $\hat{Q}$ is an abbreviation for the aggregated chain $\overline{QP_+Q}$. We can argue as in (60) to get

$$\begin{aligned}\text{Gap}(QP_+Q) &\geq \frac{1}{3}\,\text{Gap}(\widehat{Q}) \cdot \min_t \left\{ \text{Gap}\left((QP_+Q)|_{\text{Tr}^{-1}(t)}\right)\right\} \\ &\geq \frac{1}{24}\,\text{Gap}(\widehat{Q}) \cdot \min_t \left\{ \text{Gap}(Q|_{\text{Tr}^{-1}(t)}P_+|_{\text{Tr}^{-1}(t)}Q|_{\text{Tr}^{-1}(t)})\right\} \\ &\geq \frac{1}{24}\,\text{Gap}(\widehat{Q}) \cdot \min_t \left\{ \text{Gap}(P_+|_{\text{Tr}^{-1}(t)})\right\} \qquad (90)\end{aligned}$$

where the last inequality uses Lemma 3.8 again. This looks promising, as $P_+|_{\text{Tr}^{-1}(t)}$ is unimodal in each component as constructed, and thus the chain should be fast on this subset. $\widehat{Q}$ will be comparable to a very simple random walk, which is known to be rapidly mixing, thus leading to a polynomial lower bound for $\text{Gap}(QP_+Q)$.

3.3.1. *Speed of convergence of the aggregated chain $\widehat{Q}$.* We will follow in the wake of Bhatnagar and Randall [**2**, Theorem 4.4] and define the probability measure

$$\widehat{\pi}(t) := \prod_{i=1}^{M} \pi_i\Big(\text{Tr}_i^{-1}(t)\Big) \tag{91}$$

on the state space

$$\widehat{\Omega} = \prod_{i=1}^{i_c-1}\{1\} \times \prod_{i=i_c}^{M}\{0,1\}. \tag{92}$$

A simple reversible random walk $\widehat{RW1}$ with respect to $\widehat{\pi}$ to compare $\widehat{Q}$ on $\widehat{\Omega}$ to would be the following. Start at some $t \in \widehat{\Omega}$ and either switch the component t_{i_c} from 0 to 1 or vice versa with the Metropolis probabilities induced by $\widehat{\pi}$, or choose an $i \in \{i_c, ..., M-1\}$ at random and interchange components i and $i+1$ according to a Metropolis update with regard to $\widehat{\pi}$ as well, such that $t \to (i, i+1)t$. Again, for technical reasons $\widehat{RW1}$ does not act on t at all with probability $\frac{1}{2}$. In order to analyze $\widehat{RW1}$ we will compare it with an even simpler random walk $\widehat{RW2}$ on $\widehat{\Omega}$ which picks an $i \in \{i_c, ..., M\}$ at random and updates t_i by choosing t_i' exactly according to the stationary distribution $\widehat{\pi}_i$. It is apparent, that after this move, the ith component of t is in equilibrium.

Using the coupon collector's theorem (see for instance (2.7), (5.10) and (12.12) in **[25]**), we get easily

LEMMA 5.9. *Let $\widehat{R}$ denote the transition kernel of $\widehat{RW2}$. Then*

$$\mathrm{Gap}(\widehat{R}) \geq \frac{1}{4M \log M}.$$

This leads directly to

THEOREM 5.10. *The aggregated chain $\widehat{Q}$ of the Swapping Markov chain is rapidly mixing on $\widehat{\Omega}$ for $K < K_{low}$.*

Remark For why there is no explicit bound given, we would like to call the remark given for Theorem 5.4 to mind.

PROOF. The main idea is, to give a canonical path in $\widehat{RW1}$ in which every step compares well to the rapidly mixing chain $\widehat{R}$. Consider a single transition (t, t') in $\widehat{R}$, thus $t' = (t_1, ..., t_{i-1}, 1 - t_i, t_{i+1}, ..., t_M)$ for one $i \geq i_c$. Now consider the concatenation $p_1 \circ p_2 \circ p_3$ of the three paths

- p_1 consists of the $i - i_c$ swap moves from t to
$$t^{(1)} = (t_1, ..., t_{i_c-1}, t_i, t_{i_c}, ..., t_{i-1}, t_{i+1}, ..., t_M)$$
- p_2 is the one step from $t^{(1)}$ to
$$t^{(2)} = (t_1, ..., t_{i_c-1}, 1 - t_i, t_{i_c}, ..., t_M)$$
- p_3 consists of the $i - i_c$ steps needed to swap the ith component back up, thus p_2 is the path from $t^{(2)}$ to
$$t^{(3)} = (t_1, ..., t_{i_c}, ..., t_{i-1}, 1 - t_i, ..., t_M).$$

In order to use Lemma 3.5, we will establish, that

$$\widehat{\pi}(z)\widehat{RW1}(z, z') \geq \frac{1}{2}\, \widehat{\pi}(t)\widehat{R}(t, t') \tag{93}$$

holds for any transition (z, z') in the canonical path $p_1 \circ p_2 \circ p_3$.

Transition along p_1: Let $z = (t_0, ..., t_{i_c}, ..., t_{j-1}, t_i, t_j, ..., t_M)$ for a $j \in \{i_c + 1, ..., M\}$ and $z' = (j - 1, j)z$. It is easy to verify

$$\begin{aligned} \widehat{\pi}(z)\widehat{RW1}(z, z') &= \frac{\widehat{\pi}(z)}{2(M - i_c + 1)} \min\left(1, \frac{\widehat{\pi}(z')}{\widehat{\pi}(z)}\right) \\ &= \frac{1}{2(M - i_c + 1)} \min(\widehat{\pi}(z), \widehat{\pi}(z')) \end{aligned} \tag{94}$$

and for $t, t' = (t_1, ..., t_{i-1}, 1 - t_i, t_{i+1}, ..., t_M)$ for one $i \geq i_c$,

$$\begin{aligned} \widehat{\pi}(t)\widehat{R}(t, t') &= \frac{\widehat{\pi}(t)}{(M - i_c + 1)} \hat{\pi}_i(t'_i) \\ &\leq \frac{1}{(M - i_c + 1)} \widehat{\pi}(t^*) \end{aligned} \tag{95}$$

with $t^* = (t_1, ..., t_{i-1}, 0, t_{i+1}, ..., t_M)$. Thus it suffices to show $\widehat{\pi}(t^*) \leq \widehat{\pi}(z)$ and $\widehat{\pi}(t^*) \leq \widehat{\pi}(z')$. We will show this for z only, as the argument works exactly the same for both z and z'. It is useful to partition t^* into blocks of bits t_l that equal 1, separated by one or more zeros. Let $i_c \leq k < i$ be the largest value that satisfies $t_k = 0$. Using Lemma 5.7, it is straightforward to verify

$$\prod_{l=k+1}^{i} \widehat{\pi}_l(z_l) \geq \prod_{l=k+1}^{i} \widehat{\pi}_l(t_l^*).$$

Similarly , consider the next block of 1s in t^*, until the first index k' such that $t'_k = 0$,

$$\prod_{l=k'+1}^{k} \widehat{\pi}_l(z_l) \geq \prod_{l=k'+1}^{k} \widehat{\pi}_l(t_l^*).$$

Continuing in this way we find

$$\prod_{l=j}^{i} \widehat{\pi}_l(z_l) \geq \prod_{l=j}^{i} \widehat{\pi}_l(t_l^*)$$

and thus

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

In an analogous fashion one can also show

$$\widehat{\pi}(z') \geq \widehat{\pi}(t^*)$$

such that (93) holds on all transitions in p_1.

Transition along p_2: The same argument as before yields

$$\min(\widehat{\pi}(z), \widehat{\pi}(z')) \geq \widehat{\pi}(t^*)$$

for $(z, z') \in p_2$.

Transition along p_3: This is exactly as the case of p_1.

We find, that for any edge (z, z') in the canonical path equation (93) is satisfied, so what needs to be done in order to show rapid convergence of $\widehat{RW1}$ to equilibrium is to ensure that not too many paths use the same transition (z, z'). With the notation of Lemma 3.5, we can obviously bound the number of paths in $\tilde{E}(z, z')$ by M and as any path $\gamma_{t,t'}$ has at most $2M + 1$ many transitions, we can guarantee

(96)

$$A = \max_{(z,z')} \left\{ \frac{1}{\widehat{\pi}(z)\widehat{RW1}(z, z')} \sum_{\tilde{E}(z,z')} |\gamma_{t,t'}| \widehat{\pi}(t) \widehat{R}(t, t') \right\} \leq 4M^2 + 2M$$

which leads to $\mathrm{Gap}(\widehat{RW1}) \geq \left(2(2M^3 + M^2)\log(M)\right)^{-1}$.

It remains to compare $\widehat{RW1}$ with $\widehat{Q}$. We will do so by means of case differentiation. First consider the case of $z' = (i, i+1)z$ with $z_i = 1$, $z_{i+1} = 0$ in which we will show

$$\widehat{Q}(z, z') \geq \frac{1}{8M} \widehat{RW1}(z, z'). \tag{97}$$

So taking $z' = (i, i+1)z$ with $z_i = 1$, $z_{i+1} = 0$ leads to

$$\widehat{RW1}(z, z') = \frac{1}{2(M - i_c + 1)} \min\left(1, \frac{\widehat{\pi}(z')}{\widehat{\pi}(z)}\right) = \frac{1}{2(M - i_c + 1)} \tag{98}$$

as $\widehat{\pi}_i(1) \leq \widehat{\pi}_{i+1}(1)$ and $\widehat{\pi}_i(0) \geq \widehat{\pi}_{i+1}(0)$. The equivalent for $\widehat{Q}$ yields with $\mathcal{B} := \left\{x \in \Omega_{+z} \middle| x_i \in B_\varepsilon(a_g) \cap \mathcal{A}_g, x_{i+1} \in B_\varepsilon(a_l) \cap \mathcal{A}_l\right\}$

$$\begin{aligned}
\frac{1}{\widehat{\pi}(z)} \quad & \textstyle\sum_{x \in \Omega_{+z}} \sum_{y \in \Omega_{+z'}} \pi(x)\big(QP_+Q\big)(x, y) \\
& \geq \tfrac{1}{4\widehat{\pi}(z)} \textstyle\sum_{x \in \Omega_{+z}} \sum_{y \in \Omega_{+z'}} \pi(x) Q(x, y) \\
& = \tfrac{1}{4\widehat{\pi}(z)} \textstyle\sum_{x \in \Omega_{+z}} \pi(x) Q(x, (i, i+1)x)
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{4\widehat{\pi}(z)} \left(\sum_{x \in \mathcal{B}} \pi(x) Q(x, (i, i+1)x) + \sum_{x \in \Omega_{+z} \setminus \mathcal{B}} \pi(x) Q(x, (i, i+1)x) \right) \\
&\geq \frac{1}{4\widehat{\pi}(z)} \sum_{x \in \mathcal{B}} \pi(x) Q(x, (i, i+1)x) \\
&= \frac{1}{4\widehat{\pi}(z)} \frac{1}{2(M+1)} \pi(\mathcal{B}) \qquad (99) \\
&\geq \frac{1}{8(M+1)} - \frac{1}{8(M+1)} e^{-cN}. \qquad (100)
\end{aligned}$$

To get (99) is analogous to (98) for the not aggregated states. For (100), we use Theorem .9, which implies that

$$\frac{\pi_i(B_\varepsilon(a_g) \cap \mathcal{A}_g)}{\pi_i(A_g)} \frac{\pi_{i+1}(B_\varepsilon(a_l) \cap \mathcal{A}_l)}{\pi_{i+1}(A_l)} \geq 1 - e^{-cN},$$

for some $c > 0$. Second consider $z' = (i, i+1)z$ with $z_i = 0$, $z_{i+1} = 1$ which leads to

$$\widehat{RW1}(z, z') = \frac{1}{2(M - i_c + 1)} \min\left(1, \frac{\widehat{\pi}(z')}{\widehat{\pi}(z)}\right) = \frac{1}{2(M - i_c + 1)} \frac{\widehat{\pi}(z')}{\widehat{\pi}(z)} \tag{101}$$

and with $\mathcal{B}' := \big\{x \in \Omega_{+z} \big| x_i \in B_\varepsilon(a_l) \cap \mathcal{A}_l, x_{i+1} \in B_\varepsilon(a_g) \cap \mathcal{A}_g\big\}$

$$\begin{aligned}
&\frac{1}{\widehat{\pi}(z)} \sum_{x\in\Omega_{+z}} \sum_{y\in\Omega_{+z'}} \pi(x)\big(QP_+Q\big)(x,y) \\
&\qquad \geq \frac{1}{4\widehat{\pi}(z)} \sum_{x\in\Omega_{+z}} \sum_{y\in\Omega_{+z'}} \pi(x)Q(x,y) \\
&\qquad = \frac{1}{4\widehat{\pi}(z)} \sum_{x\in\Omega_{+z}} \pi(x)Q(x,(i,i+1)x) \\
&\qquad \geq \frac{1}{4\widehat{\pi}(z)} \sum_{x\in\mathcal{B}'} \pi(x)Q(x,(i,i+1)x) \\
&\qquad = \frac{1}{4\widehat{\pi}(z)} \sum_{x\in\mathcal{B}'} \frac{1}{2(M+1)} \pi\big((i,i+1)x\big) \qquad (102) \\
&\qquad = \frac{1}{4\widehat{\pi}(z)} \frac{1}{2(M+1)} \widehat{\pi}(z') \\
&\qquad \geq \frac{1}{8(M+1)} \frac{\widehat{\pi}(z')}{\widehat{\pi}(z)} - \frac{1}{8(M+1)} \frac{\widehat{\pi}(z')}{\widehat{\pi}(z)} e^{-cN}. \qquad (103)
\end{aligned}$$

The arguments for (102) and (103) are the same as above. The two remaining cases of $z' = (z_0, ..., 1 - z_{i_c}, ..., z_M)$ with $z_{i_c} \in \{0, 1\}$ is dealt with automatically, as by showing rapid mixing of P_{i_c} on $\mathcal{A}_g = \mathcal{A}$. The claim follows by using Lemma 3.6. □

3.3.2. *Rapid Mixing in $\mathcal{A}_g$ and $\mathcal{A}_l$.* It remains to show rapid convergence to equilibrium of $P_+|_{\mathrm{Tr}^{-1}(t)}$ as constructed in (90). In favor of a shorter notation and by using Theorem 3.10 we can stick to the case of

$$T := P_i|_{\mathrm{Tr}_i^{-1}(t)}$$

for fixed t and i. Using Lemma 3.7 with $m = 3$ gives us

$$\mathrm{Gap}(T) \geq \frac{1}{3}\,\mathrm{Gap}(T^3)$$

which will prove to be simpler to handle than T itself. We will only deal with the case of $\mathcal{A}_g$ as the case of $\mathcal{A}_l$ works technically the same. Consider the disjoint union

$$\mathcal{A}_g = \bigcup_{\mathcal{A}_{s,r}\subseteq\mathcal{A}_g} \mathcal{A}_{s,r} \qquad (104)$$

and decompose the state space accordingly. This leads to

$$\mathrm{Gap}(T^3) = \mathrm{Gap}(T^{\frac{1}{2}}T^2T^{\frac{1}{2}}) \geq \mathrm{Gap}(\overline{T}) \cdot \min_{s,r} \mathrm{Gap}(T^2_{s,r}) \qquad (105)$$

which may now make apparent, why choosing to deal with T^3 is an advantage over dealing with T. Here $\overline{T}$ is the aggregated chain defined as $\overline{\mathcal{Q}}$ in Theorem 3.9. Restricting T^2 to $\mathcal{A}_{s,r}$ will still give us a nontrivial

chain, whilst the restriction of T to $\mathcal{A}_{s,r}$ would deterministically stay in the originally occupied state.

THEOREM 5.11. $\mathrm{Gap}(\overline{T}) \geq \frac{1}{4}N^{-5}$

PROOF. This is already well prepared. As constructed earlier, f_β fulfills an unimodality condition on $\mathcal{A}_g$. Thus we can easily choose one path γ_{xy} for any given set x and y that is unimodal. Each such path has at most length N^2, such that the Poincaré inequality given in Lemma 3.4 simplifies to

$$\begin{aligned} A &= \max_{\langle \mathcal{A}_{s,r}, \mathcal{A}_{s',r'} \rangle} \frac{1}{\pi_i(\mathcal{A}_{s,r})\overline{T}(\mathcal{A}_{s,r}, \mathcal{A}_{s',r'})} \sum_{\gamma_{z_1 z_2} \ni \langle \mathcal{A}_{s,r}, \mathcal{A}_{s',r'} \rangle} |\gamma_{z_1 z_2}| \pi_i(z_1)\pi_i(z_2) \\ &\leq N^2 \max_{\langle \mathcal{A}_{s,r}, \mathcal{A}_{s',r'} \rangle} \frac{1}{\pi_i(\mathcal{A}_{s,r})\overline{T}(\mathcal{A}_{s,r}, \mathcal{A}_{s',r'})} \sum_{\gamma_{z_1 z_2} \ni \langle \mathcal{A}_{s,r}, \mathcal{A}_{s',r'} \rangle} \pi_i(z_1)\pi_i(z_2) \\ &= N^2 \max_{\langle \mathcal{A}_{s,r}, \mathcal{A}_{s',r'} \rangle} \sum_{\gamma_{z_1 z_2} \ni \langle \mathcal{A}_{s,r}, \mathcal{A}_{s',r'} \rangle} \frac{\pi_i(z_1)\pi_i(z_2)}{\pi_i(\mathcal{A}_{s,r})\overline{T}(\mathcal{A}_{s,r}, \mathcal{A}_{s',r'})} \end{aligned} \tag{106}$$

It is now of interest, how $\overline{T}$ behaves. Given $\mathcal{A}_{s,r} \neq \mathcal{A}_{s',r'}$ with $\overline{T}(A_{s,r}, \mathcal{A}_{s',r'}) > 0$, we first consider the case $\pi_i(\sigma) \leq \pi_i(\sigma')$ for $\sigma \in \mathcal{A}_{s,r}$ and $\sigma' \in \mathcal{A}_{s',r'}$. Note that $\pi_i(\sigma)$ is independent of the choice of $\sigma \in \mathcal{A}_{s,r}$.

$$\begin{aligned} \overline{T}(\mathcal{A}_{s,r}, \mathcal{A}_{s',r'}) &= \frac{1}{\pi_i(\mathcal{A}_{s,r})} \sum_{\sigma \in \mathcal{A}_{s,r}} \sum_{\sigma' \in \mathcal{A}_{s',r'}} \pi_i(\sigma) T(\sigma, \sigma') \\ &= \frac{1}{\pi_i(\mathcal{A}_{s,r})} \sum_{\sigma \in \mathcal{A}_{s,r}} \sum_{\sigma' \in \mathcal{A}_{s',r'}} \pi_i(\sigma) \frac{1}{4N} \\ &\geq \frac{1}{4N} \frac{1}{\pi_i(\mathcal{A}_{s,r})} \sum_{\sigma \in \mathcal{A}_{s,r}} \pi_i(\sigma) \\ &= \frac{1}{4N} \end{aligned}$$

The second case $\pi_i(\sigma) > \pi_i(\sigma')$ uses T's reversibility with

$$\begin{aligned}\overline{T}(\mathcal{A}_{s,r}, \mathcal{A}_{s',r'}) &= \frac{1}{\pi_i(\mathcal{A}_{s,r})} \sum_{\sigma \in \mathcal{A}_{s,r}} \sum_{\sigma' \in \mathcal{A}_{s',r'}} \pi_i(\sigma) T(\sigma, \sigma') \\ &= \frac{1}{\pi_i(\mathcal{A}_{s,r})} \sum_{\sigma \in \mathcal{A}_{s,r}} \sum_{\sigma' \in \mathcal{A}_{s',r'}} \pi_i(\sigma') T(\sigma', \sigma) \\ &= \frac{1}{4N} \frac{1}{\pi_i(\mathcal{A}_{s,r})} \sum_{\sigma' \in \mathcal{A}_{s',r'}} \sum_{\sigma \in \mathcal{A}_{s,r}} \pi_i(\sigma') \\ &\geq \frac{1}{4N} \frac{\pi_i(\mathcal{A}_{s',r'})}{\pi_i(\mathcal{A}_{s,r})}.\end{aligned}$$

To further analyze (106) we will take the worst case scenario $\frac{\pi_i(\mathcal{A}_{s',r'})}{\pi_i(\mathcal{A}_{s,r})} < 1$ and for inequality (107) remember that all paths are unimodal:

$$\begin{aligned}A &\leq N^2 \max_{\langle \mathcal{A}_{s,r}, \mathcal{A}_{s',r'} \rangle} \sum_{\gamma_{z_1 z_2} \ni \langle \mathcal{A}_{s,r}, \mathcal{A}_{s',r'} \rangle} \frac{\pi_i(z_1)\pi_i(z_2)}{\pi_i(\mathcal{A}_{s,r})\overline{T}(\mathcal{A}_{s,r}, \mathcal{A}_{s',r'})} \\ &\leq 4N^3 \max_{\langle \mathcal{A}_{s,r}, \mathcal{A}_{s',r'} \rangle} \sum_{\gamma_{z_1 z_2} \ni \langle \mathcal{A}_{s,r}, \mathcal{A}_{s',r'} \rangle} \frac{\pi_i(z_1)\pi_i(z_2)}{\pi_i(\mathcal{A}_{s,r})} \frac{\pi_i(\mathcal{A}_{s,r})}{\pi_i(\mathcal{A}_{s',r'})} \\ &= 4N^3 \max_{\langle \mathcal{A}_{s,r}, \mathcal{A}_{s',r'} \rangle} \sum_{\gamma_{z_1 z_2} \ni \langle \mathcal{A}_{s,r}, \mathcal{A}_{s',r'} \rangle} \frac{\pi_i(z_1)}{\pi_i(\mathcal{A}_{s,r})} \frac{\pi_i(z_2)}{\pi_i(\mathcal{A}_{s',r'})} \pi_i(\mathcal{A}_{s,r}) \\ &\leq 4N^5. \qquad (107)\end{aligned}$$

□

THEOREM 5.12. $\mathrm{Gap}(T^2_{s,r}) \geq \frac{1}{96N^6} e^{-\beta - 4K\beta}$.

PROOF. We need to consider two cases. The first is $\mathcal{A}_{N,N}$ in which case $|\mathcal{A}_{N,N}| = 1$, such that $T^2_{N,N}$ is the constant chain, and therefore rapidly mixing. The other case is $\mathcal{A}_{s,r}$ with $s \leq \min\{r, N-1\}$. Let $\sigma, \sigma' \in \mathcal{A}_{s,r}$ with $\sigma \neq \sigma'$. We will compare $T^2_{s,r}$ with the Markov chain $(X_i)_i$ given in Section 1.2. Assume $(j,k)\sigma = \sigma'$ for some $j, k \in \{1, ..., N\}$. Otherwise $T^2_{s,r}(\sigma, \sigma') = \mathbb{P}\big(X_{i+1} = \sigma' \big| X_i = \sigma\big) = 0$. We know

$$\mathbb{P}\big(X_{i+1} = \sigma' \big| X_i = \sigma\big) = \frac{1}{N^2}$$

and

$$T^2_{s,r}(\sigma, \sigma') \geq T(\sigma, \tau) T(\tau, \sigma')$$

for a fixed τ. It is obvious that either $T(\sigma, \tau) = \frac{1}{4N}$ or $T(\tau, \sigma') = \frac{1}{4N}$. Due to the symmetry assume

$$\tau := (\sigma_1, ..., \sigma_{j-1}, \sigma_k, \sigma_{j+1}, ..., \sigma_k, ...\sigma_N)$$

and conclude

$$\begin{aligned} T(\sigma,\tau) &= \frac{1}{4N}\min\left\{1, \frac{e^{\beta(N-R(\tau))-\frac{\beta K}{N}S^2(\tau)}}{e^{\beta(N-r)-\frac{\beta K}{N}s^2}}\right\} \\ &= \frac{1}{4N}\min\left\{1, e^{\beta(r-R(\tau))+\frac{\beta K}{N}\left(s^2-S^2(\tau)\right)}\right\} \\ &= \frac{1}{4N}\min\left\{1, e^{\beta(r-R(\tau))+\frac{\beta K}{N}\left(s-S(\tau)\right)\left(s+S(\tau)\right)}\right\} \\ &\geq \frac{1}{4N}e^{-\beta-4K\beta} \end{aligned}$$

such that taking $\tau = (\sigma_1, ..., \sigma_{j-1}, \sigma_k, \sigma_{j+1}, ...\sigma_N)$, where, without loss of generality, $\sigma_k > \sigma_j$ yields

$$T^2_{s,r}(\sigma,\sigma') \geq \frac{1}{16N^2}e^{-\beta-4K\beta}.$$

And we can easily deduce from Lemma 5.1 that $\mathrm{Gap}(X) \geq \frac{1}{6N^4}$ (see **[25]** for instance). Then Lemma 3.6 proves the claim. □

3.4. The Case $K_{\mathbf{low}} < K < K_c$. In this section we will prove Theorem 5.2. This is done in three parts. We first give the general idea, why slow mixing should be expected. We then support this idea with the necessary calculations in the remaining parts.

3.4.1. *The idea.* We will follow Gore and Jerrum **[21]** in order to find a bad cut in the state space of BEG for $\beta > \beta_c(K)$. Using their technique we can show, that the Metropolis chain has to overcome an exponential barrier to leave any local maximum. We will show, that an ε-stripe around the 0-axis contains such a maximum, with ε independent of β_i. Intuitively speaking this leads to the following behavior of the Tempering chain. At β_i close to 0 the chain will find the unique global maximum on the 0-axis. As of now the tempering chain is trapped in this ε-stripe, as Ellis et al. **[14]** show a discontinuous behavior of the global maximum as β_i passes through β_c. Thus the chain will never get the chance to leave this ε stripe within polynomial time at any temperature, even though, at low temperature, this stripe has exponential little mass.

3.4.2. *One bad cut for BEG's Metropolis chain.* Following the idea stated earlier, we show the existence of a bad cut within close proximity to the 0-axis in the two-phase region. It is well known, due to Ellis et al. **[14]**, that

$$a_{\max}(0) := \left(\frac{e^{-\beta}}{1+2e^{-\beta}}, \frac{1}{1+2e^{-\beta}}, \frac{e^{-\beta}}{1+2e^{-\beta}}\right) \in \Upsilon_\infty \tag{108}$$

is the unique global maximum for $\beta < \beta_c(K)$ and a local, non-global, maximum for $\beta > \beta_c(K)$. Here

$$\Upsilon_\infty := \{(a_{-1}, a_0, a_1) \in \mathbb{R}^3_+ : \sum_i a_i = 1\} \tag{109}$$

is the set of all probability measures on three points. They further show, that the phase transition for fixed K at $\beta_c(K)$ is discontinuous, thereby granting us, uniformly in β, the existence of an $\varepsilon > 0$ such that

$$\mathcal{N} := \big\{\sigma \big| |S_N(\sigma)| \leq N \cdot \varepsilon\big\} \tag{110}$$

contains only this local maximum, and f_β restricted to $B_\varepsilon(a_{\max}(0))$ is unimodal for all $\beta > 0$. It is even possible to show f_β restricted to $\mathcal{N}$ to be unimodal for all β, see Lemma .8 for details.

Recalling Section 1.1 we have

$$\begin{aligned}\pi(\sigma \text{ has type } N \cdot \mathbf{a}) &= \frac{1}{ZN} e^{-N\left(\beta\left(K(a_{-1}-a_1)^2 - a_1 - a_{-1}\right) - \sum_{i=-1}^{1} a_i \log a_i\right)} \\ &\quad \times e^{\Delta(\mathbf{a})} \\ &= \frac{1}{ZN} e^{-N f_\beta(\mathbf{a}) + \Delta(\mathbf{a})}. \end{aligned} \tag{111}$$

which implies, that every local maximum of f_β yields a locally exponential structure in π. This leads to exponentially low conductance $\Phi_{\mathcal{N}}$ for all $\beta > \beta_c(K)$, thereby implying slow mixing of the Metropolis algorithm in this regime.

3.4.3. *The bad cut for BEG's Simulated Tempering chain.* Having low conductance $\Phi_{\mathcal{N}}$ for any $\beta > \beta_c$ using the Metropolis algorithm it is easy to generalize this to the Simulated Tempering chain. To this end define

$$\mathcal{N}_{\text{edge}} := \big\{\sigma \big| N\varepsilon - 1 \leq |S_N(\sigma)| \leq N \cdot \varepsilon\big\} \tag{112}$$

and get

THEOREM 5.13. *Let $\mathcal{N}$ and $\mathcal{N}_{edge}$ be defined as in* (110) *and* (112). *There exists an $\varepsilon > 0$ such that for sufficiently large N and any $\beta \geq 0$*

$$\frac{\pi_\beta(\mathcal{N}_{edge})}{\pi_\beta(\mathcal{N})} \leq e^{-cN} \tag{113}$$

holds, with $c > 0$ only depending on K.

PROOF. Recall equation (111)

$$\pi(\sigma \text{ has type } N \cdot \mathbf{a}) = \frac{1}{ZN} e^{-N f_\beta(\mathbf{a}) + \Delta(\mathbf{a})}$$

and verify that there are only polynomially (in N) many $\mathbf{a} \in \Upsilon$ which satisfy $N \cdot \mathbf{a} \in \mathcal{N}_{\text{edge}}$. Then, considering

$$f_\beta(\mathbf{a}) = \beta\big(K(a_{-1} - a_1)^2 - a_1 - a_{-1}\big) - \sum_{i=-1}^{1} a_i \log a_i$$

and the results presented by Ellis et al. [**14**] it is clear, that f has a local maximum at $a_{\max}(0)$ (see equation (108)). Due to f being smooth in $a_{\max}$ it is clearly possible to find an $\varepsilon > 0$ such that f is unimodal on $B_\varepsilon(a_{\max})$. Due to the discontinuous behavior of the system at $\beta_c(K)$

for $K \in (K_{\text{low}}, K_c)$ and as $f_\beta(\mathbf{a})$ is smooth in all variables, including β, this ε can be chosen uniform in β.

Combining this with the exponential structure of (111) leads to the desired result

$$\frac{\pi_\beta(\mathcal{N}_{\text{edge}})}{\pi_\beta(\mathcal{N})} \leq e^{-cN}$$

with c depending only on K and sufficiently large N. □

This is the main ingredient for this section's main

THEOREM 5.14. *Define $\mathcal{N}$ and $\mathcal{N}_{edge}$ as in Theorem 5.13. The set*

$$\mathcal{S} := \{(x,i) | x \in \mathcal{N}, 0 \leq \beta_i \leq \beta\}$$

satisfies $\Phi_{\mathcal{S}} \leq e^{-cN}$ with $c > 0$.

Remark For the definition of the conductance $\Phi_{\mathcal{S}}$ of a set $\mathcal{S}$, see Theorem 2.6.

PROOF. Using Theorem 5.13 we get

$$\begin{aligned}
\Phi_{\mathcal{S}} &= \frac{\sum_{x\in\mathcal{S}, y\notin\mathcal{S}} \pi(x) QPQ(x,y)}{\pi(\mathcal{S})} \\
&= \frac{\sum_{\beta_i} \sum_{x\in\mathcal{N}_{\text{edge}}} \pi_i(x) \sum_{x'\in\mathcal{N}^c} QPQ(x,x')}{\sum_{\beta_i} \sum_{x\in\mathcal{N}} \pi_i(x)} \\
&\leq \frac{\sum_{\beta_i} \sum_{x\in\mathcal{N}_{\text{edge}}} \pi_i(x)}{\sum_{\beta_i} \sum_{x\in\mathcal{N}} \pi_i(x)} \\
&= \frac{\sum_{\beta_i} \pi_i(\mathcal{N}_{\text{edge}})}{\sum_{\beta_i} \pi_i(\mathcal{N}_{\text{edge}}) \frac{\pi_i(\mathcal{N})}{\pi_i(\mathcal{N}_{\text{edge}})}} \\
&\leq \frac{\sum_{\beta_i} \pi_i(\mathcal{N}_{\text{edge}})}{e^{cN} \sum_{\beta_i} \pi_i(\mathcal{N}_{\text{edge}})} \\
&= e^{-cN}
\end{aligned}$$

□

This concludes the proof of Theorem 5.2 by using Theorem 2.6.

CHAPTER 6

The Random-Energy-Model and the Generalized-Random-Energy-Model

The aim of the present chapter is twofold. On the one hand we want to analyze the Swapping (and Tempering) algorithm for a simple model of a spin glass. This analysis (together with [**26**]) is the first of its kind. On the other hand, we want to show that even for models with a third order phase transition like the Random Energy Model convergence of the Swapping chain can be slow. The speed of convergence of the Swapping algorithm in the first place seems to be correlated with the question, how well the underlying Markov chain is adapted to the probability distribution we want to simulate.

This chapter is organized in 4 sections. First we will define a natural Metropolis algorithm for the REM and the GREM in Section 1. In Section 2 we formulate the results – Both, the Swapping and the Tempering chain are slowly mixing for the Random Energy Model and the Generalized Random Energy Model – which are proofed in Section 3 and Section 4.

1. Defining the Metropolis chain

We will be looking at a natural and in the literature usually looked at (see [**18**]) realization of the Metropolis algorithm. Two states $\sigma \neq \sigma'$ are said to be neighbors, if they differ in exactly one component, thus their Hamming distance satisfies

$$\|\sigma - \sigma'\|_1 = 1.$$

To make the transition matrix of the proposal chain be a positive operator the chain will have a staying probability of at least $\frac{1}{2}$ and otherwise it will suggest any of the neighbors with equal probability:

$$K(\sigma,\sigma') = \begin{cases} \frac{1}{2N} & \text{if } \|\sigma - \sigma'\|_1 = 1 \\ 1 - \sum_{\tau \neq \sigma} K(\sigma,\tau) & \text{if } \sigma = \sigma' \\ 0 & \text{otherwise} \end{cases} \tag{114}$$

Verify that K is positive, irreducible and aperiodic on $\Omega = \{-1,1\}^N$. Let T denote the transition kernel of the corresponding Metropolis chain with regard to the desired Boltzmann distribution of either the REM or the GREM. Again, due to the construction, T is a positive,

irreducible and aperiodic transition kernel, which is reversible with respect to the desired distribution.

2. Results

In the rest of the chapter we are going to prove the following results, which state that Simulated Tempering and therefore also Swapping are slowly mixing for almost all realizations of the REM and the GREM.

THEOREM 6.1. *For almost all realizations of the* (X_σ), $\sigma \in \{0,1\}^N$, $N \in \mathbb{N}$ *the Simulated Tempering algorithm as well as the Swapping algorithm are slowly mixing in the REM.*

As to the GREM we will first show that the Metropolis-Hastings algorithm mixes torpidly.

THEOREM 6.2. *For almost all realizations of the* (X_σ), $\sigma \in \{0,1\}^N$, $N \in \mathbb{N}$ *the Metropolis-Hastings algorithm is slowly mixing for the GREM.*

As a consequence we also obtain torpid mixing for Swapping and Simulated Tempering in the GREM.

THEOREM 6.3. *For almost all realizations of the* (X_σ), $\sigma \in \{0,1\}^N$, $N \in \mathbb{N}$ *the Simulated Tempering algorithm as well as the Swapping algorithm are slowly mixing in the GREM.*

Concerning the proofs of Theorems 6.1 and 6.3 notice that the REM as well as the GREM satisfy the condition in Theorem 3.1. Hence we just need to show slow mixing for the Simulated Tempering algorithm.

3. Proofs for the REM

In this section we will prove Theorem 6.1. Recall that we want to simulate from the probability measure

$$\pi(\sigma) = \frac{e^{-\beta\sqrt{N}X_\sigma}}{Z(\beta)} \tag{115}$$

on $\Omega = \Omega_N = \{-1,1\}^N$.

Now for the Metropolis algorithm Fontes, Isopi, Kohayakawa, and Picco show **[18]** that for any fixed inverse temperature $\beta > 0$ it is slowly mixing. More precisely, they prove for the inverse of the spectral gap

$$\frac{1}{\mathrm{Gap}(P)} = \tau \tag{116}$$

the following inequality.

THEOREM 6.4 ([**18**] Prop. 3.1). *There exists a constant $c > 0$ such that, for all β, with $\mathbb{P}$-probability 1, for all but a finite number of indices N we have*

$$\frac{1}{N}\log(\tau) \geq \beta_c\beta - c\beta\sqrt{\frac{\log N}{N}} \tag{117}$$

with $\beta_c = \sqrt{2\log 2}$.

We will now try to translate this result to the case of the Simulated Tempering algorithm. Our proof is partially inspired by the techniques used by Bhatnagar and Randall [**2**] to show torpid mixing of Simulated Tempering on the Potts model. We consider subset

$$S := \{\underline{\sigma}\} \times \{0, ..., M\} \subset \Omega \times \{0, ..., M\},$$

where $\underline{\sigma} := \text{argmin}_\sigma\{X_\sigma\}$ is the spin with highest energy. We show that once the chain is in S, leaving it takes an exponentially long time. This is due fact that S is very narrow in the Ω-direction. Given i big enough the chain can only leave S by leaving state $\underline{\sigma}$ at the inverse temperature $\beta_{i-1} > 0$, $\beta_i > 0$ or given $i \neq M$ at $\beta_{i+1} > 0$. This should take exponential time, according to Theorem 6.4. In case i is smaller the chain could actually leave to a neighbor state of $\underline{\sigma}$, but given the chain is in S it is exponentially unlikely for the chain to be in a state of high temperature, so the chain will actually rarely get the opportunity to leave $\underline{\sigma}$ in the second way.

Define 1_S to be the indicator function in S. Using the notation of Section 2.2 we have

$$\begin{aligned}
\frac{1}{\tau(1_S)} &= \frac{\sum_{x\in S}\sum_{y\in S^c}\pi(x)P(x,y)}{\pi(S)(1-\pi(S))} \\
&= \frac{\sum_{i=0}^{k(M)}\pi((\underline{\sigma},i))\sum_{y\in S^c}P((\underline{\sigma},i),y)}{\pi(S)(1-\pi(S))} \\
&\quad + \frac{\sum_{i=k(M)+1}^{M}\pi((\underline{\sigma},i))\sum_{y\in S^c}P((\underline{\sigma},i),y)}{\pi(S)(1-\pi(S))} \\
&=: \Psi_l(S) + \Psi_h(S)
\end{aligned} \tag{118}$$

with $k(M)$ defined as in Corollary 6.6, below. Note, that the exact choice of $k(M)$ is unimportant, as long as its growth is of order N and $k(M)$ is small enough, such that C_2 in Corollary 6.6, below, is positive.

In order to bound $\Psi_l(S)$ and $\Psi_h(S)$ we derive the following consequences of Theorem 1.2

COROLLARY 6.5. *For every $\varepsilon > 0$ there exists with $\mathbb{P}$-probability 1 a $N_0 \in \mathbb{N}$ such that for all $\frac{i}{M}\beta \leq \beta_c$*

$$Z(\beta_i) \in \left\{ e^{N\ln(2)+\frac{i^2\beta^2}{2Nc_1{}^2}} e^{\delta N} \mid \delta \in (-\varepsilon, \varepsilon) \right\} \tag{119}$$

and for all $\frac{i\beta}{c_1 N} > \beta_c$

$$Z(\beta_i) \in \left\{ e^{N\left(\frac{\beta_c^2}{2}+(\frac{i\beta}{M}-\beta_c)\beta_c+\ln(2)\right)} e^{\delta N} \mid \delta \in (-\varepsilon, \varepsilon) \right\} \tag{120}$$

holds for all $N \geq N_0$.

This leads to the following

COROLLARY 6.6. *For every $\varepsilon > 0$ with $\mathbb{P}$-probability 1 there exists a N_0 such that for all $N \geq N_0$ and all $i \leq k(M) := i_0 M$*

$$\frac{\pi_i(\underline{\sigma})}{\pi_M(\underline{\sigma})} \leq e^{-C_2\beta N}\ e^{3\varepsilon N} \tag{121}$$

with i_0 and C_2 as in Lemma 6.11.

PROOF. The computation will be done in the more general framework of the GREM in the proof of Corollary 6.12 □

For an upper bound on $\Psi_h(S)$ we bound the transition probabilities of the Metropolis-Hastings chains $T_i := T_{\beta_i}$. Let g denote the indicator on $\underline{\sigma}$. We then arrive at

$$\begin{aligned} \mathrm{Gap}(T_i) &\geq \frac{\frac{1}{2}\sum_{x,y}(g(x)-g(y))^2\pi_i(x)T_i(x,y)}{\frac{1}{2}\sum_{x,y}(g(x)-g(y))^2\pi_i(x)\pi_i(y)} \\ &\geq \frac{\sum_{\langle\underline{\sigma},\sigma\rangle}\pi_i(\underline{\sigma})T_i(\underline{\sigma},\sigma')}{\pi_i(\underline{\sigma})(1-\pi_i(\underline{\sigma}))} \\ &\geq \frac{1}{1-\pi_i(\underline{\sigma})}\sum_{\langle\underline{\sigma},\sigma\rangle}T_i(\underline{\sigma},\sigma') \end{aligned} \tag{122}$$

On the other hand, we can bound the probabilities that the Simulated Tempering algorithm leaves S at temperature level i for $1 \leq i \leq M-1$

by

$$\sum_{y\in S^c} P((\underline{\sigma},i),y) = \sum_{\langle\underline{\sigma},\sigma\rangle}\sum_{j_1,j_2=-1}^{1} Q_{\underline{\sigma}}(i,i+j_1)T_{i+j_1}(\underline{\sigma},\sigma) \times Q_\sigma(i+j_1,i+j_1+j_2) \tag{123}$$

$$\leq \sum_{\langle\underline{\sigma},\sigma\rangle}\sum_{j_1,j_2=-1}^{1} T_{i+j_1}(\underline{\sigma},\sigma) \tag{124}$$

$$\leq 9\sum_{\langle\underline{\sigma},\sigma\rangle} T_{i-1}(\underline{\sigma},\sigma).$$

It is easy to see that similarly

$$\sum_{y\in S^c} P((\underline{\sigma},0),y) \leq 5\sum_{\langle\underline{\sigma},\sigma\rangle} T_0(\underline{\sigma},\sigma)$$

and

$$\sum_{y\in S^c} P((\underline{\sigma},M),y) \leq 5\sum_{\langle\underline{\sigma},\sigma\rangle} T_{M-1}(\underline{\sigma},\sigma).$$

From this, together with Theorem 6.4, we conclude, that

$$\Psi_h(S) = \frac{\sum_{i=k(M)+1}^{M}\pi((\underline{\sigma},i))\sum_{y\in S^c}P((\underline{\sigma},i),y)}{\pi(S)(1-\pi(S))}$$

$$\leq 9\sum_{i=k(M)+1}^{M}\frac{\pi_i(\underline{\sigma})}{\sum_{j=0}^{M}\pi_j(\underline{\sigma})}\frac{M+1}{\sum_{j=0}^{M}(1-\pi_j(\underline{\sigma}))}\sum_{\langle\underline{\sigma},\sigma\rangle}T_{i-1}(\underline{\sigma},\sigma)$$

$$\leq 9(M+1)\sum_{i=k(M)}^{M-1}\frac{\pi_{i+1}(\underline{\sigma})}{\sum_{j=0}^{M}\pi_j(\underline{\sigma})}\frac{1-\pi_i(\underline{\sigma})}{\sum_{j=0}^{M}(1-\pi_j(\underline{\sigma}))}\mathrm{Gap}(T_i) \tag{125}$$

$$\leq 9(M+1)\sum_{i=k(M)}^{M-1}\mathrm{Gap}(T_i)$$

$$\leq 9(M+1)\sum_{i=k(M)}^{M-1} e^{-N\beta_c\frac{i\beta}{c_1N}+\frac{ci\beta}{M}\sqrt{N\log(N)}}$$

$$\leq 9(M+1)\sum_{i=k(M)}^{M-1} e^{-Ni_0\beta_c+c\beta\sqrt{N\log(N)}}$$

$$= 9(M+1)(M-k(M))e^{-Ni_0\beta_c+c\beta\sqrt{N\log(N)}}$$

For small i the technique we just employed for estimating $\Psi_h(s)$, namely the exponentially slow mixing of the standard Metropolis algorithm of the REM, this might not be good enough, since, if we substitute $\frac{i\beta}{c_1 N}$ for β, N cancels out. However, Corollary 6.6 tells us, that – given we are in $\underline{\sigma}$ – the probability of being in a high temperature state is exponentially smaller than the probability of being at the lowest possible temperature β_M. Therefore,

$$\begin{aligned}
\Psi_l(S) &= \frac{\sum_{i=0}^{k(M)} \pi((\underline{\sigma},i)) \sum_{y\in S^c} P((\underline{\sigma},i),y)}{\pi(S)(1-\pi(S))} \\
&\leq 9\sum_{i=0}^{k(M)} \frac{\pi_i(\underline{\sigma})}{\sum_{j=0}^{M}\pi_j(\underline{\sigma})} \frac{M+1}{\sum_{j=0}^{M}(1-\pi_j(\underline{\sigma}))} \sum_{\langle\underline{\sigma},\sigma\rangle} T_{\max(i-1,0)}(\underline{\sigma},\sigma) \\
&\leq 9(M+1)\sum_{i=0}^{k(M)} \frac{\pi_i(\underline{\sigma})}{\sum_{j=0}^{M}\pi_j(\underline{\sigma})} \frac{1-\pi_{\max(i-1,0)}(\underline{\sigma})}{\sum_{j=0}^{M}(1-\pi_j(\underline{\sigma}))} \operatorname{Gap}(T_{\max(i-1,0)}) \qquad (126) \\
&\leq 9(M+1)\sum_{i=0}^{k(M)} \frac{\pi_i(\underline{\sigma})}{\sum_{j=0}^{M}\pi_j(\underline{\sigma})} \\
&\leq 9(M+1)\sum_{i=0}^{k(M)} \frac{\pi_i(\underline{\sigma})}{\pi_M(\underline{\sigma})} \\
&\leq 9(M+1)\sum_{i=0}^{k(M)} e^{-C_2\beta N}\ e^{3\varepsilon N} \\
&= 9(M+1)(k(M)+1)e^{-C_2\beta N}\ e^{3\varepsilon N}
\end{aligned}$$

for any $\varepsilon > 0$ and all $N \geq N_0$ with a $\mathbb{P}$-a.s. finite N_0. Plugging these estimates of $\Psi_l(S)$ and $\Psi_h(S)$ into (118) gives the desired result.

4. Proofs for the GREM

In this section we will use similar methods as in the case of the REM to prove Theorems 6.2 and 6.3. To give an idea, why slow mixing of both, the Metropolis-Hastings chain and Simulated Tempering are plausible, recall, that in the GREM we decompose Ω into n factors as in (16). For fixed n the size of each factor growth exponentially in N. If $\underline{\sigma} = \underline{\sigma}_1 \cdots \underline{\sigma}_n$ denotes the maximum $H(X_{\underline{\sigma}})$ there are n potentially different ways to leave this state. Lets say we tried to leave it by changing $\underline{\sigma}_i$ in $\sigma_i \neq \underline{\sigma}_i$ for $i \in \{1,...,n\}$, then all $(\alpha_i)^N\ X'_{\sigma_i\cdots\sigma_n} := \sqrt{a_i}X_{\underline{\sigma}_1\cdots\underline{\sigma}_{i-1}\sigma_i} + ... + \sqrt{a_n}X_{\underline{\sigma}_1\cdots\sigma_i\cdots\underline{\sigma}_n}$ are i.i.d. $\mathcal{N}(0, a_i + ... + a_n)$ random variables. This is, after proper rescaling, a

REM situation so taken all $i \in \{1, ..., n\}$ together it suggests slow mixing of the standard Metropolis algorithm for the GREM since $n \in \mathbb{N}$ is fixed. Since the standard Metropolis chain is slow in the REM for every $\beta > 0$ this leads to a slow Metropolis algorithm for the GREM for every $\beta > 0$. This should take care of the low temperature part of the Simulated Tempering algorithm, just as it does for the REM.

Since similar results about $X_{\underline{\sigma}} = \min\{X_\sigma\}$ are also known for the GREM, and even Z_β in the GREM behaves a lot like Z_β in the REM we can hope to be able to apply similar techniques.

We first show that the Metropolis algorithm is slowly mixing in the GREM, hence Theorem 6.2. To this end we will employ Dirichlet-forms, as in Section 2.2. Again define $\underline{\sigma} := \text{argmin}_\sigma\{X_\sigma\}$ and let $S := \{\underline{\sigma}\} \in \Omega$. We define

$$\mathscr{A}_i := \{\sigma \in \Omega | \sigma = \underline{\sigma}_1 \cdots \underline{\sigma}_{i-1} \sigma_i \underline{\sigma}_{i+1} \cdots \underline{\sigma}_n\} \setminus \{\underline{\sigma}\}$$

for $i \in \{1, ..., n\}$. It is apparent that any neighbor σ of $\underline{\sigma}$ is in $\bigcup \mathscr{A}_i \cup \{\underline{\sigma}\}$, and that the $\mathscr{A}_i$ are pairwise disjoint. Let $T := T_\beta$ denote the transition matrix of the Metropolis chain in the GREM and with $\mathscr{T}_i := T_{|\mathscr{A}_i \cup \{\underline{\sigma}\}}$ the restriction of T to $\mathscr{A}_i \cup \{\underline{\sigma}\}$. Note that π induces a probability measure π_i on $\mathscr{A}_i$ through restriction, such that

$$\pi_i(\sigma) := \frac{e^{-\beta\sqrt{\frac{\ln 2}{\ln(\alpha_i)}\sum_{j=i}^n a_j}\sqrt{N\frac{\ln(\alpha_i)}{\ln 2}}X'_\sigma}}{Z_{R(i)}(\beta)} \tag{127}$$

for $\sigma \in \mathscr{A}_i \cup \{\underline{\sigma}\}$ and $\pi_i(\sigma) = 0$ otherwise. In this context

$$Z_{R(i)}(\beta) := \sum_{\sigma \in \mathscr{A}_i \cup \{\underline{\sigma}\}} e^{-\beta\sqrt{\frac{\ln 2}{\ln(\alpha_i)}\sum_{j=i}^n a_j}\sqrt{N\frac{\ln(\alpha_i)}{\ln 2}}X'_\sigma}$$

denotes the restricted partition function and, for $\sigma \in \mathscr{A}_i \cup \{\underline{\sigma}\}$,

$$X'_\sigma := \frac{1}{\sqrt{\sum_{j=i}^n a_i}} \sum_{j=i}^n \sqrt{a_j} X_{\sigma_1 \ldots \sigma_j}$$

are i.i.d. $\mathscr{N}(0,1)$ random variables. With this notation it is apparent, that on $\mathscr{A}_i$ $\mathscr{T}_i$ is the standard Metropolis chain for the REM at inverse temperature $\beta\sqrt{\frac{\ln 2}{\ln(\alpha_i)}\sum_{j=i}^n a_j}$. From this we conclude with the help of Theorem 6.4

$$\text{Gap}(\mathscr{T}'_i) \leq e^{-\beta_c\beta\sqrt{\frac{\ln(\alpha_i)\sum_{j=i}^n a_j}{\ln 2}}N + c\beta\sqrt{\sum_{j=i}^n a_j}\sqrt{N\log\left(N\frac{\ln(\alpha_j)}{\ln 2}\right)}} \tag{128}$$

where $\mathscr{T}_i'$ denotes the Markov chain equivalent to $\mathscr{T}_i$ on $\mathscr{A}_i \cup \{\underline{\sigma}\}$. This leads to the following estimate for the Dirichlet-form.

$$\begin{aligned}
\frac{1}{\tau(1_S)} &= \frac{\sum_{\sigma\in S}\pi(\sigma)\sum_{\sigma'\in S^c}T(\sigma,\sigma')}{\pi(S)(1-\pi(S))} \\
&= \frac{1}{1-\pi(S)}\sum_{\langle\sigma,\underline{\sigma}\rangle}T(\underline{\sigma},\sigma) \\
&= \frac{1}{1-\pi(S)}\sum_{i=1}^{n}\sum_{\substack{\sigma\in\mathscr{A}_i\\ \langle\sigma,\underline{\sigma}\rangle}}T(\underline{\sigma},\sigma) \\
&\le \sum_{i=1}^{n}\frac{1-\pi_i(S)}{1-\pi(S)}\ \mathrm{Gap}(\mathscr{T}_i') \qquad \text{(i)} \\
&\le \sum_{i=1}^{n} e^{-\beta_c\beta\sqrt{\frac{\ln(\alpha_i)\sum_{j=i}^n a_j}{\ln 2}}N+c\beta\sqrt{\sum_{j=i}^n a_j}\sqrt{N\log(N)}} \qquad \text{(ii)} \\
&\le n\ e^{-\beta_c\beta\sqrt{\frac{a_n\min_i\{\ln(\alpha_i)\}}{\ln 2}}N+c\beta\sqrt{N\log(N)}}
\end{aligned}$$

(i) holds because of (122) and since $Z_{R(i)}(\beta) \le Z(\beta)$ implies $\pi_i(\underline{\sigma}) \ge \pi(\underline{\sigma})$. For (ii) note that $\frac{\ln(\alpha_i)}{\ln 2} \le 1$. This proves Theorem 6.2.

We are now armed to also prove Theorem 6.3. We have just seen, that the standard Metropolis chain for the GREM behaves a lot like the Metropolis chain of the REM. In this perspective it is not all to surprising that the same proof for slow mixing of the REM simulated tempering works for the GREM as well, with just some minor adjustments.

Recall that P denotes the transition matrix of the Simulated Tempering chain in the GREM and $\underline{\sigma} := \mathrm{argmin}_\sigma\{X_\sigma\}$. Define $S = \{\underline{\sigma}\}\times\{0,...,M\}$. Using the notation of Section 2.2 we have

$$\begin{aligned}
\frac{1}{\tau(1_S)} &= \frac{\sum_{x\in S}\sum_{y\in S^c}\pi(x)P(x,y)}{\pi(S)(1-\pi(S))} \\
&= \frac{\sum_{i=0}^{k(M)}\pi((\underline{\sigma},i))\sum_{y\in S^c}P((\underline{\sigma},i),y)}{\pi(S)(1-\pi(S))} \\
&\quad + \frac{\sum_{i=k(M)+1}^{M}\pi((\underline{\sigma},i))\sum_{y\in S^c}P((\underline{\sigma},i),y)}{\pi(S)(1-\pi(S))} \\
&=: \Psi_l(S)+\Psi_h(S)
\end{aligned}$$

with $k(M)$ defined as in Corollary 6.12. The following theorems enable us to bound $\Psi_l(S)$ and $\Psi_h(S)$ from above.

THEOREM 6.7 ([**4**] Theorem 1.5). *Define the sequence* $J_1, ..., J_m$ *by* $J_0 := 0$ *and*

$$J_l := \min\{J > J_{l-1} | A_{J_{l-1}+1,J} > A_{J+1,k} \ \forall k \geq J+1\} \tag{129}$$

with $A_{j,k} := \sum_{i=j}^{k} \frac{a_i}{2\ln(\prod_{i=j}^{k} \alpha_i)}$. *Further define*

$$\overline{a}_l := \sum_{i=J_{l-1}+1}^{J_l} a_i$$

$$\overline{\alpha}_l := \sum_{i=J_{l-1}+1}^{J_l} \alpha_i$$

With this notation

$$\max_{\sigma} \left\{ \frac{1}{\sqrt{N}} X_\sigma \right\} \longrightarrow \sum_{i=1}^{m} \sqrt{2\overline{a}_i \ln(\overline{\alpha}_i)} \tag{130}$$

holds with $\mathbb{P}$*-probability 1.*

Remark For all $l = 1, ..., m$ and all k such that $J_{l-1} + 2 \leq k \leq J_l$

$$\frac{\sum_{i=k}^{J_l} a_i}{\overline{a}_l} \geq \frac{\sum_{i=k}^{J_l} \ln(\overline{\alpha}_i)}{\ln(\overline{\alpha}_l)} \tag{131}$$

holds, which is the condition, for the concave hull as described in Figure 10.3 in [**3**].

THEOREM 6.8 ([**3**] Theorem 10.1.10). *Using the notation from Theorem 6.7 and defining*

$$\gamma_l := \sqrt{\frac{\overline{a}_l}{2\ln(\overline{\alpha}_l)}} \tag{132}$$

we get for $l(\beta) := \max\{l \geq 1 | \beta\gamma_l > 1\}$ *and* $l(\beta) := 0$ *if* $\beta\gamma_1 \leq 1$

$$\lim_{N\to\infty} \left[\frac{1}{N} \ln \left(2^{-N} \sum_{\sigma} e^{\beta H(\sigma)} \right) \right] = \beta \sum_{i=1}^{l(\beta)} \sqrt{2\overline{a}_i \ \ln(\overline{\alpha}_i)} + \sum_{i=J_{l(\beta)}+1}^{n} \frac{\beta^2 a_i}{2} \tag{133}$$

with $\mathbb{P}$*-probability 1.*

COROLLARY 6.9. *For every $\varepsilon > 0$ there exists with $\mathbb{P}$-probability 1 a $N_0 \in \mathbb{N}$ such that*

(134)
$$Z(\beta_i) \in \left\{ e^{N\ln(2)+N\beta\sum_{i=1}^{l(\beta)}\sqrt{2\overline{a}_i\ \ln(\overline{\alpha}_i)}+N\sum_{i=J_{l(\beta)}+1}^{n}\frac{\beta^2 a_i}{2}}\ e^{\delta N} \middle| \delta \in (-\varepsilon,\varepsilon) \right\}$$

holds for every $N \geq N_0$.

LEMMA 6.10. *For every $\varepsilon > 0$ with $\mathbb{P}$-probability 1 there exists a N_0 such that for all $N \geq N_0$*

$$\frac{\pi_i(\underline{\sigma})}{\pi_M(\underline{\sigma})} \leq e^{-\frac{N\beta}{2}\sum_{j=l(\beta)+1}^{m}\sqrt{2\overline{a}_j\ln(\overline{\alpha}_j)}-\frac{i^2\beta^2}{{c_1}^2N}+\frac{i\beta}{c_1}\sum_{j=1}^{m}\sqrt{2\overline{a}_j\ln(\overline{\alpha}_j)}}\ e^{3\varepsilon N} \tag{135}$$

PROOF. We compute

$$\begin{aligned}
\frac{\pi_i(\underline{\sigma})}{\pi_M(\underline{\sigma})} &= \frac{Z(\beta_M)}{Z(\beta_i)} e^{\frac{i\beta}{M}H(\underline{\sigma})-\beta H(\underline{\sigma})} \\
&\leq e^{N\ln(2)+N\beta\sum_{j=1}^{l(\beta)}\sqrt{2\overline{a}_j\ \ln(\overline{\alpha}_j)}+N\sum_{j=J_{l(\beta)}+1}^{n}\frac{\beta^2 a_j}{2}-N\ln(2)-N\sum_{j=1}^{n}\frac{i^2\beta^2 a_j}{2M^2}} \\
&\quad \times e^{\beta H(\underline{\sigma})\left(\frac{i}{M}-1\right)}\ e^{2\varepsilon N} \\
&\leq e^{N\beta\sum_{j=1}^{l(\beta)}\sqrt{2\overline{a}_j\ \ln(\overline{\alpha}_j)}+N\sum_{j=J_{l(\beta)}+1}^{n}\frac{\beta^2 a_j}{2}-\frac{i^2\beta^2}{2{c_1}^2N}} \\
&\quad \times e^{\left[\sum_{j=1}^{m}\sqrt{2\overline{a}_j\ln(\overline{\alpha}_j)}\right]\beta N\left(\frac{i}{M}-1\right)}\ e^{3\varepsilon N} \\
&\leq e^{-N\beta\left[\sum_{j=l(\beta)+1}^{m}\sqrt{2\overline{a}_j\ln(\overline{\alpha}_j)}-\beta\sum_{j=J_{l(\beta)}+1}^{n}\frac{a_j}{2}\right]-\frac{i^2\beta^2}{2{c_1}^2N}+\frac{i\beta}{c_1}\sum_{j=1}^{m}\sqrt{2\overline{a}_j\ln(\overline{\alpha}_j)}} \\
&\quad \times e^{3\varepsilon N} \\
&= e^{-N\beta\sum_{j=l(\beta)+1}^{m}\left[\sqrt{2\overline{a}_j\ln(\overline{\alpha}_j)}-\frac{\beta}{2}\overline{a}_j\right]-\frac{i^2\beta^2}{2{c_1}^2N}+\frac{i\beta}{c_1}\sum_{j=1}^{m}\sqrt{2\overline{a}_j\ln(\overline{\alpha}_j)}}\ e^{3\varepsilon N} \\
&= e^{-N\beta\sum_{j=l(\beta)+1}^{m}\sqrt{\overline{a}_j}\left[\sqrt{2\ln(\overline{\alpha}_j)}-\frac{\beta}{2}\sqrt{\overline{a}_j}\right]-\frac{i^2\beta^2}{2{c_1}^2N}+\frac{i\beta}{c_1}\sum_{j=1}^{m}\sqrt{2\overline{a}_j\ln(\overline{\alpha}_j)}}\ e^{3\varepsilon N} \\
&\leq e^{-N\beta\sum_{j=l(\beta)+1}^{m}\sqrt{\overline{a}_j}\left[\sqrt{2\ln(\overline{\alpha}_j)}-\frac{1}{2}\sqrt{2\ln(\overline{\alpha}_j)}\right]-\frac{i^2\beta^2}{2{c_1}^2N}+\frac{i\beta}{c_1}\sum_{j=1}^{m}\sqrt{2\overline{a}_j\ln(\overline{\alpha}_j)}} \\
&\quad \times e^{3\varepsilon N} \\
&= e^{-\frac{N\beta}{2}\sum_{j=l(\beta)+1}^{m}\sqrt{2\overline{a}_j\ln(\overline{\alpha}_j)}-\frac{i^2\beta^2}{2{c_1}^2N}+\frac{i\beta}{c_1}\sum_{j=1}^{m}\sqrt{2\overline{a}_j\ln(\overline{\alpha}_j)}}\ e^{3\varepsilon N} \\
&= e^{-N\beta\left(\frac{1}{2}\sum_{j=l(\beta)+1}^{m}\sqrt{2\overline{a}_j\ln(\overline{\alpha}_j)}+\frac{i^2\beta}{2{c_1}^2N^2}-\frac{i}{c_1N}\sum_{j=1}^{m}\sqrt{2\overline{a}_j\ln(\overline{\alpha}_j)}\right)}\ e^{3\varepsilon N}
\end{aligned}$$

where the last inequality follows from (132). □

LEMMA 6.11. *There exists $i_0 > 0$ and $C_2 > 0$ such that for all $i \leq i_0 c_1 N = i_0 M$*

$$\frac{1}{2}\sum_{j=l(\beta)+1}^{m}\sqrt{2\overline{a}_j\ln(\overline{\alpha}_j)}+\frac{i^2\beta}{2{c_1}^2N^2}-\frac{i}{c_1N}\sum_{j=1}^{m}\sqrt{2\overline{a}_j\ln(\overline{\alpha}_j)}\geq C_2 \tag{136}$$

PROOF. This is simply a matter of calculus. Define the continuous function

$$f(i):=\frac{1}{2}\sum_{j=l(\beta)+1}^{m}\sqrt{2\overline{a}_j\ln(\overline{\alpha}_j)}+\frac{i^2\beta}{2{c_1}^2N^2}-\frac{i}{c_1N}\sum_{j=1}^{m}\sqrt{2\overline{a}_j\ln(\overline{\alpha}_j)}. \tag{137}$$

Since $l(\beta)\leq m-1$, we have that $f(0)>0$ holds. This yields

$$f_0:=\frac{c_1N}{\beta}\left(\sum_{j=1}^{m}\sqrt{2\overline{a}_j\ln(\overline{\alpha}_j)}\right. \\ \left.-\sqrt{\Big(\sum_{j=1}^{m}\sqrt{2\overline{a}_j\ln(\overline{\alpha}_j)}\Big)^2-\beta\sum_{j=l(\beta)+1}^{m}\sqrt{2\overline{a}_j\ln(\overline{\alpha}_j)}}\right) \tag{138}$$

as the left zero of f. Clearly $f_0>0$, since

$$f'\left(\frac{c_1N\sum_{j=1}^{m}\sqrt{2\overline{a}_j\ln(\overline{\alpha}_j)}}{\beta}\right)=0$$

and thus $\frac{1}{\beta}c_1N\sum_{j=1}^{m}\sqrt{2\overline{a}_j\ln(\overline{\alpha}_j)}$ is the only extremal point (and therefore the minimum) of f, and $\frac{c_1N\sum_{j=1}^{m}\sqrt{2\overline{a}_j\ln(\overline{\alpha}_j)}}{\beta}>0$. Now, $f_0\in\mathbb{C}\setminus\mathbb{R}$ would lead to $f>0$, so we can conclude that the pair

$$i_0:=\begin{cases}\lfloor\left(\frac{f_0}{2c_1N}\right)\rfloor & \text{if } f_0\in\mathbb{R}\\ \frac{1}{2} & \text{if } f_0\notin\mathbb{R}\end{cases}$$
$$C_2:=f(i_0M)$$

satisfies all conditions. Note that neither i_0 nor C_2 depend on N. □

Lemma 6.10 and 6.11 lead to

COROLLARY 6.12. *For every $\varepsilon>0$ with $\mathbb{P}$-probability 1 there exists a N_0 such that for all $N\geq N_0$ and all $i\leq K(M):=i_0M$*

$$\frac{\pi_i(\underline{\sigma})}{\pi_M(\underline{\sigma})}\leq e^{-C_2\beta N}\ e^{3\varepsilon N} \tag{139}$$

with i_0 and C_2 as in Lemma 6.11.

Now we will give an upper bound for $\Psi_h(S)$. Recall that (122) gives

$$(1-\pi_i(\underline{\sigma}))\,\mathrm{Gap}(T_i) \geq \sum_{\langle \underline{\sigma},\sigma\rangle} T_i(\underline{\sigma},\sigma').$$

Moreover,

$$\sum_{y\in S^c} P((\underline{\sigma},i),y) \leq 9 \sum_{\langle \underline{\sigma},\sigma\rangle} T_{i-1}(\underline{\sigma},\sigma)$$

as in (123) is true for the GREM as well. We get

$$\begin{aligned}
\Psi_h(S) &= \frac{\sum_{i=k(M)+1}^{M} \pi((\underline{\sigma},i)) \sum_{y\in S^c} P((\underline{\sigma},i),y)}{\pi(S)(1-\pi(S))} \\
&\leq 9 \sum_{i=k(M)+1}^{M} \frac{\pi_i(\underline{\sigma})}{\sum_{j=1}^{M}\pi_j(\underline{\sigma})} \frac{M+1}{\sum_{j=0}^{M}(1-\pi_j(\underline{\sigma}))} \sum_{\langle \underline{\sigma},\sigma\rangle} T_{i-1}(\underline{\sigma},\sigma) \\
&\leq 9(M+1) \sum_{i=k(M)}^{M-1} \frac{\pi_{i+1}(\underline{\sigma})}{\sum_{j=1}^{M}\pi_j(\underline{\sigma})} \frac{1}{\sum_{j=0}^{M}(1-\pi_j(\underline{\sigma}))} \sum_{\langle \underline{\sigma},\sigma\rangle} T_i(\underline{\sigma},\sigma) \\
&\leq 9(M+1) \sum_{i=k(M)}^{M-1} \frac{\pi_{i+1}(\underline{\sigma})}{\sum_{j=1}^{M}\pi_j(\underline{\sigma})} \frac{1-\pi_i(\underline{\sigma})}{\sum_{j=0}^{M}(1-\pi_j(\underline{\sigma}))} \mathrm{Gap}(T_i) \\
&\leq 9(M+1) \sum_{i=k(M)}^{M-1} \mathrm{Gap}(T_i)
\end{aligned}$$

$$\begin{aligned}
&\leq 9(M+1) \sum_{i=k(M)}^{M-1} n\; e^{-\beta_c \frac{i\beta}{M}\sqrt{\frac{a_n \max\{\ln(\alpha_j)\}}{\ln 2}}N + c\frac{i\beta}{M}\sqrt{N\log(N)}} \\
&\leq 9(M+1) \sum_{i=k(M)}^{M-1} n\; e^{-\beta_c i_0 \beta\sqrt{\frac{a_n \max\{\ln(\alpha_j)\}}{\ln 2}}N + c\beta\sqrt{N\log(N)}} \\
&= 9nM(M+1)(1-i_0) e^{-\beta_c i_0 \beta\sqrt{\frac{a_n \max\{\ln(\alpha_j)\}}{\ln 2}}N + c\beta\sqrt{N\log(N)}}
\end{aligned}$$

The upper bound for $\Psi_l(S)$ follows similarly to the case of the REM.

$$\begin{aligned}
\Psi_l(S) &= \frac{\sum_{i=0}^{k(M)} \pi((\underline{\sigma}, i)) \sum_{y \in S^c} P((\underline{\sigma}, i), y)}{\pi(S)(1 - \pi(S))} \\
&\leq 9 \sum_{i=0}^{k(M)} \frac{\pi_i(\underline{\sigma})}{\sum_{j=1}^{M} \pi_j(\underline{\sigma})} \frac{M+1}{\sum_{j=0}^{M}(1 - \pi_j(\underline{\sigma}))} \sum_{\langle \underline{\sigma}, \sigma \rangle} T_{\max(i-1,0)}(\underline{\sigma}, \sigma) \\
&\leq 9(M+1) \sum_{i=0}^{k(M)} \frac{\pi_i(\underline{\sigma})}{\sum_{j=1}^{M} \pi_j(\underline{\sigma})} \frac{1 - \pi_{\max(i-1,0)}(\underline{\sigma})}{\sum_{j=0}^{M}(1 - \pi_j(\underline{\sigma}))} \operatorname{Gap}(T_{\max(i-1,0)}) \\
&\leq 9(M+1) \sum_{i=0}^{k(M)} \frac{\pi_i(\underline{\sigma})}{\sum_{j=1}^{M} \pi_j(\underline{\sigma})} \\
&\leq 9(M+1) \sum_{i=0}^{k(M)} \frac{\pi_i(\underline{\sigma})}{\pi_M(\underline{\sigma})} \\
&\leq 9(M+1) \sum_{i=0}^{k(M)} e^{-C_2 \beta N} \; e^{3\varepsilon N} \\
&= 9 i_0 M e^{-C_2 \beta N} \; e^{3\varepsilon N}
\end{aligned}$$

for any $\varepsilon > 0$ and all $N \geq N_0$ with a $\mathbb{P}$-a.s. finite N_0. This finishes the proof of Theorem 6.3.

CHAPTER 7

Equi-Energy sampling as a derivative of the Swapping algorithm

After having studied the Swapping algorithm for multiple models in the prior chapters, this chapter will deal with a derived algorithm suggested by Kou, Zhou and Wong in [**24**] called Equi-Energy sampling.

The principle observation is that a main obstacle to fast mixing is the presence of a phase transition in the model. This, in turn, may be characterized by a multimodal distribution of a macroscopic observable. Usually then the (projected) Metropolis chain enters one of the modes rapidly and stays there for an exponentially long time. The Equi-Energy method tries to avoid this behavior by introducing shortcuts in the state space. These shortcuts are created by the observations of Metropolis chains at higher temperatures where the above mentioned modes are less pronounced or possibly not even present. Additionally to the Metropolis steps one allows also for jumps to points of the same energy as the present one, given one has observed these points already at higher temperatures (otherwise, the algorithm would require the knowledge of the exact structure of the energy function, in which case simulation would probably be pointless).

Besides defining the algorithm, Kou et al. also address the convergence question and give simulations of interesting examples. The goal of the present chapter is to shed some more light on the convergence and, in particular, the speed of convergence of the Equi-Energy algorithm. To this end we will see that even under the best conditions of knowing the whole energy landscape, Equi-Energy sampling may be slow in some models. Namely we will see this for the Potts model given in Chapter 1.2. The phenomenon responsible is the same used by Bhatnagar and Randall in [**2**] to show torpid mixing of Tempering and Swapping and similar to the phenomenon encountered in the BEG model as seen in Chapter 5.

This chapter is organized in the following way: After defining the Equi-Energy algorithm in Section 1 we will see in Section 2 that Equi-Energy sampling is not well suited for the Potts-Model.

1. The Equi-Energy algorithm

In this section we introduce the base chain for the Metropolis-Hastings algorithm and give a definition of a version of the Equi-Energy sampler.

Remember the Metropolis-Hastings algorithm to by defined by

$$T_\beta(\sigma,\tau) = \begin{cases} K_{gen}(\sigma,\tau) & \text{if } \sigma \neq \tau \text{ and } H(\tau) \geq H(\sigma) \\ K_{gen}(\sigma,\tau)\frac{\pi_\beta(\tau)}{\pi_\beta(\sigma)} & \text{if } \sigma \neq \tau \text{ and } H(\tau) < H(\sigma) \\ 1 - \sum_{\upsilon \neq \sigma} T_\beta(\sigma,\upsilon) & \text{otherwise.} \end{cases} \tag{140}$$

for an aperiodic, symmetric and irreducible Markov chain K_{gen} on Ω. As in the Curie-Weiss model the proposal chain for the Potts model selects one coordinate at random and suggests to change the coordinates color to any other color with equal probability, thus K_{gen} is given by

$$K_{gen}(\sigma,\tau) := \begin{cases} \frac{1}{2} & \text{if } \sigma = \tau \\ \frac{1}{2N(q-1)} & \text{if } \|\sigma - \tau\| = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{141}$$

As mentioned in Chapter 2.3.2 the induced Metropolis-Hastings sampler is torpidly mixing on the Potts model for β in the multiple-phase region.

To speed up its convergence, we first introduce a sequence of energy levels:

$$\inf_\sigma H(\sigma) := h_0 < h_1 < \ldots < h_M = \sup H(\sigma) \tag{142}$$

and a sequence of inverse temperature levels

$$0 = \beta_0 < \beta_1 < \ldots < \beta_M = \beta$$

where we assume that β is the temperature we want to sample from. For the same reasons as for the Swapping algorithm it will often be convenient and necessary to take $\beta_i = \frac{i\beta}{M}$. Note that M may and will depend on N, which is not made explicit in **[24]**.

Moreover, we will need a dummy state ι and define

$$\tilde{\Omega} := \Omega \cup \{\iota\}.$$

Let $\mathcal{M}$ be an $M \times |\Omega|$ matrix over $\tilde{\Omega}$, which is initially filled with ι's, only.

The Equi-Energy algorithm consists of alternating between two steps. One is a Metropolis update at a random temperature level β_i. The other one is an Equi-Energy jump again at a randomly selected temperature β_i with $i \geq 1$. At temperature 0 there are only Metropolis moves. We keep record of the results of each single one of the states we see at temperature β_i by entering them into the i'th row of the matrix $\mathcal{M}$, if it has not been seen before. In this case it replaces one of the ι's.

The Equi-Energy step works as follows: Assume, the chain at temperature $\beta_i, i \geq 1$ is in state σ. Then an Equi-Energy jump consists of determining the energy level k, such that $h_{k-1} < H(\sigma) \leq h_k$ and choosing with equal probability a state τ from all states υ with $h_{k-1} < H(\upsilon) \leq h_k$, which have already been observed at temperature level β_{i-1}. This new state is accepted with probability $\min\left\{1, \frac{\pi_{\beta_i}(\tau)\pi_{\beta_{i-1}}(\sigma)}{\pi_{\beta_i}(\sigma)\pi_{\beta_{i-1}}(\tau)}\right\}$. Otherwise, in particular, if there is no state in the same energy band in the $i-1$st row of $\mathcal{M}$, we stay where we are. We denote the corresponding transition matrix (on Ω) as well as the corresponding operator by Q_i. Of course, we can also consider the process, that describes the movement of all the particles simultaneously, which is a process on Ω^{M+1}. The transition matrix on Ω^{M+1} that moves the i'th coordinate according to Q_i and lets all others rest, i.e. the matrix corresponding to the operator

$$\bigotimes_{j\leq i-1} I \otimes Q_i \otimes \bigotimes_{j=i+1}^{M} I$$

is denoted by $\mathcal{Q}_i$ (where I is the identity). Similarly, we will denote by T_i the Metropolis chain T_{β_i} at temperature β_i and by $\mathcal{T}_i$ the chain on Ω^{M+1} corresponding to the operator

$$\bigotimes_{j\leq i-1} I \otimes T_i \otimes \bigotimes_{j=i+1}^{M} I.$$

The Equi-Energy sampler is then defined as

$$\mathcal{R} = \frac{1}{(M+1)^3} \sum_{j,k,l=0} \mathcal{Q}_j \mathcal{T}_k \mathcal{Q}_l.$$

Note that even though we are just interested in what happens with the last coordinate of this process, it might still be worth and easier to consider the entire process. Also note that $\mathcal{R}$ is not a Markov chain. To consider a Markov chain, will however be useful sometimes. To this end, we define the natural extension $\tilde{\mathcal{R}}$ to the space $\Omega^{M+1} \times \tilde{\Omega}^{(M+1)\times|\Omega|}$, which in the first coordinate moves according to $\mathcal{R}$ and in the second coordinate contains the elements of $\mathcal{M}$. The latter in turn has in the i'th row all the states visited by T_{i-1} up to the present time entered in some fixed, deterministic order.

The role of this complicated Markov chain becomes apparent, when we prove the following theorem.

THEOREM 7.1. *The distribution of the M+1'st coordinate of $\mathcal{R}$ converges to π_β.*

PROOF. By reversibility, for each i, T_i is reversible with respect to the measure π_{β_i} and the entries in $\mathcal{M}$ do not play any role for T_i. Second, note that in the second coordinate of $\tilde{\mathcal{R}}$ the state M_0

in which each row of the matrix $\mathcal{M}$ is filled with an entry different from ι is absorbing (since we fill the matrix in some fixed order, M_0 in unique. The state M_0 is reached in finite time almost surely, for almost all realizations of $\tilde{R}$ there is $n_0 = n_0(\omega)$, such that the second coordinate of $\tilde{\mathcal{R}}$ at n_0 is M_0. However, once the second coordinate of $\tilde{\mathcal{R}}$ is M_0 all the processes Q_i are reversible with respect to π_{β_i}. This is basically since, $Q_i(\sigma, \tau) > 0$, then σ and τ are in the same energy shell $(h_k, h_{k+1}]$, and hence Q_i starting from σ draws from the same points as when starting from τ. The rest of the reversibility assertion follows from the definition of the transition probabilities.

Hence we are in the following situation: The random walk $\tilde{\mathcal{R}}$ eventually concentrates on $\Omega^{M+1} \times \{M_0\}$. The restriction $\tilde{\mathcal{R}}|_{\Omega^{M+1}\times\{M_0\}}$ is reversible with respect to the measure $\pi \times \delta_{M_0}$. Here for $x = (x_0, x_1, \ldots, x_M) \in \Omega^{M+1}$

$$\pi(x) = \prod_{i=0}^{M} \pi_{\beta_i}(x_i),$$

and δ_{M_0} is the Dirac measure concentrated in the matrix M_0. Hence by the ergodic theorem for Markov chains $\tilde{\mathcal{R}}$ converges in distribution to $\pi \times \delta_{M_0}$, which in particular implies that the $M+1$st coordinate of $\mathcal{R}$ converges to π_β.

□

Remark From the above proof one may get the impression, that for a state space Ω that is exponentially large in some parameter N, to ask for polynomial convergence is completely hopeless, since we first have to fill $\mathcal{M}$, before we can apply the ergodic theorem. We will see an even worse effect for the Potts model. The convergence to the desired probability measure may still be slow, even if the entire energy landscape has already been saved to the matrix $\mathcal{M}$.

2. Equi-Energy for the mean-field Potts model

We will see, that the Equi-Energy algorithm is not well suited for the three-color-mean-field-Potts model, namely we will show

THEOREM 7.2. *The speed of convergence to equilibrium of the Equi-Energy sampler is exponentially slow for the mean-field Potts model with $q = 3$ colors.*

Remark The proof given also works for any number $q > 3$ of colors. In favor of an easier notation we will confine ourselves to the case of $q = 3$.

Torpid mixing of the Equi-Energy sampler is slow for the same reasons why the simulated tempering algorithm is slowly mixing for the mean-field Potts model as shown in [**2**]. As stated in Section 2 of Chapter 1, the mean-field Potts model undergoes a discontinuous phase

transition. As Gore and Jerrum [**21**] show, the model shows, for any $\beta > \beta_c$, besides the three modes for each color, a fourth mode in which every color appears almost equally often. This is the main difference to the Curie-Weiss model which has, for any $\beta > \beta_c$, exactly one mode for each color. Suppose the chain $\mathcal{R}$ has all components started in a center state. Component x_0 can only do metropolis updates. It will very likely stay close to the center, thus only saving center states to the matrix. Component x_1 then could either do metropolis or Equi-Energy updates. Metropolis alone would prefer the center states as the center is, for any temperature, a locally likely state. The only chance to leave the center would thus be an Equi-Energy jump. But as the chain x_0 did not save any states distant to the center, it will never get proposals to jump away from the center.

The proof will follow a slightly different strategy, even though the idea behind the argument is the one given above.

2.1. System specific preparation. In order to handle the state space more easily we will follow the approach by Gore and Jerrum [**21**]. Remember that for any $\sigma \in \Omega$

$$\pi_\beta(\sigma) = \frac{1}{Z(\beta)} e^{\beta H(\sigma)}.$$

As we have a mean-field model we thus get for any $a = (a_1, a_2, a_3)$ with $\sum a_i = 1$ and $Na_i \in \mathbb{N}_0$ for any $i \in \{1, 2, 3\}$

$$\pi_\beta\Big(\big\{\sigma \big| \mathfrak{m}_N(\sigma) = a\big\}\Big) = \binom{N}{Na_1, Na_2, Na_3} \frac{1}{Z(\beta)} e^{\beta H(\sigma_a)}. \tag{143}$$

Here σ_a is one representative with $\mathfrak{m}_N(\sigma_a) = a$. Using Stirling's approximation we get

$$\pi_\beta\Big(\big\{\sigma \big| \mathfrak{m}_N(\sigma) = a\big\}\Big) = \frac{e^{Nf(a)+\Delta(a)}}{NZ(\beta)}. \tag{144}$$

with

$$f(a) := \sum_{i=1}^{3} \left(\frac{\beta}{2} a_i^2 - a_i \log a_i \right) \tag{145}$$

and $\Delta(a) \in o(N)$. Define $\mathcal{D} = \{a \in [0,1]^3 | \sum a_i = 1\}$ to be the domain of f. Gore and Jerrum were able to show

LEMMA 7.3 (Proposition 1 in [**21**]). *Let $a = (a_1, ..., a_q)$ be a local maximum point of f. Then a satisfies the following properties:*

1. *a lies in the interior of $\mathcal{D}$.*
2. *Either $a_i = q^{-1}$ for all i, or there are α and α' such that $0 < \alpha < \beta^{-1} < \alpha' < 1$, and $a_i \in \{\alpha, \alpha'\}$ for all i.*
3. *If a is such, that the a_i are not all equal, then there is a unique component a_j such that $a_j = \alpha'$; the other components satisfy $a_j = \alpha$ for all $j \neq i$.*

This immediately leads to

LEMMA 7.4. *Fix $\beta > \beta_c$. Then there exists a $\delta > 0$ and a constant $c > 0$ such that*

$$\pi_\beta\Big(\Big\{\sigma\Big|\Big\|\mathfrak{m}_N(\sigma) - a_{\frac{1}{3}}\Big\| > \delta\Big\}\Big) \leq e^{-cN}. \tag{146}$$

Here $a_{\frac{1}{3}} := (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$.

PROOF. Let a_{q1} denote the local maximum with dominant color 1 and remember $a_{\frac{1}{3}}$ to be the local maximum in the center of the state space. Either $a_{\frac{1}{3}}$ or a_{q1} is a global maximum point. Assume first, that $f(a_{q1}) > f(a_{\frac{1}{3}})$. Define the set

$$\mathfrak{M} := \big\{a \in \mathcal{D}\big|f(a) > f(a_{\frac{1}{3}}), Na \in \{0, ..., N\}^3\big\}$$

and note that there exists a $\delta > 0$ such that

$$\Big\{\sigma\Big|\Big\|\mathfrak{m}_N(\sigma) - a_{\frac{1}{3}}\Big\| > \delta\Big\} \cap \mathfrak{M} = \emptyset$$

as $a_{\frac{1}{3}}$ is a local maximum point of a smooth function f. (Without loss of generality one could choose δ small enough such that

$$\Big\{\sigma\Big|\Big\|\mathfrak{m}_N(\sigma) - a_{\frac{1}{3}}\Big\| > \delta\Big\}$$

only contains $a_{\frac{1}{3}}$ as a local maximum point.) As there are at most N^2 many different points a with $Na \in \{0, ..., N\}^3$ and $\sum a_i = 1$ it is easily checked that

$$\begin{aligned}
\pi_\beta\Big(\Big\{\sigma\Big|\Big\|\mathfrak{m}_N(\sigma) - a_{\frac{1}{3}}\Big\| > \delta\Big\}\Big) &\leq \pi_\beta(\mathfrak{M}^c) \\
&= \frac{\sum_{\sigma \notin \mathfrak{M}} e^{\beta H(\sigma)}}{\sum_{\sigma \in \Omega} e^{\beta H(\sigma)}} \\
&\leq \frac{N^2 e^{Nf(a_{\frac{1}{3}})}}{e^{Nf(a_{q1})}} \\
&= N^2 e^{-N(f(a_{q1}) - f(a_{\frac{1}{3}}))} \qquad (147)
\end{aligned}$$

holds. Using the same argument yields that the case $f(a_{\frac{1}{3}}) \geq f(a_{q1})$ contradicts $\beta > \beta_c$ as the only macrostate would then be the center state. □

2.2. Proof for Theorem 7.2. We want to use a conductance argument (see Theorem 2.6) in order to show Theorem 7.2. Assume the Matrix $\mathcal{M}$ to be filled completely, thus $\mathcal{M} = M_0$. Then the Equi-Energy algorithm is reversible to the desired distribution and Theorem 2.6 can be applied. Now assume the last coordinate x_M of x is in the center of the state space, thus close to $a_{\frac{1}{3}}$. For $\beta > \beta_c$ the chain would need to see at least one of the modes containing a_{q1}, a_{q2} or a_{q3} in reasonable time. To accomplish this the chain has two possibilities,

either using metropolis steps, or using Equi-Energy steps in order to leave the center and get drawn to any of the other modes. Metropolis itself is slow on the Potts model for $\beta > \beta_c$ as can be seen in a similar way to the one given in Theorem 2.7 for the Curie-Weiss model, by using either of the four modes, especially the center mode, as a bad cut. Thus the only hope to have rapid convergence would be the Equi-Energy component of the algorithm. The Hamilton function grows as the distance to the center increases. As the Equi-Energy step can only get proposals which are in the same energy band as the current state the chain cannot increase its distance to the center by any considerable amount, using an Equi-Energy update. This effectively traps x_M close to the center and thus the center is a bad cut for the Equi-Energy algorithm. Note that widening the energy bands would not increase the speed of convergence either, as there are exponentially many more states in the center then on the outer circle. Having wide energy bands would lead to many proposals close to the center which the chain could either accept, if close to the center anyhow, or, in case the chain is in a distance to the center, reject as the new states energy is exponentially much smaller than the current states energy. Especially the hope of switching from the mode containing a_{q1} to any other mode is disappointed if the energy band is too wide, as too many Equi-Energy proposals will be flat out rejected if x_M is currently close to a_{q1}.

Remember (142) and note that we will henceforth only consider $M = cN$ to be linearly dependent on N and the h_i to be equi-distanced in the domain. Choosing M super-linearly would lead to empty energy bands and thus effectively lead to an exact-Equi-Energy sampling. On the other hand, having M fixed leads to almost non-interactive components and Equi-Energy sampling stands no chance of increasing the speed of convergence compared to the standard Metropolis algorithm.

LEMMA 7.5. *For every $\varepsilon > 0$ and every $\varepsilon > \delta > 0$ there exists a N_0 such that for all $N > N_0$ and for all $\sigma, \tau \in \Omega$ which satisfy $\|\mathfrak{m}_N(\sigma) - a_{\frac{1}{3}}\| < \delta$ and $\|\mathfrak{m}_N(\tau) - a_{\frac{1}{3}}\| > \varepsilon$*

$$Q_M(\sigma, \tau) = 0.$$

PROOF. First note that $a_{\frac{1}{3}}$ is the (unique) minimum for

$$a \mapsto \frac{1}{N} H(a) := \frac{1}{N} H(\sigma_a) = \frac{1}{2} \sum_{i=1}^{3} a_i^2$$

on $\mathcal{D}$. Further note that $\frac{1}{N} H(\sigma_a) = \frac{1}{2} \|a\|_2^{\,2}$ is basically the square of the two-norm of the vector a. Now $Q_M(\sigma, \tau) > 0$ only if there is an index i such that $h_i < H(\sigma) < h_{i+1}$ and $h_i < H(\tau) < h_{i+1}$. This leads

to

$$\begin{aligned} |H(\sigma) - H(\tau)| &= \frac{h_M - h_0}{M} \\ &= \frac{N - \frac{N}{6}}{2M} \\ &= \frac{5}{12c} \end{aligned} \tag{148}$$

thus in order for σ_a and $\sigma_{a'}$ to be in the same energy band, a and a' must satisfy

$$\|a - a'\|_2{}^2 \leq \frac{5}{12cN}.$$

Using the equivalence of the one norm and the two norm on $\mathcal{D}$ concludes the proof. □

We now define a set $\tilde{\mathcal{S}}$ which guarantees for a small conductance. First define for every $\varepsilon > 0$

$$S_\varepsilon := \left\{ x \in \Omega^{M+1} \middle| \left\| \mathfrak{m}_N(x_M) - a_{\frac{1}{3}} \right\| < \varepsilon \right\}. \tag{149}$$

Fix ε such that $\left\{ a \in \mathcal{D} \middle| \|a - a_{\frac{1}{3}}\| < \varepsilon \right\}$ only contains points of the center mode. Now choose $\delta' > 0$ and N_0 according to Lemma 7.5 and the remark thereafter. As $\tilde{R}$ consists of possibly two Equi-Energy jumps we need to iterate this construction once. For this choose $0 < \delta'' < \delta'$ and afterwards choose $\delta > 0$ – and again N even larger – according to Lemma 7.5 this time with δ'' as input.

LEMMA 7.6. *There exists a $c' > 0$ such that*

$$\frac{\pi_M(\mathcal{S}_\varepsilon \setminus \mathcal{S}_\delta)}{\pi_M(\mathcal{S}_\varepsilon)} < e^{-c'N}.$$

PROOF. The proof works exactly as the proof for Lemma 7.4. □

This directly leads to

THEOREM 7.7. *Consider $\tilde{R}$ and its state space. The conductance Φ satisfies for a $c' > 0$*

$$\Phi \leq e^{-c'N}.$$

PROOF. With the notation introduced above, it is clear by Lemma 7.4 that $\pi_M(\mathcal{S}_\varepsilon) < \frac{1}{2}$ for N large enough. Remember

$$\Omega_{\text{EE}} := \Omega^{M+1} \times \tilde{\Omega}^{(M+1)\times|\Omega|}$$

to be the state space for $\tilde{R}$. Then it is clear that $\pi\big((\Omega^M \times \mathcal{S}_\varepsilon) \times \{M_0\}\big) < \frac{1}{2}$ holds for all N large enough as well.

$$\begin{aligned}
\Phi_{(\Omega^M \times \mathcal{S}_\varepsilon)\times\{M_0\}} &= \frac{\sum_{\substack{\tilde{\sigma}\in(\Omega^M \times \mathcal{S}_\varepsilon)\times\{M_0\} \\ \tilde{\tau}\notin(\Omega^M \times \mathcal{S}_\varepsilon)\times\{M_0\}}} \pi(\tilde{\sigma})\tilde{R}(\tilde{\sigma},\tilde{\tau})}{\pi\big((\Omega^M \times \mathcal{S}_\varepsilon) \times \{M_0\}\big)} \\
&= \frac{\sum_{\substack{\tilde{\sigma}\in(\Omega^M \times \mathcal{S}_\varepsilon\setminus\mathcal{S}_\delta)\times\{M_0\} \\ \tilde{\tau}\notin(\Omega^M \times \mathcal{S}_\varepsilon)\times\{M_0\}}} \pi(\tilde{\sigma})\tilde{R}(\tilde{\sigma},\tilde{\tau})}{\pi\big((\Omega^M \times \mathcal{S}_\varepsilon) \times \{M_0\}\big)} \\
&\leq \frac{\sum_{\tilde{\sigma}\in(\Omega^M \times \mathcal{S}_\varepsilon\setminus\mathcal{S}_\delta)\times\{M_0\}} \pi(\tilde{\sigma})}{\pi\big((\Omega^M \times \mathcal{S}_\varepsilon) \times \{M_0\}\big)} \qquad (150) \\
&= \frac{\pi_M(\mathcal{S}_\varepsilon \setminus \mathcal{S}_\delta)}{\pi_M(\mathcal{S}_\varepsilon)} \\
&< e^{-c'N}
\end{aligned}$$

Here (150) holds because of Lemma 7.5: The first Equi-Energy jump started in the circle $B_\delta(a_{\frac{1}{3}})$ will not leave $B_{\delta''}(a_{\frac{1}{3}})$, the following metropolis update will not leave $B_{\delta'}(a_{\frac{1}{3}})$ thus the second Equi-Energy jump cannot leave $B_\varepsilon(a_{\frac{1}{3}})$. □

Using Theorem 7.7 concludes the proof for Theorem 7.2, as the chain $\tilde{R}$ is reversible once the matrix has been filled.

Appendix

1. Appendix to the BEG Model

1.1. Analysis of f_β. This appendix contains a detailed analysis of the function f_β given in (111). The first result is needed for the slow convergence case.

LEMMA .8. *There exists an $\varepsilon_0 > 0$ such that for any $0 < \varepsilon \leq \varepsilon_0$ on the set*

$$\mathcal{N} = \left\{\sigma \middle| |S_N(\sigma)| \leq N \cdot \varepsilon\right\}$$

as defined in (110) *the free energy f_β is unimodal for all β.*

PROOF. The claim is true, if we find an $\varepsilon_0 > 0$ such that

$$f_\beta((a_{-1}, a_0, a_1)) \text{ is unimodal on } |a_{-1} - a_1| < \varepsilon_0.$$

Consider

$$-f_\beta(a_1, a_0, a_1) = 2\beta a_1 + 2a_1 \log(a_1) + (1 - 2a_1)\log(1 - 2a_1) \tag{151}$$

$$-f_\beta(a_1, a_0, a_1)' = 2\beta - \log\left(\frac{1}{a_1} - 2\right) \tag{152}$$

which tells us, that there is exactly one mode on the $a_1 = a_{-1}, a_0 = 1 - 2a_1$ line. As f_β is smooth this generalizes for all lines $a_1 = a_{-1} + 2\varepsilon_0$ for sufficiently small ε_0. This yields the desired result by using Theorem .9 as all that could happen, are maxima on the boundary. □

THEOREM .9. *f_β has at most three local maxima on*

$$\Upsilon_\infty := \{(a_{-1}, a_0, a_1) \in \mathbb{R}^3_+ : \sum_{i=-1}^{1} a_i = 1\}.$$

There are no further maxima on the boundary of Υ_∞.

PROOF. We first change coordinates. Let $r = \frac{x}{x+z}$ and $t = x + z$. Then the mapping is

$$T : \Upsilon_\infty \to (0,1)^2 \text{ with } (a_{-1}, a_0, a_1) \mapsto (r, t)$$

bijective. Hence, instead of investigating the maxima of f_β, we can analyze the minima of $F(r,t) := F_\beta(r,t) := -f_\beta \circ T^{-1}(r,t)$. Here $F : (0,1)^2 \to \mathbb{R}$ is given by

$$F(r,t) = \beta t(1 - Kt(1-2r)^2) + tH(r) + H(t),$$

with $H(r) = r\log + (1-r)\log(1-r)$.

Minimums at the boundary: For fixed $r \in [0,1]$ the function F is the sum of a polynomial in t and the entropy function $H(t)$. Now $H(t)$ is steep at $t = 0$ and $t = 1$, hence there are no local minima in these points.

If, on the other hand, $t \in (0,1)$ is fixed, the same argument yields that there are no local minima in $r = 0$ and $r = 1$, either.

Global and local Minimums: We take derivatives of F for $r, t \in (0,1)$.

$$\begin{aligned}
\partial_r F(r,t) &= 4\beta K t^2(1-2r) + t\log\tfrac{r}{1-r} \\
\partial_t F(r,t) &= \beta - 2\beta K t(1-2r)^2 + H(r) + \log\tfrac{t}{1-t} \\
\partial_r^2 F(r,t) &= -8\beta K t^2 + \tfrac{t}{r(1-r)} \\
\partial_{rt}^2 F(r,t) &= 8\beta K t(1-2r) + \log\tfrac{r}{1-r} \\
\partial_t^2 F(r,t) &= -2\beta K(1-2r)^2 + \tfrac{1}{t(1-t)}
\end{aligned}$$

Hence the equations for potential minima are

$$4\beta K t(2r-1) = \log\frac{r}{1-r} \tag{153}$$

$$\frac{1}{t} - 1 = e^{\beta}\cdot\sqrt{r(1-r)}\ , \tag{154}$$

where we have used (153) to solve $\partial_t F = 0$ and obtain (154). Taking the Taylor expansion of F in a critical point (r_0, t_0) up to second order we see that

$$F(r,t) = F(r_0,t_0) + \frac{1}{2}A$$

where

$$A = \partial_r^2 F(r_0,t_0)(r-r_0)^2 + 2\partial_{rt}^2 F(r_0,t_0)(r-r_0)(t-t_0) + \partial_t^2 F(r_0,t_0)(t-t_0)^2.$$

Putting $w := \sqrt{r_0(1-r_0)}$ we see that $t_0 = (1+e^{\beta}w)^{-1}$ and therefore

$$\partial_r^2 F(r_0,t_0) = \frac{t_0^2}{w^2}(1 + e^{\beta}w - 8\beta K w^2). \tag{155}$$

Due to (153) we have in critical points (r_0, t_0)

$$\partial_{rt}^2 F(r_0,t_0) = 4\beta K t_0(1-2r_0)$$

and the determinant of the Hessian M in (r_0, t_0) is given by

$$\det M = \Big(\frac{t_0}{w^2} - 8\beta K t_0^2\Big)\Big(\frac{1}{t_0(1-t_0)} - 2\beta K(1-4w^2)\Big) - (4\beta K t_0)^2(1-4w^2).$$

This can be simplified to

$$\begin{aligned}
\det M &= \Big(\frac{1}{w^2} - 8\beta K t_0\Big)\frac{1}{1-t_0} - 2\beta K\frac{t_0}{w^2}(1-4w^2) \\
&= \frac{1 - 2\beta K t_0 + 2\beta K t_0^2(1-4w^2)}{w^2(1-t_0)}
\end{aligned}$$

and by replacing t_0 we obtain:

$$\begin{aligned}\det M &= \frac{(1+e^\beta w)^2 - 2\beta K(1+e^\beta w) + 2\beta K(1-4w^2)}{w^2(1-t_0)(1+e^\beta w)^2} \\ &= \frac{1+2e^\beta w(1-\beta K) + w^2(e^{2\beta} - 8\beta K)}{w^2(1-t_0)(1+e^\beta w)^2}. \qquad (156)\end{aligned}$$

Note that the sign of $\det M$ is determined by the sign of the nominator, which is important, since M is positive definite in (r_0, t_0), if $\partial_r^2 F > 0$ and $\det M > 0$ in that point.

Investigating which points are critical, we see the following

(1) Obviously, $r_0 = \frac{1}{2}$, $t_0 = \frac{2}{2+e^\beta}$ is critical. Here $\partial_r^2 F(r_0, t_0) = 2t_0^2(2 + e^\beta - 4\beta K)$ and hence

$$A = 2t_0^2(2 + e^\beta - 4\beta K)(r - r_0)^2 + \frac{1}{t_0(1-t_0)}(t - t_0)^2.$$

Thus there is a local minimum of F in (r_0, t_0), if and only if $4\beta K \leq 2 + e^\beta$. If $4\beta K > 2 + e^\beta$, (r_0, t_0) as defined above is not an extremal point.

(2) For $r \neq \frac{1}{2}$, we only consider $r \in I := (\frac{1}{2}, 1)$, since F is symmetric in r around $\frac{1}{2}$.

Combining (153) and (154) we see that a necessary condition for (r, t) to be a local minimum is

$$h(r) := \log \frac{r}{1-r} = \frac{4\beta K(2r-1)}{1 + e^\beta \sqrt{r(1-r)}} := \phi(r), \qquad (157)$$

which we will investigate for solutions in I. Let $w(r) := \sqrt{r(1-r)}$. We compute

$$h'(r) = \frac{1}{r} + \frac{1}{1-r} = \frac{1}{w^2(r)}$$
$$h''(r) = -\frac{1}{r^2} + \frac{1}{(1-r)^2} = \frac{2r-1}{w^4(r)}$$

and

$$\phi'(r) = 4\beta K \frac{2 + 2e^\beta w(r) - (2r-1)e^\beta \frac{(1-2r)}{2w(r)}}{(1 + e^\beta w(r))^2} = 2\beta K \frac{4w(r) + e^\beta}{w(r)(1 + e^\beta w(r))^2}$$

and eventually

$$\begin{aligned}\phi''(r) &= 2\beta K \frac{4w'(r)w(r)(1+e^\beta w(r))^2 - (4w(r)+e^\beta)[w(r)(1+e^\beta w(r))^2]'}{w^2(r)(1+e^\beta w(r))^4} \\ &= 2\beta K w'(r) \frac{4w(r)(1+e^\beta w(r)) - (4w(r)+e^\beta)[1 + e^\beta w(r) + 2w(r)e^\beta]}{w^2(r)(1+e^\beta w(r))^3} \\ &= \beta K e^\beta \frac{(2r-1)(8w^2(r) + 3e^\beta w(r) + 1)}{w^3(r)(1+e^\beta w(r))^3}.\end{aligned}$$

Now $h'(r) \lessgtr \phi'(r)$ implies

$$(e^{2\beta} - 8\beta K)w^2(r) + 2e^{\beta}(1 - \beta K)w(r) + 1 \lessgtr 0. \tag{158}$$

Hence there are at most two solutions $r_1, r_2 \in I$ with $\phi' = h'$, because w is injective on I. Therefore, according to Rolle's theorem also the equation $\phi = h$ has at most two further solutions in I (next to $r = 1/2$). Moreover, we see that the left hand side of (158) equals the nominator of $\det M$ in (156). In a critical point we thus have $h' < \phi'$ (or $h' > \phi'$, respectively) if and only if in this point it holds $\det M < 0$ (or $\det M > 0$, respectively).

Again we distinguish different cases:

If $4\beta K > 2 + e^{\beta}$, then $\phi'(1/2) > h'(1/2)$ and thus $\phi > h$ on $(1/2, 1/2+\delta)$ for an appropriate $\delta > 0$. Now, close to $r = 1$ we always have $\phi < h$, which means, there is at least one solution $\phi = h$ in I. However, there cannot be two such solutions: If there were $\frac{1}{2} < r_1 < r_2 < 1$ with $\phi = h$, then $\phi - h$ cannot change sign in both solutions, otherwise we would have $\phi > h$ also in a right neighborhood of r_2 and we would need a third solution r_3 to the right of r_2, in contradiction to the above conclusion. If, on the other hand, $\phi - h$ cannot change sign in both solutions, then at least one of r_1 and r_2 also solves $\phi' = h'$. But this again leads to a contradiction. Again using Rolle's theorem we see that $\phi = h$ for $\frac{1}{2} < r_1 < r_2$ implies that there exist ξ_1, ξ_2 with $\phi' = h'$ and

$$\frac{1}{2} < \xi_1 < r_1 < \xi_2 < r_2$$

and there cannot be more than two solutions of $\phi' = h'$.

Hence there is exactly one solution $r_1 \in I$ and from (154) one obtains the corresponding t_1, such that $(r_1, t_1), (1 - r_1, t_1)$ and (r_0, t_0) are the only critical points of F. However, we already know that here we have $4\beta K > 2 + e^{\beta}$ and hence (r_0, t_0) is not a minimum of F. Moreover, minima at the boundary do not exist. But F is continuous on $[0, 1]^2$, therefore has a minimum, thus the points (r_1, t_1) and $(1 - r_1, t_1)$ are global minima.

If, on the other hand $4\beta K = 2 + e^{\beta}$ and $e^{\beta} > 4$, then $\phi'(1/2) = h'(1/2)$ and of course $\phi''(1/2) = h''(1/2) = 0$, however we still have $\phi'''(1/2) > h'''(1/2)$, hence again $\phi > h$ on $(1/2, 1/2 + \delta)$ for an appropriate $\delta > 0$. $\phi'''(1/2) > h'''(1/2)$ can be seen as follows: Write

$$v(u) := \frac{8u^2 + 3e^{\beta}u + 1}{(u + e^{\beta}u^2)^3}.$$

Then $\phi''(r) = \beta K e^\beta (2r-1) v \circ w(r)$ and hence

$$\phi'''(r) = \beta K e^\beta (2v \circ w(r) - (2r-1)^2 \frac{1}{2w(r)} v' \circ w(r)). \tag{159}$$

Thus

$$\phi'''(1/2) = \frac{1}{2}(2+e^\beta)e^\beta v(1/2) = 48 \frac{e^\beta}{2+e^\beta}.$$

Due to $h'''(1/2) = 32$ we have $\phi'''(1/2) > h'''(1/2)$ if and only if $e^\beta > 4$.

Analogously to our arguments above we see that there is only one solution $r_1 \in I$ of $\phi = h$, and again the corresponding t_1 can be computed from (154). Indeed there is a local minimum of F in (r_1, t_1) and $(1-r_1, t_1)$. This can be seen by showing that the Hessian is positive definite. However, as this is not part of our assertion, we will refrain from doing so.

If, finally $4\beta K = 2 + e^\beta$ and $e^\beta \leq 4$, then $\phi'(1/2) = h'(1/2)$ and $\phi'''(1/2) \leq h'''(1/2)$ and $\phi^{(5)}(1/2) < h^{(5)}(1/2)$, such that again $\phi < h$ on $(1/2, 1/2+\delta)$ for an appropriate $\delta > 0$.

For $\phi^{(5)}(1/2) < h^{(5)}(1/2)$ one argues: Because of (159) we have

$$\begin{aligned}
\phi^{(5)}(1/2) &= \beta K e^\beta (2(v \circ w)''(1/2) - 8v'(1/2)) \\
&= 2\beta K e^\beta \Big(((v' \circ w) \cdot w')'(1/2) - 4v'(1/2) \Big) \\
&= 2\beta K e^\beta \Big(- \frac{v' \circ w}{w}(1/2) - 4v'(1/2) \Big) \\
&= -12 \beta K e^\beta v'(1/2)
\end{aligned}$$

and

$$\begin{aligned}
v'(1/2) &= \frac{(8+3e^\beta)\frac{1}{4}(2+e^\beta) - 3(1+e^\beta)(3+\frac{3}{2}e^\beta)}{(\frac{1}{2}+\frac{1}{4}e^\beta)^4} \\
&= -320 \frac{2+3e^\beta}{(2+e^\beta)^3},
\end{aligned}$$

thus

$$\phi^{(5)}(1/2) = 960 e^\beta \frac{2+3e^\beta}{(2+e^\beta)^2}.$$

Because of $h^{(5)}(1/2) = 4! \cdot 2^6$ one has $\phi^{(5)}(1/2) < h^{(5)}(1/2)$ if and only if $5e^\beta(2+3e^\beta) < 8(2+e^\beta)^2$, thus $7e^{2\beta} - 22e^\beta - 32 < 0$ and this is true for all $0 < e^\beta \leq 4$.

The same is of course also true, when $4\beta K < 2 + e^\beta$, since then we already have $\phi'(1/2) < h'(1/2)$.

Summarizing we see that in all possible cases we have at most three local minima of F and none at the boundary. Of course, we could discuss how many minima there are exactly

in certain cases. However, we will refrain from doing so, since this is not needed.

□

Bibliography

[1] D. J. Aldous and J. A. Fill. *Reversible Markov Chains and Random Walks on Graphs*. Book in preparation, `http://www.stat.berkeley.edu/~aldous/book.html`, 200X.

[2] N. Bhatnagar and D. Randall. Torpid mixing of simulated tempering on the Potts model. In *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 478–487 (electronic), New York, 2004. ACM.

[3] A. Bovier. *Statistical Mechanics of Disordered Systems - A Mathematical Perspective*. Cambridge Series in Statistical and Probabilistic Mathematics, 2006.

[4] A. Bovier and I. Kurkova. Derrida's generalized random energy models I: Poisson cascades and extremal processes. 2004.

[5] A. Bovier and I. Kurkova. Derrida's generalized random energy models. II: Gibbs measures and probability cascades. 2004.

[6] A. Bovier, I. Kurkova, and M. Löwe. Fluctuations of the free energy in the REM and the p-spin SK models. *Ann. Probab.*, 30(2):605–651, 2002.

[7] B. Derrida. Random-energy model: an exactly solvable model of disordered systems. *Phys. Rev. B (3)*, 24(5):2613–2626, 1981.

[8] B. Derrida. A generalization of the random energy model which includes correlations between energies. *J. Physique Lett.*, 46(9):401–407, 1985.

[9] P. Diaconis and L. Saloff-Coste. Comparison theorems for reversible Markov chains. *Ann. Appl. Probab.*, 3(3):696–730, 1993.

[10] P. Diaconis and L. Saloff-Coste. Logarithmic Sobolev inequalities for finite Markov chains. *Ann. Appl. Probab.*, 6(3):695–750, 1996.

[11] P. Diaconis and D. Stroock. Geometric bounds for eigenvalues of Markov chains. *Ann. Appl. Probab.*, 1(1):36–61, 1991.

[12] M. Ebbers. Der Swapping Algorithmus im Curie-Weiss und im Potts Modell. Master's thesis, Westfälische Wilhelms-Universität Münster, 2007.

[13] T. Eisele and R. S. Ellis. Multiple phase transitions in the generalized Curie-Weiss model. *J. Statist. Phys.*, 52(1-2):161–202, 1988.

[14] R. S. Ellis, P. T. Otto, and H. Touchette. Analysis of phase transitions in the mean-field Blume-Emery-Griffiths model. *Ann. Appl. Probab.*, 15(3):2203–2254, 2005.

[15] R. S. Ellis, H. Touchette, and B. Turkington. Thermodynamic versus statistical nonequivalence of ensembles for the mean-field blume-emery-griffiths model. *Physica A: Statistical and Theoretical Physics*, 335(3-4):518 – 538, 2004.

[16] R. S. Ellis and K. Wang. Limit theorems for the empirical vector of the Curie-Weiss-Potts model. *Stochastic Process. Appl.*, 35(1):59–79, 1990.

[17] Ellis, Richard S. and Haven, Kyle and Turkington, Bruce. Large deviation principles and complete equivalence and nonequivalence results for pure and mixed ensembles. *J. Stat. Phys.*, 101(5-6):999–1064, 2000.

[18] L. Fontes, M. Isopi, Y. Kohayakawa, and P. Picco. The spectral gap of the REM under Metropolis dynamics. *Ann. Appl. Probab.*, 8(3):917–943, 1998.

[19] C. J. Geyer. Markov chain monte carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of 23rd Symposium on the Interface Interface Foundation*, pages 156–163. Fairfax Station, 1991.

[20] C. J. Geyer and E. A. Thompson. Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Am. Stat. Assoc.*, 90(431):909–920, 1995.

[21] V. K. Gore and M. R. Jerrum. The Swendsen-Wang process does not always mix rapidly. *J. Statist. Phys.*, 97(1-2):67–86, 1999.

[22] O. Häggström. *Finite Markov chains and algorithmic applications.* London Mathematical Society Student Texts. 52. Cambridge: Cambridge University Press. ix, 114 p. 14.95; $ 21.00/pbk; 40.00; $ 60.00/hbk, 2002.

[23] M. Jerrum and A. Sinclair. Approximate counting, uniform generation and rapidly mixing Markov chains. *Inform. and Comput.*, 82(1):93–133, 1989.

[24] S. C. Kou, Q. Zhou, and W. H. Wong. Equi-energy sampler with applications in statistical inference and statistical mechanics. *Ann. Statist.*, 34(4):1581–1652, 2006. With discussions and a rejoinder by the authors.

[25] D. A. Levin, Y. Peres, and E. L. Wilmer. *Markov chains and mixing times.* American Mathematical Society, Providence, RI, 2009. With a chapter by James G. Propp and David B. Wilson.

[26] M. Löwe and F. Vermet. The swapping algorithm for the Hopfield model with two patterns. *Stochastic Process. Appl.*, 119(10):3471–3493, 2009.

[27] N. Madras and M. Piccioni. Importance sampling for families of distributions. *Ann. Appl. Prob.*, 9(4):1202–1225, 1999.

[28] N. Madras and D. Randall. Markov chain decomposition for convergence rate analysis. *Ann. Appl. Probab.*, 12(2):581–606, 2002.

[29] N. Madras and Z. Zheng. On the swapping algorithm. *Random Struct. Algorithms*, 22(1):66–97, 2003.

[30] M. Mézard, G. Parisi, and M. A. Virasoro. *Spin glass theory and beyond*, volume 9 of *World Scientific Lecture Notes in Physics*. World Scientific Publishing Co. Inc., Teaneck, NJ, 1987.

[31] A. Sinclair. *Algorithms for random generation and counting: a Markov chain approach.* Birkhauser Verlag, 1993.

[32] D. B. Woodard, S. C. Schmidler, and M. Huber. Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. *Ann. Appl. Probab.*, 19(2):617–640, 2009.

[33] D. B. Woodard, S. C. Schmidler, and M. Huber. Sufficient conditions for torpid mixing of parallel and simulated tempering. *Electron. J. Probab.*, 14:no. 29, 780–804, 2009.

[34] Z. Zheng. *Analysis of swapping and tempering Monte Carlo algorithms.* PhD thesis, York University Ontario, 1999.

[35] Z. Zheng. On swapping and simulated tempering algorithms. *Stochastic Process. Appl.*, 104(1):131–154, 2003.