Theoretical Computer Science
doi:10.1016/j.tcs.2015.05.019

A Bound for the Convergence Rate of Parallel Tempering for Sampling Restricted Boltzmann Machines

Asja Fischer[a,b], Christian Igel[b]

[a]*Institut für Neuroinformatik, Ruhr-Universität Bochum, 44780 Bochum, Germany*
[b]*Department of Computer Science, University of Copenhagen, 2100 Copenhagen, Denmark*

Abstract

Sampling from restricted Boltzmann machines (RBMs) is done by Markov chain Monte Carlo (MCMC) methods. The faster the convergence of the Markov chain, the more efficiently can high quality samples be obtained. This is also important for robust training of RBMs, which usually relies on sampling. Parallel tempering (PT), an MCMC method that maintains several replicas of the original chain at higher temperatures, has been successfully applied for RBM training. We present the first analysis of the convergence rate of PT for sampling from binary RBMs. The resulting bound on the rate of convergence of the PT Markov chain shows an exponential dependency on the size of one layer and the absolute values of the RBM parameters. It is minimized by a uniform spacing of the inverse temperatures, which is often used in practice. Similar as in the derivation of bounds on the approximation error for contrastive divergence learning, our bound on the mixing time implies an upper bound on the error of the gradient approximation when the method is used for RBM training.

Keywords: restricted Boltzmann machines, parallel tempering, mixing time

1. Introduction

Restricted Boltzmann machines (RBMs) are probabilistic graphical models corresponding to stochastic neural networks [1, 2] (see [3] for a recent review). They are applied in many machine learning tasks, notably they serve as building blocks of deep belief networks [4]. Markov chain Monte

Carlo (MCMC) methods are used to sample from RBMs, and chains that quickly converge to their stationary distribution are desirable to efficiently get high quality samples. Adaptation of the RBM model parameters typically corresponds to gradient-based likelihood maximization given training data. As computing the exact gradient is usually computationally not tractable, sampling-based methods are employed to approximate the likelihood gradient. It has been shown that inaccurate approximations can deteriorate the learning process (e.g., [5]), and for the most popular learning scheme *contrastive divergence learning* (CD, [2]) the approximation quality has been analyzed [6, 7]. The quality of the approximation depends, among other things, on how quickly the Markov chain approaches the stationary distribution, that is, on its mixing rate.

To improve RBM learning, *parallel tempering* (PT, [8]) has successfully been used as a sampling method in RBM training [9, 10, 11, 3]. Parallel tempering introduces supplementary Gibbs chains that sample from smoothed replicas of the original distribution—with the goal of improving the mixing rate. However, so far there exist no published attempts to analyze the mixing rate of PT applied to RBMs. Based on the work by Woodard et al. [12], we provide the first such analysis. After introducing the basic concepts, section 3 states our main result. Section 4 summarizes general theorems required for our proof in section 5, which is followed by a discussion and our conclusions.

2. Background

In the following, we will give a brief introduction to RBMs and the relation between the mixing rate and the spectral gap of a Markov chain. Afterwards we will describe the parallel tempering algorithm and its application to sampling from RBMs.

2.1. Restricted Boltzmann machines

Restricted Boltzmann machines (RBMs) are probabilistic undirected graphical models (Markov random fields). Their structure is a bipartite graph connecting a set of m visible random variables $\boldsymbol{V} = (V_1, V_2, \ldots, V_m)$ modeling observations to n hidden (latent) random variables $\boldsymbol{H} = (H_1, H_2, \ldots, H_n)$ capturing dependencies between the visible variables. In binary RBMs the state space of one single variable is given by $\Omega = \{0, 1\}$ and accordingly $(\boldsymbol{V}, \boldsymbol{H}) \in \{0, 1\}^{m+n}$. The joint probability distribution of $(\boldsymbol{V}, \boldsymbol{H})$ is given by the Gibbs distribution

$$\pi(\boldsymbol{v},\boldsymbol{h}) = \frac{\exp(-E(\boldsymbol{v},\boldsymbol{h}))}{Z} \quad , \tag{1}$$

with energy

$$E(\boldsymbol{v},\boldsymbol{h}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} h_i w_{ij} v_j - \sum_{j=1}^{m} b_j v_j - \sum_{i=1}^{n} c_i h_i$$

and real-valued connection weights w_{ij} and bias parameters b_j and c_i for $i \in \{1,\dots,n\}$ and $j \in \{1,\dots,m\}$. The normalization constant Z, also called the partition function, is given by $Z = \sum_{\boldsymbol{v},\boldsymbol{h}} \exp(-E(\boldsymbol{v},\boldsymbol{h}))$.

Training an RBM means adapting its parameters such that the distribution of $\boldsymbol{V}$ models a distribution underlying some observed data. In practice, this training corresponds to performing stochastic gradient ascent on the log-likelihood of the weight and bias parameters given sample (training) data. The gradient of the log-likelihood given a single training sample $\boldsymbol{v}_{\text{train}}$ is given by

$$\frac{\partial \ln \pi(\boldsymbol{v}_{\text{train}} \,|\, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -\sum_{\boldsymbol{h}} \pi(\boldsymbol{h} \,|\, \boldsymbol{v}_{\text{train}}) \frac{\partial E(\boldsymbol{v}_{\text{train}},\boldsymbol{h})}{\partial \boldsymbol{\theta}} + \sum_{\boldsymbol{v},\boldsymbol{h}} \pi(\boldsymbol{v},\boldsymbol{h}) \frac{\partial E(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}} \quad , \tag{2}$$

where $\boldsymbol{\theta}$ is the vector collecting all parameters. Since the expectation under the model distribution in the second term on the right hand side can not be computed efficiently (it is exponential in $\min(n,m)$), it is approximated by MCMC methods in RBM training algorithms. Typically, the expected value under the model distribution is approximated by $\sum\limits_{\boldsymbol{h}} \pi(\boldsymbol{h} \,|\, \boldsymbol{v}^{(k)}) \frac{\partial E(\boldsymbol{v}^{(k)},\boldsymbol{h})}{\partial \boldsymbol{\theta}}$ given a sample $\boldsymbol{v}^{(k)}$ obtained by running a Markov chain for k steps. Alternatively, we can consider $\sum\limits_{\boldsymbol{v}} \pi(\boldsymbol{v} \,|\, \boldsymbol{h}^{(k)}) \frac{\partial E(\boldsymbol{v},\boldsymbol{h}^{(k)})}{\partial \boldsymbol{\theta}}$ given a sample $\boldsymbol{h}^{(k)}$ to save computation time if $m < n$.

2.2. Mixing rates and the spectral gap

A homogeneous Markov chain on a discrete state space Ω can be described by a transition matrix $P = (p_{\boldsymbol{x},\boldsymbol{y}})_{\boldsymbol{x},\boldsymbol{y}\in\Omega}$, where $p_{\boldsymbol{x},\boldsymbol{y}}$ is the probability to move from $\boldsymbol{x}$ to $\boldsymbol{y}$ in one step of the Markov chain. We also refer to this probability as $P(\boldsymbol{x},\boldsymbol{y})$, and accordingly $P^k(\boldsymbol{x},\boldsymbol{y})$ gives the probability to move from $\boldsymbol{x}$ to $\boldsymbol{y}$ in k steps of the chain.

Markov chain Monte Carlo methods make use of the fact that an ergodic Markov chain on Ω with transition matrix P and equilibrium distribution π satisfies $P^k(\boldsymbol{x}, \boldsymbol{y}) \to \pi(\boldsymbol{y})$ as $k \to \infty$ for all $\boldsymbol{x}, \boldsymbol{y} \in \Omega$. It is important to study the convergence rate of the Markov chain in order to find out how large k has to be to ensure that $P^k(\boldsymbol{x}, \boldsymbol{y})$ is suitably close to $\pi(\boldsymbol{y})$. One way to measure the closeness to stationarity is the total variation distance $\|P^k(\boldsymbol{x}, \cdot) - \pi\| = \frac{1}{2} \sum_{\boldsymbol{y}} |P^k(\boldsymbol{x}, \boldsymbol{y}) - \pi(\boldsymbol{y})|$ for an arbitrary starting state $\boldsymbol{x}$. Reversibility of the chain implies that the eigenvalues of P are real-valued, and we sort them by value $1 = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r \geq -1$. The value $\mathrm{Gap}(P) = 1 - \lambda_2$ is called the *spectral gap* of P, and the eigenvalue with the second largest absolute value $\lambda_{\mathrm{SLEM}} = \max\{\lambda_2, |\lambda_r|\}$ is often referred to as *second largest eigenvalue modulus* (SLEM).

Many bounds on convergence rates are based on the SLEM. Diaconis and Saloff-Coste [13], for example, prove

$$\|P^k(\boldsymbol{x}, \cdot) - \pi\| \leq \frac{1}{2\sqrt{\pi(\boldsymbol{x})}} {\lambda_{\mathrm{SLEM}}}^k \ . \tag{3}$$

If P is positive definite all eigenvalues are non-negative. In this case $\lambda_{\mathrm{SLEM}} = \lambda_2$ and we can replace λ_{SLEM} in (3) (and similar bounds) by $1 - \mathrm{Gap}(P)$. We can make an arbitrary transition matrix Q positive definite by skipping the move each time with probability $\frac{1}{2}$, that is, by considering the transition matrix $P = \frac{1}{2}I + \frac{1}{2}Q$. This makes P in the long run two times slower than Q.

A typical application of MCMC methods (e.g., in the training of RBMs) is to approximate the expected value of a given function under the stationary distribution based on samples obtained after running the Markov chain for k steps. A bound on the chain's convergence rate directly leads to a bound on the bias of this approximation (see Appendix A for a proof).

2.3. Parallel tempering

Consider a multi modal target density π on a state space Ω from which one would like to sample via a Markov chain. The transition steps of the Metropolis-Hastings algorithm move only locally in space so that the resulting Markov chain may move between the modes of π only infrequently. The PT algorithm [8] tries to overcome this by introducing supplementary Markov chains, which are "flattened" or "smoothed" versions of π. These (and the original density) are referred to as *tempered* densities π_t, $t = 0, \ldots, N$ fulfilling $\pi_t(\boldsymbol{z}) \propto \pi(\boldsymbol{z})^{\beta_t}$ for $\boldsymbol{z} \in \Omega$ with *inverse temperatures* β_t. The inverse temperatures satisfy $0 \le \beta_0 < \cdots < \beta_N = 1$. Parallel tempering now constructs a Markov chain on the joint state space $\Omega_{\mathrm{PT}} = \Omega^{N+1}$ of the tempered distributions. The chain takes values $\boldsymbol{x} = (\boldsymbol{x}_{[0]}, \dots, \boldsymbol{x}_{[N]}) \in \Omega_{\mathrm{PT}}$ and converges to the stationary distribution

$$\pi_{\mathrm{PT}}(\boldsymbol{x}) = \prod_{t=0}^{N} \pi_t(\boldsymbol{x}_{[t]}) \ .$$

The PT transition matrix consists out of two components, one for updating the states of the individual tempered chains and one allowing swaps between states of adjacent tempered chains. We will denote the corresponding transition matrices by T and Q, respectively. For the considerations in this paper, we add a probability of $\frac{1}{2}$ of staying in the current state each time T or Q is applied to guarantee positive definiteness and thus positive eigenvalues. We define the PT matrix as $P = QTQ$. A transition matrix constructed like this assures reversibility (unlike TQ) and is used in the analysis by Madras and Zheng [14] and Woodard et al. [12].

Here the matrices T and Q are defined as follows: T chooses $t \in \{0, \dots, N\}$ uniformly and updates the state of π_t based on some transition matrix T_t, which is reversible with respect to π_t. This can, for example, be the Metropolis-Hastings or a Gibbs matrix [15]. Thus, the probability of moving from $\boldsymbol{x} = (\boldsymbol{x}_{[0]}, \dots, \boldsymbol{x}_{[N]}) \in \Omega_{\mathrm{PT}}$ to $\boldsymbol{y} = (\boldsymbol{y}_{[0]}, \dots, \boldsymbol{y}_{[N]}) \in \Omega_{\mathrm{PT}}$ under T is given by

$$T(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2(N+1)} \sum_{t=0}^{N} T_t(\boldsymbol{x}_{[t]}, \boldsymbol{y}_{[t]}) I_{\{\boldsymbol{x}_{[-t]}\}}(\boldsymbol{y}_{[-t]}) + \frac{1}{2} I_{\{\boldsymbol{x}\}}(\boldsymbol{y}) \ ,$$

where $\boldsymbol{x}_{[-t]} = (\boldsymbol{x}_{[0]}, \dots, \boldsymbol{x}_{[t-1]}, \boldsymbol{x}_{[t+1]}, \dots, \boldsymbol{x}_{[N]}) \in \Omega^N$ collects all components of $\boldsymbol{x}$ except the t-th one and, for any set B, $I_B(\boldsymbol{y}) = 1$ if $\boldsymbol{y} \in B$ and $I_B(\boldsymbol{y}) = 0$, otherwise. The $\frac{1}{2} I_{\{\boldsymbol{x}\}}(\boldsymbol{y})$ ensures non-negative eigenvalues as discussed above.

The swapping matrix Q proposes to swap state $\boldsymbol{x}_{[t]}$ and $\boldsymbol{x}_{[t+1]}$ by sampling t uniformly from $\{0, \dots, N-1\}$. The proposed swap is accepted with the Metropolis probability

$$a(\boldsymbol{x}, t) = \min \left\{ 1, \frac{\pi_t(\boldsymbol{x}_{[t+1]}) \pi_{t+1}(\boldsymbol{x}_{[t]})}{\pi_t(\boldsymbol{x}_{[t]}) \pi_{t+1}(\boldsymbol{x}_{[t+1]})} \right\} \ , \tag{4}$$

which guarantees detailed balance. So the transition probability from $\boldsymbol{x} = (\boldsymbol{x}_{[0]}, \dots, \boldsymbol{x}_{[N]})$ to $\boldsymbol{y} = (\boldsymbol{x}_{[0]}, \dots, \boldsymbol{x}_{[t+1]}, \boldsymbol{x}_{[t]}, \dots \boldsymbol{x}_{[N]})$ is given by

$$Q(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2N} a(\boldsymbol{x}, t) \ ,$$

the probability to stay in $\boldsymbol{x}$ is

$$Q(\boldsymbol{x}, \boldsymbol{x}) = 1 - \sum_{t=0}^{N-1} \frac{1}{2N} a(\boldsymbol{x}, t) \ ,$$

and the transition probability between all other states is 0.

By construction both T and Q are reversible with respect to π_{PT} and strictly positive definite due to their $\frac{1}{2}$ holding probability. It follows from the definition of P that both properties also hold for P. Thus, P has only positive real-valued eigenvalues and the convergence of the PT chain to the stationary distribution π_{PT} can be bounded in terms of (3) by replacing λ_{SLEM} by $1 - \alpha$, where α is a lower bound for $\mathrm{Gap}(P)$.

2.4. Training RBMs with PT

The PT algorithm is often the sampling procedure of choice for RBM learning algorithms to approximate the gradient of the log-likelihood given in equation (2)[10, 11]. In this setting, the tempered distributions are defined as

$$\pi_t(\boldsymbol{v}, \boldsymbol{h}) = \frac{\exp(-E(\boldsymbol{v}, \boldsymbol{h})\beta_t)}{Z_t} \tag{5}$$

with $Z_t = \sum_{\boldsymbol{v},\boldsymbol{h}} \exp(-E(\boldsymbol{v}, \boldsymbol{h})\beta_t)$ and β_0 is typically set to 0, which we will also assume in our analysis. The transition matrices T_t for the single tempered chains are set to a 'randomized' version of blockwise Gibbs sampling: we choose to update either $\boldsymbol{V}$ or $\boldsymbol{H}$ with equal probability and then update all neurons in this layer simultaneously. The random choice of the layer leads to reversibility of T_t.

In this paper, we want to find a lower bound for the spectral gap of PT sampling as defined above applied to RBMs. In the PT algorithm most often used for RBM training in practice the transition operator differs from the one analyzed in this paper. It consists out of an update step followed by a swapping step. During the update step, one step of blockwise Gibbs sampling is performed in all tempered chains in parallel, where one step of blockwise Gibbs sampling corresponds to sampling a new state $\boldsymbol{h}'$ of $\boldsymbol{H}$ based on the conditional probability $\pi(\boldsymbol{h}'|\boldsymbol{v})$ given the current state of the visible variables $\boldsymbol{v}$, followed by sampling a new state $\boldsymbol{v}'$ of $\boldsymbol{V}$ based on $\pi(\boldsymbol{v}'|\boldsymbol{h}')$. During the swapping step, the states at all temperatures may be swapped based on the Metropolis probability $a(\boldsymbol{x}, t)$ as defined in (4), with $\boldsymbol{x} = (\boldsymbol{v}, \boldsymbol{h})$. The swapping is often organized in two substeps, where even temperatures are considered in the first and odd temperatures in the second substep. While this approach seems to work in practice, the corresponding transition matrix need not be reversible, which makes it difficult to obtain theoretical results on its mixing behavior, see our discussion in section 6.

3. Main Result

Let π denote the Gibbs distribution of an RBM as defined in (1) and for $0 = \beta_0 < \cdots < \beta_N = 1$ let π_t be the corresponding tempered Gibbs distribution with inverse temperature β_t as given in (5). Then the stationary distribution of PT sampling for RBMs as defined above is given by $\pi_{\mathrm{PT}}((\boldsymbol{x}_{[0]}, \ldots, \boldsymbol{x}_{[N]})) = \prod_{t=0}^{N} \pi_t(\boldsymbol{x}_{[t]})$ and for the spectral gap of the corresponding transition matrix it holds:

Theorem 1. *For the PT transition matrix P of an RBM with m visible and n hidden variables*

$$\begin{aligned}\mathrm{Gap}(P) \geq \min \Bigg\{ & \frac{\exp(-\Delta_c)}{2^{n+10}(N+1)^4} \prod_{j=1}^{m} f_j^2, \frac{\exp(-\Delta_c - \max_t(\beta_{t+1} - \beta_t)\Delta)}{2^8(N+1)^4} \prod_{j=1}^{m} f_j^2 \\ & \frac{\exp(-\Delta_b)}{2^{m+10}(N+1)^4} \prod_{i=1}^{n} g_i^2, \frac{\exp(-\Delta_b - \max_t(\beta_{t+1} - \beta_t)\Delta)}{2^8(N+1)^4} \prod_{i=1}^{n} g_i^2 \Bigg\} \ .\end{aligned} \tag{6}$$

with

$$\Delta = \sum_{i,j} |w_{ij}| + \sum_j |b_j| + \sum_i |c_i| \ ,$$

being the sum over the absolute values of parameters of the RBM,

$$\Delta_b = \sum_j |b_j| \quad \textit{and} \quad \Delta_c = \sum_i |c_i|$$

being the sum of over the absolute values of the bias parameters of the visible and hidden variables, respectively, and

$$f_j = \frac{\min_{\boldsymbol{h}} \pi(v_j = 1|\boldsymbol{h})}{\max_{\boldsymbol{h}} \pi(v_j = 1|\boldsymbol{h})} \qquad \textit{and} \qquad g_i = \frac{\min_{\boldsymbol{v}} \pi(h_i = 1|\boldsymbol{v})}{\max_{\boldsymbol{v}} \pi(h_i = 1|\boldsymbol{v})}$$

being the fraction of the minimal and maximal activation of the j-th visible and the i-th hidden neuron.

The proof is given in section 5. By combining this result with equation (3), we arrive at an upper bound on the rate of convergence of the PT chain.

Corollary 1. *Let P be the PT transition matrix for an RBM with m visible and n hidden variables and let π_{PT} be the joint distribution over all tempered chains. Then for any starting point of the Markov chain $\boldsymbol{x} = (\boldsymbol{x}_{[0]}, \dots, \boldsymbol{x}_{[N]})$, with $\boldsymbol{x}_{[i]} = (\boldsymbol{v}_{[i]}, \boldsymbol{h}_{[i]})$, the distance in variation to π_{PT} is bounded by*

$$\|P^k(\boldsymbol{x}, \cdot) - \pi_{PT}\| \leq \frac{1}{2\sqrt{\pi_{PT}(\boldsymbol{x})}}(1-\alpha)^k \tag{7}$$

with

$$\alpha = \min\Bigg\{ \frac{\exp(-\Delta_c)}{2^{n+10}(N+1)^4} \prod_{j=1}^{m} f_j^2, \frac{\exp(-\Delta_c - \max_t(\beta_{t+1} - \beta_t)\Delta)}{2^8(N+1)^4} \prod_{j=1}^{m} f_j^2 \\ \frac{\exp(-\Delta_b)}{2^{m+10}(N+1)^4} \prod_{i=1}^{n} g_i^2, \frac{\exp(-\Delta_b - \max_t(\beta_{t+1} - \beta_t)\Delta)}{2^8(N+1)^4} \prod_{i=1}^{n} g_i^2 \Bigg\}$$

and Δ, Δ_b, Δ_c, f_j, and g_i as defined above.

Theorem 1 bounds the gap of the PT product chain. As the original chain (the chain with inverse temperature β_N) never converges slower than this product chain (see Appendix B for a proof), corollary 1 also leads to an upper bound for the convergence rate of the chain of interest. Bounding the product chain leads to the dependency on the number N of supplementary tempered chains. However, to our knowledge all approaches analyzing PT suffer from this drawback (e.g., [12, 14, 16]).

Now, we can make use of the fact that a bound on the convergence rate implies a bound on the bias of an MCMC estimate and obtain a bound on the bias of the PT based estimator when used in RBM training. Recall that for training RBMs the log-likelihood gradient given in equation (2) is approximated by replacing the expected value under the model distribution in the second term, $\sum_{\boldsymbol{v},\boldsymbol{h}} \pi(\boldsymbol{v}, \boldsymbol{h}) \frac{\partial E(\boldsymbol{v},\boldsymbol{h})}{\partial \boldsymbol{\theta}}$, by the value $\sum_{\boldsymbol{h}} \pi(\boldsymbol{h} \mid \boldsymbol{v}^{(k)}) \frac{\partial E(\boldsymbol{v}^{(k)},\boldsymbol{h})}{\partial \boldsymbol{\theta}}$ for a sample $\boldsymbol{v}^{(k)}$ obtained by running the PT-chain for k steps. The value of $\sum_{\boldsymbol{h}} \pi(\boldsymbol{h} \mid \boldsymbol{v}^{(k)}) \frac{\partial E(\boldsymbol{v}^{(k)},\boldsymbol{h})}{\partial \boldsymbol{\theta}}$ is bounded by 1 and thus, using the result in Appendix A with $\max t(\boldsymbol{x}) = 1$, we get the following corollary:

Corollary 2. *Given an RBM with m visible and n hidden variables and a PT chain starting from $\boldsymbol{x} = (\boldsymbol{x}_{[0]}, \dots, \boldsymbol{x}_{[N]})$, with $\boldsymbol{x}_{[i]} = (\boldsymbol{v}_{[i]}, \boldsymbol{h}_{[i]})$. Let $\delta_{PT}(\boldsymbol{x}^{(k)})$ be the approximation of the log-likelihood derivative w.r.t some RBM parameter θ obtained by estimating the second term in equation* (2) *based on a PT-sample $\boldsymbol{x}^{(k)} = (\boldsymbol{x}^{(k)}_{[0]}, \dots, \boldsymbol{x}^{(k)}_{[N]})$, with $\boldsymbol{x}^{(k)}_{[i]} = (\boldsymbol{v}^{(k)}_{[i]}, \boldsymbol{h}^{(k)}_{[i]})$, gained by running the chain for k steps. That is, for a training sample $\boldsymbol{v}_{train}$*

$$\delta_{PT}(\boldsymbol{x}^{(k)}) = -\sum_{\boldsymbol{h}} \pi(\boldsymbol{h} \,|\, \boldsymbol{v}_{train}) \frac{\partial E(\boldsymbol{v}_{train}, \boldsymbol{h})}{\partial \theta} + \sum_{\boldsymbol{h}} \pi(\boldsymbol{h} \,|\, \boldsymbol{v}^{(k)}_{[0]}) \frac{\partial E(\boldsymbol{v}^{(k)}_{[0]}, \boldsymbol{h})}{\partial \theta} \; . \tag{8}$$

Then we can bound the absolute value of the bias of $\delta_{PT}(\boldsymbol{x}^{(k)})$ by

$$\Big| \frac{\partial \ln \pi(\boldsymbol{v}_{train}|\theta)}{\partial \theta} - \sum_{\boldsymbol{x}^{(k)}} P^k(\boldsymbol{x}, \boldsymbol{x}^{(k)}) \delta_{PT}(\boldsymbol{x}^{(k)}) \Big| \leq \frac{1}{\sqrt{\pi_{PT}(\boldsymbol{x})}} (1-\alpha)^k \; , \tag{9}$$

where α is defined as in corollary 1.

Similar considerations lead to an alternative proof of the approximation error of contrastive divergence learning [7] by combining the result in Appendix A with standard bounds on the periodic Gibbs sampler (e.g. [20], page 289).

4. Bounding the Spectral Gap of PT

This section summarizes definitions and results required for the proof of theorem 1. We state important results for bounding the spectral gap of general Markov chains and recall the link between the SLEM and *Dobrushin's coefficient*.

4.1. Definitions

For any transition matrix P reversible with respect to a distribution π and any subset A of the state space Ω of P, the *restriction* of P to A is defined as:

$$P|_A(\boldsymbol{x}, B) = P(\boldsymbol{x}, B) + I_B(\boldsymbol{x}) P(\boldsymbol{x}, A^c) \tag{10}$$

for $\boldsymbol{x} \in A, B \subset A$. In this way transitions that would leave A are prohibited and the probabilities to stay in A are increased correspondingly.

Given a partition $\mathcal{A} = \{A_j : j = 1, \dots, J\}$ of a finite Ω such that

$$\pi(A_j) = \sum_{\boldsymbol{x} \in A_j} \pi(\boldsymbol{x}) > 0$$

for all j, the *projection matrix* of P with respect to $\mathcal{A}$ is given by the transition operator $\bar{P}$ with

$$\bar{P}(i, j) = \frac{1}{\pi(A_i)} \sum_{\boldsymbol{x} \in A_i} \sum_{\boldsymbol{y} \in A_j} \pi(\boldsymbol{x}) P(\boldsymbol{x}, \boldsymbol{y}) \tag{11}$$

for $i, j \in \{1, \dots, J\}$.

4.2. General results

Given a partition $\mathcal{A}$ of the state space as defined above, the following theorem allows to lower bound the spectral gap of the PT transition matrix in terms of the spectral gap of the projection matrix and the minimum over the spectral gaps of the restrictions of P to the subsets A_j.

Theorem 2. *Let P be a transition matrix reversible with respect to a distribution π on a state space Ω. Let $\{A_j : i = 1, \dots, J\}$ be any partition of Ω such that $\pi(A_j) > 0$ for all j. Define $P_{|A_j}$ as in* (10) *and $\bar{P}$ as in* (11)*. If P is nonnegative definite, it holds*

$$\mathrm{Gap}(P) \geq \frac{1}{2} \mathrm{Gap}(\bar{P}) \min_{j=1,\dots,J} \mathrm{Gap}(P_{|A_j}) \ .$$

As Woodard et al. [12] show this theorem can be proven based on the results from Caracciolo et al. [17] (which were first published in [18]) and the results from Madras and Zheng [14].

Often, one wishes to bound the spectral gap of one chain based on the spectral gap of another chain, which, for example, is easier to estimate. In this case the following theorem can be applied. Combining the theorem for the comparison of the Dirichlet-forms of two Markov chains given by Diaconis and Saloff-Coste [19] with Reyleigh's theorem (e.g., [20], p. 205) we get:

Theorem 3. *Let P and Q be transition matrices on a finite state space Ω, reversible with respect to densities π_P and π_Q respectively. Denote by $E_P = \{(\boldsymbol{x}, \boldsymbol{y}) : \pi_P(\boldsymbol{x}) P(\boldsymbol{x}, \boldsymbol{y}) > 0\}$ and $E_Q = \{(\boldsymbol{x}, \boldsymbol{y}) : \pi_Q(\boldsymbol{x}) Q(\boldsymbol{x}, \boldsymbol{y}) > 0\}$ the edge sets of the corresponding transition graphs. For each pair $\boldsymbol{x} \neq \boldsymbol{y}$*

such that $(\boldsymbol{x}, \boldsymbol{y}) \in E_Q$ fix a path $\gamma_{\boldsymbol{x},\boldsymbol{y}} = (\boldsymbol{x} = \boldsymbol{x}^0, \boldsymbol{x}^1, \dots, \boldsymbol{x}^k = \boldsymbol{y})$ of length $|\gamma_{\boldsymbol{x},\boldsymbol{y}}| = k$ such that $(\boldsymbol{x}^i, \boldsymbol{x}^{i+1}) \in E_P$ for $i \in \{0, \dots, k-1\}$ and define

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

Then it holds $\operatorname{Gap}(Q) \leq c \operatorname{Gap}(P)$*.*

If both chains are reversible with respect to the same distribution, the following theorem proven in [12] can be applied:

Theorem 4. *Let P and Q be transition matrices on a state space Ω reversible with respect π. If $Q(\boldsymbol{x}, A \setminus \{\boldsymbol{x}\}) \leq P(\boldsymbol{x}, A \setminus \{\boldsymbol{x}\})$ for every $\boldsymbol{x} \in \Omega$ and every $A \subset \Omega$ then* $\operatorname{Gap}(Q) \leq \operatorname{Gap}(P)$*.*

Recall that the PT transition matrix T combines $N+1$ transition operators $T_0, \dots, T_N$ by taking values $\boldsymbol{x} = (\boldsymbol{x}_{[0]}, \dots, \boldsymbol{x}_{[N]})$ in the joint state space Ω_{PT} and performing a transition by randomly picking $t \in \{0, \dots, N\}$ and sampling a new value $\boldsymbol{x}_{[t]}$ for the t-th chain from T_t. Such a Markov chain is called a product chain. The following theorem deals with product chains and gives the dependency between the spectral gap of the product chain and the spectral gaps of all single chains it subsumes ([21], lemma 3.2):

Theorem 5. *For any natural number N and $t = 0, \dots, N$, let P_t be a π_t-reversible transition matrix on a state space Ω_t. Let P be the transition matrix on $\Omega = \prod_t \Omega_t$ given by*

$$P(\boldsymbol{x}, \boldsymbol{y}) = \sum_{t=0}^{N} b_t P_t(\boldsymbol{x}_{[t]}, \boldsymbol{y}_{[t]}) I_{\{\boldsymbol{x}_{[-t]}\}}(\boldsymbol{y}_{[-t]}) \ \ , \boldsymbol{x}, \boldsymbol{y} \in \Omega$$

for some set of $b_t > 0$ such that $\sum_t b_t = 1$ and where $\boldsymbol{x}_{[t]}$ denotes the t-th component of $\boldsymbol{x}$ (corresponding to the state of the t-th chain) and $\boldsymbol{x}_{[-t]}$ all components except the t-th one. A Markov chain with transition matrix P is called a product chain. It is reversible with respect to $\pi(\boldsymbol{x}) = \prod_t \pi_t(\boldsymbol{x}_{[t]})$ and

$$\operatorname{Gap}(P) = \min_{t=0,\dots,N} b_t \operatorname{Gap}(P_t) \ \ .$$

The next theorem is a well known result that gives an upper bound on the SLEM (e.g., see [20], p. 237).

Theorem 6. *The second largest eigenvalue modulus* λ_{SLEM} *of a transition matrix* $P = (p_{\boldsymbol{x},\boldsymbol{y}})_{\boldsymbol{x},\boldsymbol{y}\in\Omega}$ *can be bounded from above by* Dobrushin's coefficient*:*

$$D(P) = \frac{1}{2}\max_{\boldsymbol{x},\boldsymbol{y}\in\Omega}\sum_{\boldsymbol{z}\in\Omega}|p_{\boldsymbol{x},\boldsymbol{z}} - p_{\boldsymbol{y},\boldsymbol{z}}| = 1 - \min_{\boldsymbol{x},\boldsymbol{y}\in\Omega}\sum_{\boldsymbol{z}\in\Omega}\min\{p_{\boldsymbol{x},\boldsymbol{z}}, p_{\boldsymbol{y},\boldsymbol{z}}\}$$

5. Proof of the main result

For proving the main result we make use of an approach inspired by Woodard et al. [12]. The basic idea behind this approach is to partition the state space of the PT chain into subsets, such that the Markov chain mixes well inside the single subsets at all temperatures and between the subsets at the lowest temperature. To begin with, a suitable partitioning of the RBM state space is needed. For a binary RBM with m visible and n hidden neurons the state space is $\Omega = \{0,1\}^{n+m}$. Let $A = \{A_j : j = 1,\dots,2^n\}$ be the partition of Ω where each subset A_j contains all states having the same state of the hidden neurons, that is, $A_j = \{(\boldsymbol{v},\boldsymbol{h}) \in \Omega | \boldsymbol{h} = \boldsymbol{h}_j\}$ if $\boldsymbol{h}_j$ denotes the j-th state of the hidden random variables. Thus, we get 2^n subsets $A_1,\dots,A_{2^n}$ and each A_j contains 2^m elements. Using this partition, we can now prove theorem 1 based on the results given in the previous section. Our proof can be divided into five steps, where steps 1 and 2 are taken from Woodard et al. [12].

The basic thoughts behind the steps of the proof can be outlined as follows. The gap of the PT transition matrix P depends on (a) how well the single tempered chains mix inside the single subsets $A_1,\dots,A_{2^n}$, (b) how well the chain at the highest temperature (i.e. the chain with inverse temperature β_0) mixes between the subsets, and (c) the mixing properties between the chains at the different temperatures. In step 1, $\mathrm{Gap}(P)$ is bounded in terms of the gaps of two transition matrices, one depending on (a) and the other depending on (b) and (c). The first one is further bounded in steps 2 and 3 and the second one in steps 4. Step 5 puts everything together.

5.1. *Step 1: Bounding* $\mathrm{Gap}(P)$ *in terms of* $\mathrm{Gap}(\bar{P})$ *and* $\mathrm{Gap}(P_{\boldsymbol{\sigma}})$

Consider the state $\boldsymbol{x} = (\boldsymbol{x}_{[0]},\dots,\boldsymbol{x}_{[N]}) \in \Omega^{N+1}$ of the PT chain. Let the signature be the vector $s(\boldsymbol{x}) = (\sigma_0,\dots,\sigma_N)$ with $\sigma_t = j$ if $\boldsymbol{x}_{[t]} \in A_j$ for $t = 0,\dots,N$. Since the partition of Ω consists out of 2^n subsets A_j and we have $N+1$ temperatures, a signature lives in $\Sigma = \{1,\dots,2^n\}^{N+1}$.

For a fixed $\boldsymbol{\sigma} \in \Sigma$ let us now define $\Omega_{\boldsymbol{\sigma}} = \{\boldsymbol{x} \in \Omega^{N+1} : s(\boldsymbol{x}) = \boldsymbol{\sigma}\}$. Then all possible $\boldsymbol{\sigma} \in \Sigma$ induce a partition $\{\Omega_{\boldsymbol{\sigma}}\}_{\boldsymbol{\sigma}\in\Sigma}$ of the PT-state space $\Omega_{\mathrm{PT}} = \Omega^{N+1}$.

Let $P_{\boldsymbol{\sigma}} = P_{\rceil\Omega_{\boldsymbol{\sigma}}}$ now be the restriction of P to $\Omega_{\boldsymbol{\sigma}}$ as defined in equation (10). And let $\bar{P}$ denote the projection matrix of P with respect to $\{\Omega_{\boldsymbol{\sigma}}\}_{\boldsymbol{\sigma}\in\Sigma}$ as defined in (11). Now we can apply theorem 2 and get

$$\operatorname{Gap}(P) \geq \frac{1}{2} \operatorname{Gap}(\bar{P}) \min_{\boldsymbol{\sigma}\in\Sigma} \operatorname{Gap}(P_{\boldsymbol{\sigma}}) \ . \tag{12}$$

5.2. Step 2: Bounding $\operatorname{Gap}(P_{\boldsymbol{\sigma}})$ *in terms of* $\min_{t,j} \operatorname{Gap}(T_t|_{A_j})$

Woodard et al. now proceed by bounding $\operatorname{Gap}(P_{\boldsymbol{\sigma}})$ and $\operatorname{Gap}(\bar{P})$ separately. For bounding $\operatorname{Gap}(P_{\boldsymbol{\sigma}})$ they note that, since $P = QTQ$ and Q has a $\frac{1}{2}$ holding probability, $P_{\boldsymbol{\sigma}}(\boldsymbol{x}, \boldsymbol{y}) \geq \frac{1}{4} T_{\boldsymbol{\sigma}}(\boldsymbol{x}, \boldsymbol{y}) \ \forall \boldsymbol{x}, \boldsymbol{y} \in \Omega_{\boldsymbol{\sigma}}$. With theorem 4 it follows $\operatorname{Gap}(P_{\boldsymbol{\sigma}}) \geq \frac{1}{4} \operatorname{Gap}(T_{\boldsymbol{\sigma}})$. As $T_{\boldsymbol{\sigma}}$ is a product chain, theorem 5 gives

$$\operatorname{Gap}(T_{\boldsymbol{\sigma}}) = \frac{1}{2(N+1)} \min_t \operatorname{Gap}(T_t|_{\sigma_t}) \geq \frac{1}{2(N+1)} \min_{t,j} \operatorname{Gap}(T_t|_{A_j}) \ .$$

Thus, it follows

$$\operatorname{Gap}(P_{\boldsymbol{\sigma}}) \geq \frac{1}{8(N+1)} \min_{t,j} \operatorname{Gap}(T_t|_{A_j}) \ . \tag{13}$$

5.3. Step 3: Bounding $\min_{t,j} \operatorname{Gap}(T_t|_{A_j})$

We will now drive a lower bound for $\min_{t,j} \operatorname{Gap}(T_t|_{A_j})$. The transition matrix T_t analyzed here for (tempered) RBMs corresponds to randomized blockwise Gibbs sampling as described above. The transition probability from $(\boldsymbol{v}, \boldsymbol{h})$ to $(\boldsymbol{v}', \boldsymbol{h})$ is given by

$$T_t((\boldsymbol{v}, \boldsymbol{h})), (\boldsymbol{v}', \boldsymbol{h})) = \frac{1}{2} \pi_t(\boldsymbol{v}'|\boldsymbol{h}) \ ,$$

and the probability to change the state of the hidden variables is accordingly

$$T_t((\boldsymbol{v}, \boldsymbol{h})), (\boldsymbol{v}, \boldsymbol{h}')) = \frac{1}{2} \pi_t(\boldsymbol{h}'|\boldsymbol{v}) \ .$$

Based on (10) for the restriction of T_t to A_j it holds:

$$T_t|_{A_j}((\boldsymbol{v}, \boldsymbol{h}_j), (\boldsymbol{v}', \boldsymbol{h}_j)) = T_t((\boldsymbol{v}, \boldsymbol{h}_j), (\boldsymbol{v}', \boldsymbol{h}_j)) + I_{\{(\boldsymbol{v}, \boldsymbol{h}_j)\}}((\boldsymbol{v}', \boldsymbol{h}_j)) T_t((\boldsymbol{v}, \boldsymbol{h}_j), A_j^c)$$

for $(\boldsymbol{v}, \boldsymbol{h}_j), (\boldsymbol{v}', \boldsymbol{h}_j) \in A_j$. Thus, for $\boldsymbol{v} \neq \boldsymbol{v}'$

$$T_t|_{A_j}((\boldsymbol{v}, \boldsymbol{h}_j), (\boldsymbol{v}', \boldsymbol{h}_j)) = \frac{1}{2}\pi_t(\boldsymbol{v}'|\boldsymbol{h}_j) \ \ ,$$

and the probability to stay in the same state is

$$T_t|_{A_j}((\boldsymbol{v}, \boldsymbol{h}_j), (\boldsymbol{v}, \boldsymbol{h}_j)) \ = \ \frac{1}{2}\pi_t(\boldsymbol{v}|\boldsymbol{h}_j) + \sum_{\boldsymbol{h}} \frac{1}{2}\pi_t(\boldsymbol{h}|\boldsymbol{v}) \ = \ \frac{1}{2}\pi_t(\boldsymbol{v}|\boldsymbol{h}_j) + \frac{1}{2} \ \ .$$

Let us denote the SLEM of $T_t|_{A_j}$ by $\lambda^{T_t}_{\text{SLEM}}$ and the second largest eigenvalue by $\lambda^{T_t}_2$. Based on theorem 6 it holds

$$\text{Gap}(T_t|A_j) = 1 - \lambda^{T_t}_2 \geq 1 - \lambda^{T_t}_{\text{SLEM}} \geq 1 - D(T_t|_{A_j}) \ \ . \tag{14}$$

To upper bound the Dobrushin's coefficient of $T_t|_{A_j}$ first note that for all $\boldsymbol{v}$, $\boldsymbol{v}'$:

$$\begin{aligned}
&\sum_{\hat{\boldsymbol{v}}} \min\{T_t|_{A_j}((\boldsymbol{v}, \boldsymbol{h}_j), (\hat{\boldsymbol{v}}, \boldsymbol{h}_j)), T_t|_{A_j}((\boldsymbol{v}', \boldsymbol{h}_j), (\hat{\boldsymbol{v}}, \boldsymbol{h}_j))\} \\
&\qquad = \sum_{\hat{\boldsymbol{v}}:\hat{\boldsymbol{v}}\neq\boldsymbol{v}\wedge\hat{\boldsymbol{v}}\neq\boldsymbol{v}'} \frac{1}{2}\pi_t(\hat{\boldsymbol{v}}|\boldsymbol{h}_j) + \min\left\{\frac{1}{2}\pi_t(\boldsymbol{v}|\boldsymbol{h}_j), \frac{1}{2}\pi_t(\boldsymbol{v}|\boldsymbol{h}_j) + \frac{1}{2}\right\} \\
&\qquad\qquad + \min\left\{\frac{1}{2}\pi_t(\boldsymbol{v}'|\boldsymbol{h}_j), \frac{1}{2}\pi_t(\boldsymbol{v}'|\boldsymbol{h}_j) + \frac{1}{2}\right\} \\
&\qquad\qquad\qquad = \sum_{\hat{\boldsymbol{v}}} \frac{1}{2}\pi_t(\hat{\boldsymbol{v}}|\boldsymbol{h}_j) = \frac{1}{2} \ \ .
\end{aligned}$$

Now it is easy to see that $D(T_t|_{A_j}) = 1 - \frac{1}{2} = \frac{1}{2}$ and insertion into (14) gives $\text{Gap}(T_t|_{A_j}) \geq 1 - D(T_t|_{A_j}) \geq \frac{1}{2}$. Thus, we finally get by insertion into equation (13)

$$\text{Gap}(P_\sigma) \geq \frac{1}{16(N+1)} \ \ . \tag{15}$$

5.4. Step 4: Bounding Gap($\bar{P}$)

For bounding Gap($\bar{P}$) first note that $\bar{P}$ is reversible with respect to the probability mass function

$$\pi^*(\boldsymbol{\sigma}) = \pi_{\text{PT}}(\Omega_{\boldsymbol{\sigma}}) = \prod_{t=0}^{N} \pi_t(A_{\sigma_t}) \ \ , \forall \boldsymbol{\sigma} \in \Sigma \ \ ,$$

and that according to definition (11) for any $\boldsymbol{\sigma}$, $\boldsymbol{\tau} \in \Sigma$, the probability of moving from $\Omega_{\boldsymbol{\sigma}}$ to $\Omega_{\boldsymbol{\tau}}$ under $\bar{P}$ is given by

$$\bar{P}(\boldsymbol{\sigma}, \boldsymbol{\tau}) = \frac{1}{\pi_{\mathrm{PT}}(\Omega_{\boldsymbol{\sigma}})} \sum_{\boldsymbol{x} \in \Omega_{\boldsymbol{\sigma}}} \sum_{\boldsymbol{y} \in \Omega_{\boldsymbol{\tau}}} \pi_{\mathrm{PT}}(\boldsymbol{x}) P(\boldsymbol{x}, \boldsymbol{y}) \ . \tag{16}$$

We proceed by bounding $\mathrm{Gap}(\bar{P})$ based on theorem 3 by comparing $\bar{P}$ to another π^*-reversible transition matrix T^*. The transition matrix T^* chooses t uniformly from $\{0, \ldots, N\}$ and then draws σ_t with the probability $\pi_t(A_{\sigma_t})$.

To ease the notation let us now denote $\boldsymbol{\sigma}_{[i,j]} = (\sigma_0, \ldots, \sigma_{i-1}, j, \sigma_{i+1}, \ldots, \sigma_N)$. Now we can define the transition matrix T^* by

$$T^*(\boldsymbol{\sigma}, \boldsymbol{\sigma}_{[i,j]}) = \frac{1}{N+1} \pi_i(A_j)$$

and $T^*(\boldsymbol{\sigma}, \boldsymbol{\sigma'}) = 0$ if $\forall i, j : \boldsymbol{\sigma'} \neq \boldsymbol{\sigma}_{[i,j]}$. For the application of theorem 3, for each edge $(\boldsymbol{\sigma}, \boldsymbol{\sigma}_{[i,j]})$ in the transition graph of T^* let us define a path $\gamma_{\boldsymbol{\sigma},\boldsymbol{\sigma}_{[i,j]}}$ in the transition graph of $\bar{P}$ as follows:

Stage 1. change σ_0 to j;

Stage 2. swap j "up" to level i;

Stage 3. swap new σ_{i-1} (formerly σ_i) "down" to level 0;

Stage 4. change value at level 0 to σ_0 (from former σ_i);

We derive an upper bound of the constant c of theorem 3 by splitting it into three terms and bounding each term separately. Here c is the maximum with respect to $\boldsymbol{\tau}$ and $\boldsymbol{\xi}$ (with $\pi^*(\boldsymbol{\tau})\bar{P}(\boldsymbol{\tau}, \boldsymbol{\xi}) > 0$) of

$$\underbrace{\sum_{\gamma_{\boldsymbol{\sigma},\boldsymbol{\sigma}_{[i,j]}} : (\boldsymbol{\tau},\boldsymbol{\xi}) \in \gamma_{\boldsymbol{\sigma},\boldsymbol{\sigma}_{[i,j]}}} |\gamma_{\boldsymbol{\sigma},\boldsymbol{\sigma}_{[i,j]}}|}_{\text{Term I}} \ \underbrace{\frac{\pi^*(\boldsymbol{\sigma})}{\pi^*(\boldsymbol{\tau})}}_{\text{Term II}} \ \underbrace{\frac{T^*(\boldsymbol{\sigma}, \boldsymbol{\sigma}_{[i,j]})}{\bar{P}(\boldsymbol{\tau}, \boldsymbol{\xi})}}_{\text{Term III}} \ .$$

The product of the three upper bounds on the three terms I, II, and III will be used as an upper bound on c.

Bounding Term I. First, we want to find an upper bound for

$$\sum_{\gamma_{\boldsymbol{\sigma},\boldsymbol{\sigma}_{[i,j]}} : (\boldsymbol{\tau},\boldsymbol{\xi}) \in \gamma_{\boldsymbol{\sigma},\boldsymbol{\sigma}_{[i,j]}}} |\gamma_{\boldsymbol{\sigma},\boldsymbol{\sigma}_{[i,j]}}| \tag{17}$$

for the above-defined paths and for any edge $(\boldsymbol{\tau}, \boldsymbol{\xi})$ in the graph of $\bar{P}$.

Let us start with finding an upper bound for the length $|\gamma_{\boldsymbol{\sigma},\boldsymbol{\sigma}_{[i,j]}}|$ of a path. Since $i \in \{0,\dots,N\}$ stage 2 and 3 each can at most consist out of N swapping moves. Stage 1 and 4 each contain only one transition. Thus $|\gamma_{\boldsymbol{\sigma},\boldsymbol{\sigma}_{[i,j]}}| \leq 2N + 2 = 2(N+1)$.

We will now upper bound the number of paths that go through any edge $(\boldsymbol{\tau}, \boldsymbol{\xi})$. If the edge corresponds to stage 1 of the path $\boldsymbol{\tau} = \boldsymbol{\sigma}$ and $\boldsymbol{\xi} = \boldsymbol{\sigma}_{[0,j]}$, and since $i \in \{0,\dots,N\}$ there are not more that $N+1$ paths containing this edge. A similar argumentation holds for edges in stage 4.

If the edge is in stage 2 of the path $\boldsymbol{\tau} = (\sigma_1,\dots,\sigma_l, j, \sigma_{l+1},\dots,\sigma_N)$ for some $l \in \{0,\dots,i\}$ and $\boldsymbol{\xi} = (\sigma_1,\dots,\sigma_l,\sigma_{l+1}, j,\dots,\sigma_N)$. Here, σ_0 is unknown and has 2^n possible values. With $i \in \{0,\dots,N\}$ there are not more then $2^n(N+1)$ paths containing an edge in stage 2 of the path. Similarly, there are no more than $2^n(N+1)$ paths containing a certain edge in stage 3 of the path. Since the same edge could either be found in stage 1 or 4 of the path or in stage 2 or 3 each edge can be in two stages, each corresponding to at most $2^n(N+1)$ paths. Therefore, the total number of paths containing a single edge is no more than $2^{n+1}(N+1) = 2\max\{N+1, 2^n(N+1)\}$ and thus

$$\sum_{\gamma_{\boldsymbol{\sigma},\boldsymbol{\sigma}_{[i,j]}} : (\boldsymbol{\tau},\boldsymbol{\xi}) \in \gamma_{\boldsymbol{\sigma},\boldsymbol{\sigma}_{[i,j]}}} |\gamma_{\boldsymbol{\sigma},\boldsymbol{\sigma}_{[i,j]}}| \leq 2^{n+2}(N+1)^2. \tag{18}$$

Bounding Term II. For bounding $\frac{\pi^*(\boldsymbol{\sigma})}{\pi^*(\boldsymbol{\tau})}$ first note that any state in the stages 1 or 2 of the path from $\boldsymbol{\sigma}$ to $\boldsymbol{\sigma}_{[i,j]}$ as given above is of the form $\boldsymbol{\tau} = (\sigma_1,\dots,\sigma_l, j, \sigma_{l+1},\dots,\sigma_N)$ for some $l \in \{0,\dots,i\}$. Therefore,

$$\frac{\pi^*(\boldsymbol{\sigma})}{\pi^*(\boldsymbol{\tau})} = \left[\prod_{k=1}^{l} \frac{\pi_k(A_{\sigma_k})}{\pi_{k-1}(A_{\sigma_k})}\right] \frac{\pi_0(A_{\sigma_0})}{\pi_l(A_j)} \quad . \tag{19}$$

For the Boltzmann distribution of RBMs we have

$$\frac{\pi_0(A_{\sigma_0})}{\pi_l(A_j)} = \frac{Z_l}{Z_0} \frac{\sum_{\boldsymbol{v}} \exp(-E(\boldsymbol{v}, \boldsymbol{h}_{\sigma_1})\beta_0)}{\sum_{\boldsymbol{v}} \exp(-E(\boldsymbol{v}, \boldsymbol{h}_j)\beta_l)}$$

and for the first term on the left side of equation (19)

$$\prod_{k=1}^{l} \frac{\pi_k(A_{\sigma_k})}{\pi_{k-1}(A_{\sigma_k})} = \frac{Z_0}{Z_l} \prod_{k=1}^{l} \frac{\sum_{\boldsymbol{v}} \exp(-E(\boldsymbol{v}, \boldsymbol{h}_{\sigma_k})\beta_k)}{\sum_{\boldsymbol{v}} \exp(-E(\boldsymbol{v}, \boldsymbol{h}_{\sigma_k})\beta_{k-1})} \quad .$$

Now consider that

$$\frac{\sum_{\boldsymbol{v}} \exp(-E(\boldsymbol{v},\boldsymbol{h}_{\sigma_k})\beta_k)}{\sum_{\boldsymbol{v}} \exp(-E(\boldsymbol{v},\boldsymbol{h}_{\sigma_k})\beta_{k-1})} \leq \frac{\sum_{\boldsymbol{v}} \exp(-\min_{\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h})\beta_k)}{\sum_{\boldsymbol{v}} \exp(-\min_{\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h})\beta_{k-1})} \tag{20}$$

because we can write

$$\frac{\sum_{\boldsymbol{v}} \exp(-E(\boldsymbol{v},\boldsymbol{h}_{\sigma_k})\beta_k)}{\sum_{\boldsymbol{v}} \exp(-E(\boldsymbol{v},\boldsymbol{h}_{\sigma_k})\beta_{k-1})} = \frac{\sum_{\boldsymbol{v}} \exp(-E(\boldsymbol{v},\boldsymbol{h}_{\sigma_k})\beta_k + \max_{\boldsymbol{v},\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h}))}{\sum_{\boldsymbol{v}} \exp(-E(\boldsymbol{v},\boldsymbol{h}_{\sigma_k})\beta_{k-1} + \max_{\boldsymbol{v},\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h}))} \ ,$$

which makes all arguments of the exponential function in the terms of denominator and numerator nonnegative. The function $\frac{R_k+\exp(-x\beta_k+y)}{R_{k-1}+\exp(-x\beta_{k-1}+y)}$ is monotonically decreasing in x for $x \leq y$ for $1 \geq \beta_k \geq \beta_{k-1} \geq 0$, and for each value of $\boldsymbol{v}$ we can write $x = E(\boldsymbol{v},\boldsymbol{h}_{\sigma_k})$ and $y = \max_{\boldsymbol{v},\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h})$ and fix R_k and R_{k-1} to the remaining terms in numerator and denominator, respectively. Thus, the expression gets maximal if we replace x by $\min_{\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h})$. This can be done for all values of $\boldsymbol{v}$. So we can write:

$$\begin{aligned}
&\frac{\pi^*(\boldsymbol{\sigma})}{\pi^*(\boldsymbol{\tau})} \\
&\quad \leq \frac{Z_0}{Z_l}\left[\prod_{k=1}^{l} \frac{\sum_{\boldsymbol{v}} \exp(-\min_{\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h})\beta_k)}{\sum_{\boldsymbol{v}} \exp(-\min_{\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h})\beta_{k-1})}\right] \frac{Z_l}{Z_0} \frac{\sum_{\boldsymbol{v}} \exp(-E(\boldsymbol{v},\boldsymbol{h}_{\sigma_1})\beta_0)}{\sum_{\boldsymbol{v}} \exp(-E(\boldsymbol{v},\boldsymbol{h}_j)\beta_l)} \\
&\qquad = \frac{\sum_{\boldsymbol{v}} \exp(-\min_{\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h})\beta_l)}{\sum_{\boldsymbol{v}} \exp(-\min_{\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h})\beta_0)} \frac{\sum_{\boldsymbol{v}} \exp(-E(\boldsymbol{v},\boldsymbol{h}_{\sigma_1})\beta_0)}{\sum_{\boldsymbol{v}} \exp(-E(\boldsymbol{v},\boldsymbol{h}_j)\beta_l)} \\
&\qquad \overset{\beta_0=0}{=} \frac{\sum_{\boldsymbol{v}} \exp(-\min_{\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h})\beta_l)}{\sum_{\boldsymbol{v}} \exp(-E(\boldsymbol{v},\boldsymbol{h}_j)\beta_l)} \leq \frac{\sum_{\boldsymbol{v}} \exp(-\min_{\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h}))}{\sum_{\boldsymbol{v}} \exp(-\max_{\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h}))}
\end{aligned}$$

Any state $\boldsymbol{\tau}$ in stage 3 of the path is of the form

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

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

$$\frac{\pi^*(\boldsymbol{\sigma})}{\pi^*(\boldsymbol{\tau})} = \left[\prod_{k=1}^{l} \frac{\pi_k(A_{\sigma_k})}{\pi_{k-1}(A_{\sigma_k})}\right] \frac{\pi_0(A_{\sigma_0})}{\pi_i(A_j)} \frac{\pi_i(A_{\sigma_i})}{\pi_l(A_{\sigma_i})} \ .$$

In an analogous way as above we get

$$\begin{aligned}\frac{\pi^*(\boldsymbol{\sigma})}{\pi^*(\boldsymbol{\tau})} \leq{} & \frac{\sum_{\boldsymbol{v}} \exp(-\min_{\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h})\beta_l)}{\sum_{\boldsymbol{v}} \exp(-\min_{\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h})\beta_0)} \times \\ & \frac{\sum_{\boldsymbol{v}} \exp(-\min_{\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h})\beta_0)}{\sum_{\boldsymbol{v}} \exp(-E(\boldsymbol{v},\boldsymbol{h}_j)\beta_i)} \frac{\sum_{\boldsymbol{v}} \exp(-\min_{\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h})\beta_i)}{\sum_{\boldsymbol{v}} \exp(-\min_{\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h})\beta_l)} \\ & \quad = \frac{\sum_{\boldsymbol{v}} \exp(-\min_{\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h})\beta_i)}{\sum_{\boldsymbol{v}} \exp(-E(\boldsymbol{v},\boldsymbol{h}_j)\beta_i)} \leq \frac{\sum_{\boldsymbol{v}} \exp(-\min_{\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h}))}{\sum_{\boldsymbol{v}} \exp(-\max_{\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h}))} \quad .\end{aligned}$$

Any state $\boldsymbol{\tau}$ in stage 4 is given by $\boldsymbol{\tau} = (\sigma_i, \sigma_1, \ldots, \sigma_{i-1}, j, \sigma_{i+1}, \ldots, \sigma_N)$. Therefore,

$$\begin{aligned}\frac{\pi^*(\boldsymbol{\sigma})}{\pi^*(\boldsymbol{\tau})} ={} & \frac{\pi_0(A_{\sigma_0})}{\pi_0(A_i)} \frac{\pi_i(A_{\sigma_i})}{\pi_i(A_j)} \\ & = \frac{\sum_{\boldsymbol{v}} \exp(-E(\boldsymbol{v},\boldsymbol{h}_{\sigma_0})\beta_0)}{\sum_{\boldsymbol{v}} \exp(-E(\boldsymbol{v},\boldsymbol{h}_i)\beta_0)} \frac{\sum_{\boldsymbol{v}} \exp(-E(\boldsymbol{v},\boldsymbol{h}_{\sigma_i})\beta_i)}{\sum_{\boldsymbol{v}} \exp(-E(\boldsymbol{v},\boldsymbol{h}_j)\beta_i)} \\ & \quad \overset{\beta_0=0}{=} \frac{\sum_{\boldsymbol{v}} \exp(-E(\boldsymbol{v},\boldsymbol{h}_{\sigma_i})\beta_i)}{\sum_{\boldsymbol{v}} \exp(-E(\boldsymbol{v},\boldsymbol{h}_j)\beta_i)} \leq \frac{\sum_{\boldsymbol{v}} \exp(-\min_{\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h}))}{\sum_{\boldsymbol{v}} \exp(-\max_{\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h}))} \quad .\end{aligned}$$

Thus, by now we have shown that for all states $\boldsymbol{\tau}$ in the path we have

$$\frac{\pi^*(\boldsymbol{\sigma})}{\pi^*(\boldsymbol{\tau})} \leq \frac{\sum_{\boldsymbol{v}} \exp(-\min_{\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h}))}{\sum_{\boldsymbol{v}} \exp(-\max_{\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h}))} \quad . \tag{21}$$

Bounding Terms III. Now we will bound the remaining term $\frac{T^*(\boldsymbol{\sigma},\boldsymbol{\sigma}_{[i,j]})}{\bar{P}(\boldsymbol{\tau},\boldsymbol{\xi})}$. We have

$$T^*(\boldsymbol{\sigma},\boldsymbol{\sigma}_{[i,j]}) = \frac{1}{N+1}\pi_i(A_j) \quad \leq \frac{1}{N+1} \frac{\sum_{\boldsymbol{v}} \exp(-\min_{\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h}))}{2^n \sum_{\boldsymbol{v}} \exp(-\max_{\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h}))} \quad . \tag{22}$$

For bounding $\bar{P}(\boldsymbol{\tau},\boldsymbol{\xi})$, we consider two cases. Any edge $(\boldsymbol{\tau},\boldsymbol{\xi})$ on the path $\gamma_{[\boldsymbol{\sigma},\boldsymbol{\sigma}_{[i,j]}]}$ can be one of the following two types:

Case 1. It is an edge where we get $\boldsymbol{\xi}$ from $\boldsymbol{\tau} = (\tau_0, \ldots, \tau_N)$ by replacing τ_0 by any new state ξ_0 and, thus, $\boldsymbol{\xi} = \boldsymbol{\tau}_{[0,\xi_0]} = (\xi_0, \tau_1, \ldots, \tau_N)$.

Case 2. It is an edge which we obtain by swapping two elements τ_t and τ_{t+1} and, thus, $\boldsymbol{\xi} = (\tau_0, \ldots, \tau_{t+1}, \tau_t, \ldots, \tau_N)$.

Let us analyse these two cases separately:

Case 1: Based on (16)

$$\bar{P}(\boldsymbol{\tau},\boldsymbol{\xi}) = \frac{1}{\pi_{\mathrm{PT}}(\Omega_{\boldsymbol{\tau}})} \sum_{\boldsymbol{x}\in\Omega_{\boldsymbol{\tau}}} \sum_{\boldsymbol{y}\in\Omega_{\boldsymbol{\xi}}} \pi_{\mathrm{PT}}(\boldsymbol{x}) P(\boldsymbol{x},\boldsymbol{y})$$

and by

$$P(\boldsymbol{x},\boldsymbol{y}) \geq Q(\boldsymbol{x},\boldsymbol{x})T(\boldsymbol{x},\boldsymbol{y})Q(\boldsymbol{x},\boldsymbol{x}) = \frac{1}{4}T(\boldsymbol{x},\boldsymbol{y}) \ ,$$

we get

$$\bar{P}(\boldsymbol{\tau},\boldsymbol{\xi}) \geq \frac{1}{4}\frac{1}{\pi_{\mathrm{PT}}(\Omega_{\boldsymbol{\tau}})} \sum_{\boldsymbol{x}\in\Omega_{\boldsymbol{\tau}}} \sum_{\boldsymbol{y}\in\Omega_{\boldsymbol{\xi}}} \pi_{\mathrm{PT}}(\boldsymbol{x}) T(\boldsymbol{x},\boldsymbol{y}) \ .$$

Because we consider the first case, we can restrict the inner sum to $\boldsymbol{y} \in \Omega_{\boldsymbol{\xi}}$ that differ from $\boldsymbol{x}$ only in the state of the lowest chain, since $T(\boldsymbol{x},\boldsymbol{y}) = 0$ otherwise. That is, for $\boldsymbol{x} = (\boldsymbol{x}_{[0]}, \boldsymbol{x}_{[1]}, \ldots, \boldsymbol{x}_{[N]}) \in \Omega_{\boldsymbol{\tau}}$ we restrict the inner sum to all $\boldsymbol{y} = (\boldsymbol{y}_{[0]}, \boldsymbol{x}_{[1]}, \ldots, \boldsymbol{x}_{[N]})$ with $\boldsymbol{y}_{[0]} \in A_{\xi_0}$. In this case we have

$$T(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{2(N+1)} T_0(\boldsymbol{x}_{[0]}, \boldsymbol{y}_{[0]}) \ .$$

Furthermore, for $\boldsymbol{x}_{[0]} = (\boldsymbol{v}_{[0]}, \boldsymbol{h}_{[0]})$ there is only one $\boldsymbol{y}_{[0]} \in A_{\xi_0}$ that can be reached by one transition of T_0, namely $\boldsymbol{y}_{[0]} = (\boldsymbol{v}_{[0]}, \boldsymbol{h}_{\xi_0})$ with probability $T_0(\boldsymbol{v}_{[0]}, \boldsymbol{h}_{[0]}) = \frac{1}{2}\frac{1}{2^n}$ for $\beta_0 = 0$ (probability of $\frac{1}{2}$ to sample a new state of the 2^n hidden variables, which are uniformly distributed). And thus

$$\bar{P}(\boldsymbol{\tau},\boldsymbol{\xi}) \geq \frac{1}{4}\frac{1}{\pi_{\mathrm{PT}}(\Omega_{\boldsymbol{\tau}})} \sum_{\boldsymbol{x}\in\Omega_{\boldsymbol{\tau}}} \pi_{\mathrm{PT}}(\boldsymbol{x}) \frac{1}{4(N+1)}\frac{1}{2^n} = \frac{1}{16(N+1)}\frac{1}{2^n} \ ,$$

where we use $\sum_{\boldsymbol{x}\in\Omega_{\boldsymbol{\tau}}} \pi_{PT}(\boldsymbol{x}) = \pi_{\mathrm{PT}}(\Omega_{\boldsymbol{\tau}})$. Thus, we get

$$\frac{T^*(\boldsymbol{\sigma},\boldsymbol{\sigma}_{[i,j]})}{\bar{P}(\boldsymbol{\tau},\boldsymbol{\xi})} \leq 16 \cdot 2^n (N+1) T^*(\boldsymbol{\sigma},\boldsymbol{\sigma}_{[i,j]}) \ .$$

Using (22) this is bounded from above by

$$\frac{2^4 \sum_{\boldsymbol{v}} \exp(-\min_{\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h}))}{\sum_{\boldsymbol{v}} \exp(-\max_{\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h}))} \ . \tag{23}$$

Case 2: For bounding $\bar{P}(\boldsymbol{\xi}, \boldsymbol{\tau})$ in the second case, we can make use of $P(\boldsymbol{x}, \boldsymbol{y}) \geq \frac{1}{2} Q(\boldsymbol{x}, \boldsymbol{y})$, because $P = QTQ$ and the swap can either occur under the first or the second Q and the probability of holding under T and the other Q is $\frac{1}{4}$. Thus, we get

$$
\begin{aligned}
\bar{P}(\boldsymbol{\tau}, \boldsymbol{\xi}) \geq & \frac{1}{2} \frac{1}{\pi_{\mathrm{PT}}(\Omega_{\boldsymbol{\tau}})} \sum_{\boldsymbol{x} \in \Omega_{\boldsymbol{\tau}}} \sum_{\boldsymbol{y} \in \Omega_{\boldsymbol{\xi}}} \pi_{\mathrm{PT}}(\boldsymbol{x}) Q(\boldsymbol{x}, \boldsymbol{y}) \\
= & \frac{1}{2} \frac{1}{\prod_{t=0}^{N} \pi_t(A_{\tau_{[t]}})} \sum_{\boldsymbol{x} \in \Omega_{\boldsymbol{\tau}}} \sum_{\boldsymbol{y} \in \Omega_{\boldsymbol{\xi}}} \prod_{t=0}^{N} \pi_t(\boldsymbol{x}_{[t]}) Q(\boldsymbol{x}, \boldsymbol{y}) \\
= & \frac{1}{4N} \frac{1}{\pi_t(A_{\tau_{[t]}}) \pi_{t+1}(A_{\tau_{[t+1]}})} \\
& \sum_{\boldsymbol{x}_{[t]} \in A_{\tau_{[t]}}} \sum_{\boldsymbol{x}_{[t+1]} \in A_{\tau_{[t+1]}}} \pi_t(\boldsymbol{x}_{[t]}) \pi_{t+1}(\boldsymbol{x}_{[t+1]}) \min\big\{1, \frac{\pi_{t+1}(\boldsymbol{x}_{[t]}) \pi_t(\boldsymbol{x}_{[t+1]})}{\pi_t(\boldsymbol{x}_{[t]}) \pi_{t+1}(\boldsymbol{x}_{[t+1]})}\big\} \ ,
\end{aligned}
$$

where in the last step we made use of the facts that $\tau_{[i]} = \xi_{[i]}$ for $i \in \{0, \ldots, t-1, t+2, \ldots, N\}$ and that the probabilities to propose and accept a certain swap according to Q are $\frac{1}{2N}$ and $\min\big\{1, \frac{\pi_{t+1}(\boldsymbol{x}_{[t]}) \pi_t(\boldsymbol{x}_{[t+1]})}{\pi_t(\boldsymbol{x}_{[t]}) \pi_{t+1}(\boldsymbol{x}_{[t+1]})}\big\}$, respectively. We further have

$$
\begin{aligned}
& \sum_{\boldsymbol{x}_{[t]} \in A_{\tau_{[t]}}} \sum_{\boldsymbol{x}_{[t+1]} \in A_{\tau_{[t+1]}}} \pi_t(\boldsymbol{x}_{[t]}) \pi_{t+1}(\boldsymbol{x}_{[t+1]}) \min\big\{1, \frac{\pi_{t+1}(\boldsymbol{x}_{[t]}) \pi_t(\boldsymbol{x}_{[t+1]})}{\pi_t(\boldsymbol{x}_{[t]}) \pi_{t+1}(\boldsymbol{x}_{[t+1]})}\big\} \\
& \quad = \frac{1}{Z_t Z_{t+1}} \sum_{\boldsymbol{v}} \sum_{\hat{\boldsymbol{v}}} \exp(-E(\boldsymbol{v}, \boldsymbol{h}_{\tau_{[t]}}) \beta_t) \exp(-E(\hat{\boldsymbol{v}}, \boldsymbol{h}_{\tau_{[t+1]}}) \beta_{t+1}) \\
& \qquad \min\big\{1, \frac{\exp(-E(\boldsymbol{v}, \boldsymbol{h}_{\tau_{[t]}}) \beta_{t+1}) \exp(-E(\hat{\boldsymbol{v}}, \boldsymbol{h}_{\tau_{[t+1]}}) \beta_t)}{\exp(-E(\boldsymbol{v}, \boldsymbol{h}_{\tau_{[t]}}) \beta_t) \exp(-E(\hat{\boldsymbol{v}}, \boldsymbol{h}_{\tau_{[t+1]}}) \beta_{t+1})}\big\} \\
& \quad = \frac{1}{Z_t Z_{t+1}} \sum_{\boldsymbol{v}} \sum_{\hat{\boldsymbol{v}}} \exp(-E(\boldsymbol{v}, \boldsymbol{h}_{\tau_{[t]}}) \beta_t) \exp(-E(\hat{\boldsymbol{v}}, \boldsymbol{h}_{\tau_{[t+1]}}) \beta_{t+1}) \\
& \qquad \min\{1, \exp\big(- (E(\boldsymbol{v}, \boldsymbol{h}_{\tau_{[t]}}) - E(\hat{\boldsymbol{v}}, \boldsymbol{h}_{\tau_{[t+1]}}))(\beta_{t+1} - \beta_t)\big)\} \\
& \quad \geq \frac{1}{Z_t Z_{t+1}} \exp\big(- (\max_{\boldsymbol{v}, \boldsymbol{h}} E(\boldsymbol{v}, \boldsymbol{h}) - \min_{\boldsymbol{v}, \boldsymbol{h}} E(\boldsymbol{v}, \boldsymbol{h}))(\beta_{t+1} - \beta_t)\big) \\
& \qquad \qquad \sum_{\boldsymbol{v}} \exp(-E(\boldsymbol{v}, \boldsymbol{h}_{\tau_{[t]}}) \beta_t) \sum_{\hat{\boldsymbol{v}}} \exp(-E(\hat{\boldsymbol{v}}, \boldsymbol{h}_{\tau_{[t+1]}}) \beta_{t+1})
\end{aligned}
$$

and

$$\frac{1}{\pi_t(A_{\tau_{[t]}})\pi_{t+1}(A_{\tau_{[t+1]}})} = \frac{Z_t Z_{t+1}}{(\sum_{\boldsymbol{v}} \exp(-E(\boldsymbol{v}, \boldsymbol{h}_{\tau_{[t]}})\beta_t))(\sum_{\hat{\boldsymbol{v}}} \exp(-E(\hat{\boldsymbol{v}}, \boldsymbol{h}_{\tau_{[t+1]}})\beta_{t+1}))}$$

and thus a bound of $\bar{P}(\boldsymbol{\tau}, \boldsymbol{\xi})$ is given by

$$\bar{P}(\boldsymbol{\tau}, \boldsymbol{\xi}) \geq \frac{1}{4N} \exp\big(-\max_t(\beta_{t+1} - \beta_t)(\max_{\boldsymbol{v},\boldsymbol{h}} E(\boldsymbol{v}, \boldsymbol{h})) - \min_{\boldsymbol{v},\boldsymbol{h}} E(\boldsymbol{v}, \boldsymbol{h}))\big) \ .$$

Using

$$\max_{\boldsymbol{v},\boldsymbol{h}} E(\boldsymbol{v}, \boldsymbol{h}) - \min_{\boldsymbol{v},\boldsymbol{h}} E(\boldsymbol{v}, \boldsymbol{h})) \leq \Delta \ ,$$

with $\Delta = \sum_{i,j} |w_{ij}| + \sum_j |b_j| + \sum_i |c_i|$ summing the absolute values of parameters of the RBM, we get

$$\bar{P}(\boldsymbol{\tau}, \boldsymbol{\xi}) \geq \frac{\exp(-\max_t(\beta_{t+1} - \beta_t)\Delta)}{4N} \ . \tag{24}$$

From (24) and the upper bound for $T^*(\boldsymbol{\sigma}, \boldsymbol{\sigma}_{[i,j]})$ given in (22) it follows in the current case that

$$\frac{T^*(\boldsymbol{\sigma}, \boldsymbol{\sigma}_{[i,j]})}{\bar{P}(\boldsymbol{\tau}, \boldsymbol{\xi})} \leq \frac{4}{2^n \exp(-\max_t(\beta_{t+1} - \beta_t)\Delta)} \frac{\sum_{\boldsymbol{v}} \exp(-\min_{\boldsymbol{h}} E(\boldsymbol{v}, \boldsymbol{h}))}{\sum_{\boldsymbol{v}} \exp(-\max_{\boldsymbol{h}} E(\boldsymbol{v}, \boldsymbol{h}))} \ . \tag{25}$$

Combining Terms I, II, III. By joining the results (23) and (25) for term III with the upper bound for term II given in (21) we get in case 1

$$\frac{\pi^*(\boldsymbol{\sigma})T^*(\boldsymbol{\sigma}, \boldsymbol{\sigma}_{[i,j]})}{\pi^*(\boldsymbol{\tau})\bar{P}(\boldsymbol{\tau}, \boldsymbol{\xi})} \leq \frac{2^4(\sum_{\boldsymbol{v}} \exp(-\min_{\boldsymbol{h}} E(\boldsymbol{v}, \boldsymbol{h}))^2}{(\sum_{\boldsymbol{v}} \exp(-\max_{\boldsymbol{h}} E(\boldsymbol{v}, \boldsymbol{h})))^2} \tag{26}$$

and in case 2

$$\frac{\pi^*(\boldsymbol{\sigma})T^*(\boldsymbol{\sigma}, \boldsymbol{\sigma}_{[i,j]})}{\pi^*(\boldsymbol{\tau})\bar{P}(\boldsymbol{\tau}, \boldsymbol{\xi})} \leq \frac{4}{2^n \exp(-\max_t(\beta_{t+1} - \beta_t)\Delta)} \frac{(\sum_{\boldsymbol{v}} \exp(-\min_{\boldsymbol{h}} E(\boldsymbol{v}, \boldsymbol{h})))^2}{(\sum_{\boldsymbol{v}} \exp(-\max_{\boldsymbol{h}} E(\boldsymbol{v}, \boldsymbol{h})))^2} \ . \tag{27}$$

If we combine these results for the two cases (26) and (27) we get

$$\begin{aligned}\frac{\pi^*(\boldsymbol{\sigma})T^*(\boldsymbol{\sigma},\boldsymbol{\sigma}_{[i,j]})}{\pi^*(\boldsymbol{\tau})\bar{P}(\boldsymbol{\tau},\boldsymbol{\xi})} &\leq \max\Bigg\{\frac{2^4(\sum_{\boldsymbol{v}}\exp(-\min_{\boldsymbol{h}}E(\boldsymbol{v},\boldsymbol{h}))^2}{(\sum_{\boldsymbol{v}}\exp(-\max_{\boldsymbol{h}}E(\boldsymbol{v},\boldsymbol{h})))^2},\\ &\qquad \frac{4}{2^n\exp(-\max_t(\beta_{t+1}-\beta_t)\Delta)}\frac{(\sum_{\boldsymbol{v}}\exp(-\min_{\boldsymbol{h}}E(\boldsymbol{v},\boldsymbol{h})))^2}{(\sum_{\boldsymbol{v}}\exp(-\max_{\boldsymbol{h}}E(\boldsymbol{v},\boldsymbol{h})))^2}\Bigg\}\\ &= \frac{4(\sum_{\boldsymbol{v}}\exp(-\min_{\boldsymbol{h}}E(\boldsymbol{v},\boldsymbol{h}))^2}{(\sum_{\boldsymbol{v}}\exp(-\max_{\boldsymbol{h}}E(\boldsymbol{v},\boldsymbol{h})))^2}\cdot\max\left\{4,\frac{1}{2^n\exp(-\max_t(\beta_{t+1}-\beta_t)\Delta)}\right\}\ .\end{aligned} \tag{28}$$

Putting this and (18) together, we arrive at an upper bound for the constant c from theorem 3:

$$\begin{aligned}c &\leq \frac{16(N+1)^2 2^n(\sum_{\boldsymbol{v}}\exp(-\min_{\boldsymbol{h}}E(\boldsymbol{v},\boldsymbol{h})))^2}{(\sum_{\boldsymbol{v}}\exp(-\max_{\boldsymbol{h}}E(\boldsymbol{v},\boldsymbol{h})))^2}\\ &\qquad \max\left\{2^2,\frac{1}{2^n\exp(-\max_t(\beta_{t+1}-\beta_t)\Delta)}\right\}\\ &= \frac{16(N+1)^2(\sum_{\boldsymbol{v}}\exp(-\min_{\boldsymbol{h}}E(\boldsymbol{v},\boldsymbol{h})))^2}{(\sum_{\boldsymbol{v}}\exp(-\max_{\boldsymbol{h}}E(\boldsymbol{v},\boldsymbol{h})))^2}\\ &\qquad \max\left\{2^{n+2},\frac{1}{\exp(-\max_t(\beta_{t+1}-\beta_t)\Delta)}\right\}\end{aligned}$$

Thus, applying theorem 3 and theorem 5, from which follows that $\mathrm{Gap}(T^*)=(N+1)^{-1}$ because we have $b_0=b_1=\cdots=b_N=1/(N+1)$ and all component chains of the product chain T^* have a spectral gap of 1, we get:

$$\begin{aligned}\mathrm{Gap}(\bar{P}) \geq \frac{1}{c}\,\mathrm{Gap}(T^*) &= \frac{(\sum_{\boldsymbol{v}}\exp(-\max_{\boldsymbol{h}}E(\boldsymbol{v},\boldsymbol{h})))^2}{16(N+1)^3(\sum_{\boldsymbol{v}}\exp(-\min_{\boldsymbol{h}}E(\boldsymbol{v},\boldsymbol{h})))^2}\times\\ &\qquad \min\left\{\frac{1}{2^{n+2}},\exp(-\max_t(\beta_{t+1}-\beta_t)\Delta)\right\}\ .\end{aligned} \tag{29}$$

5.5. Step 5: Putting it all together

Using (29) and (15) in (12) leads to

$$\begin{aligned}\mathrm{Gap}(P) &\geq \frac{1}{2}\,\mathrm{Gap}(\bar{P})\min_{\sigma\in\Sigma}\mathrm{Gap}(P_\sigma)\\ &\geq \frac{(\sum_{\boldsymbol{v}}\exp(-\max_{\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h})))^2}{2^8(N+1)^4(\sum_{\boldsymbol{v}}\exp(-\min_{\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h})))^2}\times\\ &\qquad\qquad \min\left\{\frac{1}{2^{n+2}},\exp(-\max_t(\beta_{t+1}-\beta_t)\Delta)\right\}\ ,\end{aligned}$$

which we further bound by making use of the factorisation of the marginal distribution of the hidden RBM variables. As can, for example, be seen from equation (20) in [3], the marginal distribution of the hidden units can be written as $\pi(\boldsymbol{h}) = \frac{1}{Z}\sum_{\boldsymbol{v}}\exp(-E(\boldsymbol{v},\boldsymbol{h})) = \frac{1}{Z}\prod_{i=1}^n \exp(c_i h_i)\prod_{j=1}^m(1+\exp(b_j+\sum_{i=1}^n w_{ij}h_i))$. This leads to

$$\begin{aligned}&\frac{\sum_{\boldsymbol{v}}\exp(-\max_{\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h})))}{\sum_{\boldsymbol{v}}\exp(-\min_{\boldsymbol{h}} E(\boldsymbol{v},\boldsymbol{h}))}\\ &\qquad\leq \frac{\prod_{i=1}^n\exp(\min_{h_i} c_i h_i)}{\prod_{i=1}^n\exp(\max_{h_i} c_i h_i)}\frac{\prod_{j=1}^m(1+\exp(b_j+\sum_{i=1}^n\min_{h_i} w_{ij}h_i))}{\prod_{j=1}^m(1+\exp(b_j+\sum_{i=1}^n\max_{h_i} w_{ij}h_i))}\\ &\qquad= \exp(-\sum_i^n |c_i|)\prod_{j=1}^m\frac{(1+\exp(b_j+\sum_{i=1}^n w_{ij}I_{[-\infty,0]}(w_{ij})))}{(1+\exp(b_j+\sum_{i=1}^n w_{ij}I_{[0,+\infty]}(w_{ij})))}\\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad = \exp(-\Delta_c)\prod_{j=1}^m f_j\ .\end{aligned}$$

with $\Delta_c = \sum_i^n |c_i|$ and

$$\begin{aligned}f_j &= \frac{(1+\exp(b_j+\sum_{i=1}^n w_{ij}I_{[-\infty,0]}(w_{ij})))}{(1+\exp(b_j+\sum_{i=1}^n w_{ij}I_{[0,+\infty]}(w_{ij})))} = \frac{\mathrm{sig}(-b_j-\sum_{i=1}^n w_{ij}I_{[0,+\infty]}(w_{ij}))}{\mathrm{sig}(-b_j-\sum_{i=1}^n w_{ij}I_{[-\infty,0]}(w_{ij}))}\\ &= \frac{\min_{\boldsymbol{h}}\pi(v_j=1|\boldsymbol{h})}{\max_{\boldsymbol{h}}\pi(v_j=1|\boldsymbol{h})}\ ,\end{aligned}$$

where $\mathrm{sig}(x) = \frac{1}{1+\exp(-x)}$. Therefore, f_j is just the fraction of the minimal and maximal activation of the jth visible unit of the RBM.

Thus, we get

$$
\begin{aligned}
\operatorname{Gap}(P) &\geq \frac{\exp(-\Delta_c)}{2^8(N+1)^4}\prod_{j=1}^{m} f_j^2 \times \min\left\{\frac{1}{2^{n+2}}, \exp(-\max_t(\beta_{t+1}-\beta_t)\Delta)\right\} \\
&= \min\left\{\frac{\exp(-\Delta_c)}{2^{n+10}(N+1)^4}\prod_{j=1}^{m} f_j^2, \frac{\exp(-\Delta_c - \max_t(\beta_{t+1}-\beta_t)\Delta)}{2^8(N+1)^4}\prod_{j=1}^{m} f_j^2\right\} \ .
\end{aligned}
$$

It is possible to repeat the proof while changing the roles of hidden and visible variables, which finally leads to (6) and proves our main result.

6. Discussion

Theorem 1 and corollary 1 are the first non-trivial results on the convergence rate of PT for sampling from binary RBMs. As shown in corollary 2, they lead to an upper bound on the approximation error when applying the analyzed PT method to gradient estimation in RBM training.

The upper bound depends exponentially on the size of one layer and the absolute values of the RBM parameters. The fewer the number of nodes and/or the smaller the amount of the parameters, the faster the convergence. This intuitive result resembles the bounds on the approximation bias in contrastive divergence learning [7]. Actually, our current hypothesis is that RBM PT chains are in general not rapidly mixing, and that it is in general not possible to get rid of the exponential dependencies in the RBM sampling complexity. This is the case for some related models such as variants of the Potts model [16, 22].

Because our analysis considers the convergence to the stationary distribution of the product chain consisting of all replica chains, we get an undesired additional linear dependency on the number of replicas. However, such a dependency seems to be inevitable for this type of analysis, and we are not aware of any approach bounding the original chain.

The terms f_j, $j = 1, \ldots, m$, and g_i, $i = 1, \ldots, n$, are the fractions of the minimal and maximal probability to be activated of visible neuron j and hidden neuron i given states of the respective other layer. They are measures of how strongly neurons vary with changing input. The observed dependency is in accordance with the expectation that highly variable chains require more samples to reach the stationary distribution. Furthermore, the minimal and maximal probabilities to be activated give information about how close the conditional distributions are to being uniformly distributed. Gibbs sampling relies on these conditional distributions, and it is optimal if all neurons are activated with probability 0.5 (it reaches the stationary distribution in only one sampling step in this case) and gets more and more deterministic as closer these probabilities get to zero or one—and, thus, mixing slows down as our bound indicates.

When inspecting the role of the inverse temperatures in (6), we see that the term $\max_t(\beta_{t+1}-\beta_t)$ gets minimal, if $\beta_0, \ldots, \beta_N$ are uniformly distributed between 0 and 1. In this case $\max_t(\beta_{t+1}-\beta_t) = \frac{1}{N}$. This theoretical result is in accordance with the heuristic choice of a uniform inverse temperature spacing as used in many applications (e.g., see [9, 11]). So far, to our knowledge, the only theoretically justified heuristic for spacing the inverse temperatures in the tempered transitions literature is choose the $\beta_0, \ldots, \beta_N$ geometrically (i.e., β_{i+1}/β_i = const) [23, 24]. Geometric temperature spacing has been shown to be optimal for sampling from Gaussian distributions by Neil [23], and this result has been extended to a more general class of distribution by Behrens et al. [24]. However, RBMs do not fall into this class and a geometric spacing is sub-optimal for sampling RBMs [25]. Desjardins et al. suggest to use an adaptation scheme that dynamically adjusts the inverse temperatures during RBM training (starting from a uniform distribution). In their experiments, the higher inverse temperatures are adapted to values that are much closer together than a geometric spacing would suggest [25], which also supports rather a uniform than a geometric static inverse temperature spacing.

We regard the PT variant we analyzed as the typical algorithm considered in theoretical analyses of PT, for instance by Madras and Zheng [14] and Woodard et al. [12]. Still, it differs from PT most often used for RBMs in practice. However, as mentioned in section 2.4, in this PT variant the corresponding transition operator is not reversible because of several reasons. Firstly, it consists of an update step in which blockwise Gibbs sampling is performed in all tempered chains. This is not reversible due to the fixed order of first sampling $\boldsymbol{h}^{k+1}$ based on $p(\boldsymbol{h}|\boldsymbol{v}^k)$ followed by sampling $\boldsymbol{v}^{k+1}$ based on $p(\boldsymbol{v}|\boldsymbol{h}^{k+1})$, where the superscripts denote the sampling steps. Secondly, the swapping is usually organized in two substeps, where even temperatures are considered in the first and odd temperatures in the second substep. And, last but not least, we have $P = TQ$ instead of $P = QTQ$.

Lacking reversibility prohibits our type of analysis, and we are not aware of other strategies not requiring the property. Of course, one can design PT operators that are closer to the one used for RBM sampling in practice but still reversible. For example, we can define $P = QTQ$, choose T to perform an update of all visible or all hidden variables (each picked with probability 0.5) in all tempered chains simultaneously, and Q to trying to swap with equal probability either all even or all odd chains in parallel. We have adapted our analysis to this PT variant. However, the resulting bound was actually less tight (among others due to the fact that T is not defined as an product chain anymore, but is of the form $T = \prod_{t=0}^{N} T_t$, and thus theorem 5 can not be applied) and the proof did not provide additional insights.

7. Conclusion

Parallel tempering (PT) is state-of-the-art for sampling restricted Boltzmann machines (RBMs). We presented a first analysis of the convergence rate of the Markov chains of the PT algorithm as considered by Madras and Zheng [14] and Woodard et al. [12] for sampling RBMs by deriving—an arguably loose, but non-trivial—bound on the spectral gap.

We find an exponential dependency on the maximum size of the two layers and the absolute values of the RBM parameters in accordance with existing bounds on the approximation bias in contrastive divergence learning. We hypothesise that RBM PT chains are in general not rapidly mixing, and one can in general not get rid of the exponential dependencies on the number of variables in the RBM sampling complexity. We regard proving conditions under which RBMs are torpid mixing as an interesting question for future research.

Our bound on the spectral gap gets minimal if the inverse temperatures are spaced uniformly, which is a common choice in practice. We do not claim that this spacing is optimal, in particular an adaptive spacing seems to be a good idea. Still, the result matches the practical experience that a uniform spacing (either fixed or as starting point for adaptation) is preferable to a geometric one when using PT for sampling RBMs.

The bound on the convergence rate naturally leads to a bound on the gradient approximation error of the method when used during RBM training, which resembles bounds on the approximation error of contrastive divergence learning [7].

References

[1] P. Smolensky, Information Processing in Dynamical Systems: Foundations of Harmony Theory, in: D. E. Rumelhart, J. L. McClelland (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1: Foundations, MIT Press, 194–281, 1986.

[2] G. E. Hinton, Training Products of Experts by Minimizing Contrastive Divergence, Neural Computation 14 (2002) 1771–1800.

[3] A. Fischer, C. Igel, Training Restricted Boltzmann Machines: An Introduction, Pattern Recognition 47 (2013) 25–39.

[4] G. E. Hinton, R. R. Salakhutdinov, Reducing the Dimensionality of Data with Neural Networks, Science 313 (5786) (2006) 504–507.

[5] A. Fischer, C. Igel, Empirical Analysis of the Divergence of Gibbs Sampling Based Learning Algorithms for Restricted Boltzmann Machines, in: K. Diamantaras, W. Duch, L. S. Iliadis (Eds.), International Conference on Artificial Neural Networks (ICANN 2010), vol. 6354 of *LNCS*, Springer-Verlag, 208–217, 2010.

[6] Y. Bengio, O. Delalleau, Justifying and Generalizing Contrastive Divergence, Neural Computation 21 (6) (2009) 1601–1621.

[7] A. Fischer, C. Igel, Bounding the Bias of Contrastive Divergence Learning, Neural Computation 23 (2011) 664–673.

[8] C. J. Geyer, Markov chain Monte Carlo maximum likelihood, in: E. Kerami (Ed.), Proceedings of the 23rd Symposium on the Interface of Computing Science and Statistics, Interface Foundation of North America, 156–163, 1991.

[9] R. Salakhutdinov, Learning in Markov Random Fields using Tempered Transitions, in: Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, A. Culotta (Eds.), Advances in Neural Information Processing Systems (NIPS 22), 1598–1606, 2009.

[10] G. Desjardins, A. Courville, Y. Bengio, P. Vincent, O. Dellaleau, Parallel Tempering for Training of Restricted Boltzmann Machines, Journal of Machine Learning Research Workshop and Conference Proceedings 9 (AISTATS 2010) (2010) 145–152.

[11] K. Cho, T. Raiko, A. Ilin, Parallel tempering is efficient for learning restricted Boltzmann machines, in: Proceedings of the International Joint Conference on Neural Networks (IJCNN 2010), IEEE Press, 3246–3253, 2010.

[12] D. B. Woodard, S. C. Schmidler, M. Huber, Conditions for Rapid Mixing of Parallel and Simulated Tempering on Multimodal Distributions, The Annals of Applied Probability 19 (2009) 617–640.

[13] P. Diaconis, L. Saloff-Coste, What do we know about the Metropolis algorithm?, Journal of Computer and System Sciences 57 (1998) 20–36.

[14] N. Madras, Z. Zheng, On the swapping algorithm, Random Structures Algorithms 22 (2003) 66–97.

[15] K. Brügge, A. Fischer, C. Igel, The flip-the-state transition operator for Restricted Boltzmann Machines, Machine Learning 13 (2013) 53–69.

[16] N. Bhatnagar, D. Randall, Torpid Mixing of Simulated Tempering on the Potts Model, in: Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 478–487, 2004.

[17] S. Caracciolo, A. Pelissetto, A. D. Sokal, Two remarks on simulated tempering, Unpublished manuscript .

[18] N. Madras, D. Randall, Markov chain decomposition for convergence rate analysis, The Annals of Applied Probability 12 (581–606).

[19] P. Diaconis, L. Saloff-Coste, Comparison theorems for reversible Markov chains, The Annals of Applied Probability 3 (1993) 696–730.

[20] P. Brémaud, Markov chains: Gibbs fields, Monte Carlo simulation, and queues, Springer-Verlag, 1999.

[21] P. Diaconis, L. Saloff-Coste, Logarithmic Sobolev inequalities for finite Markov chains, The Annals of Applied Probability 6 (1996) 695–750.

[22] D. Woodard, S. Schmidler, M. Huber, Sufficient Conditions for Torpid Mixing of Parallel and Simulated Tempering, Electronic Journal of Probability 14 (2009) 780–804.

[23] R. Neal, Sampling from multimodal distributions using tempered transitions, Statistics and Computing 6 (1996) 353366.

[24] G. Behrens, N. Friel, M. Hurn, Tuning tempered transitions, Statistics and Computing 22 (1) (2012) 65–78.

[25] G. Desjardins, A. Courville, Y. Bengio, Adaptive Parallel Tempering for Stochastic Maximum Likelihood Learning of RBMs, in: NIPS 2010 Workshop on Deep Learning and Unsupervised Feature Learning, 2010.

Appendix A. Bounding the bias of a MCMC-estimate in terms of the convergence rate

The convergence rate of a Markov chain can be used to bound the bias of an approximation of the expected value of a function under the model distribution based on samples from the chain. This is demonstrated in the following.

Consider a Markov chain on the state space Ω. Let π be the stationary distribution of the chain and $P^k(\boldsymbol{z}, \cdot)$ be the distribution obtained by running the Markov chain for k steps starting from a fixed sample $\boldsymbol{z}$. Furthermore, let $t : \Omega \rightarrow \mathbb{R}$ be an arbitrary function. Then, a bound on the bias of an estimate that approximates the expected value of $t(\boldsymbol{x})$ under π based on samples from $P^k(\boldsymbol{z}, \cdot)$ is given by

$$\begin{aligned}
\Big| \sum_{\boldsymbol{x}} \pi(\boldsymbol{x}) t(\boldsymbol{x}) - \sum_{\boldsymbol{x}} P^k(\boldsymbol{z}, \boldsymbol{x})) t(\boldsymbol{x}) \Big| &= \Big| \sum_{\boldsymbol{x}} (\pi(\boldsymbol{x}) - P^k(\boldsymbol{z}, \boldsymbol{x}) t(\boldsymbol{x}) \Big| \\
&\leq \sum_{\boldsymbol{x}} \big| (\pi(\boldsymbol{x}) - P^k(\boldsymbol{z}, \boldsymbol{x})) t(\boldsymbol{x}) \big| \\
&\leq \max \big| t(\boldsymbol{x}) \big| \sum_{\boldsymbol{x}} \big| \pi(\boldsymbol{x}) - P^k(\boldsymbol{z}, \boldsymbol{x} \big| \\
&= 2 \max \big| t(\boldsymbol{x}) \big| \| \pi - P^k(\boldsymbol{z}, \cdot) \|
\end{aligned}$$

where $\| \pi - P^k(\boldsymbol{z}, \cdot) \|$ is the variation distance, which can be bounded, for example, using equation (3).

Appendix B. Relation between the convergence rates of the PT product chain and the original chain

In the following, we proof that bounding the convergence rate of the PT product chain also bounds the convergence rate of the original chain (i.e., the chain with inverse temperature $\beta_N = 1$). Assume that for the product chain holds

$$\frac{1}{2}\sum_{\boldsymbol{x}} |P^k(\boldsymbol{y},\boldsymbol{x}) - \prod_{t=0}^{N} \pi_t(\boldsymbol{x}_{[t]})| < \epsilon \ ,$$

for an arbitrary starting point $\boldsymbol{y}$. Then we can write

$$\frac{1}{2}\sum_{\boldsymbol{x}_{[N]}} \sum_{\boldsymbol{x}_{[-N]}} |P^k(\boldsymbol{y},\boldsymbol{x}) - \prod_{t=0}^{N} \pi_t(\boldsymbol{x}_{[t]})| < \epsilon \ ,$$

which implies

$$\epsilon > \frac{1}{2}\sum_{\boldsymbol{x}_{[N]}} \sum_{\boldsymbol{x}_{[-N]}} \Big|P^k(\boldsymbol{y},\boldsymbol{x}) - \prod_{t=0}^{N} \pi_t(\boldsymbol{x}_{[t]})\Big| \geq \frac{1}{2}\sum_{\boldsymbol{x}_{[N]}} \Big| \sum_{\boldsymbol{x}_{[-N]}} (P^k(\boldsymbol{y},\boldsymbol{x}) - \prod_{t=0}^{N} \pi_t(\boldsymbol{x}_{[t]}))\Big| \ .$$

Thus, ϵ also upper bounds the variation distance for the original chain at temperature 1:

$$\begin{aligned}\frac{1}{2}\sum_{\boldsymbol{x}_{[N]}} \Big| \sum_{\boldsymbol{x}_{[-N]}} (P^k(\boldsymbol{y},\boldsymbol{x}) - \prod_{t=0}^{N} \pi_t(\boldsymbol{x}_{[t]}))\Big| &= \frac{1}{2}\sum_{\boldsymbol{x}_{[N]}} \Big| \sum_{\boldsymbol{x}_{[-N]}} P^k(\boldsymbol{y},\boldsymbol{x}) - \sum_{\boldsymbol{x}_{[-N]}} \prod_{t=0}^{N} \pi_t(\boldsymbol{x}_{[t]})\Big| \\ &= \frac{1}{2}\sum_{\boldsymbol{x}_{[N]}} \Big| P_N^k(\boldsymbol{y},\boldsymbol{x}_{[0]}) - \sum_{\boldsymbol{x}_{[-N]}} \prod_{t=0}^{N} \pi_t(\boldsymbol{x}_{[t]})\Big| \\ &= \frac{1}{2}\sum_{\boldsymbol{x}_{[N]}} \Big| P_N^k(\boldsymbol{y},\boldsymbol{x}_{[0]}) - \pi_N(\boldsymbol{x}_{[0]}) \sum_{\boldsymbol{x}_{[-N]}} \prod_{t=0}^{N-1} \pi_t(\boldsymbol{x}_{[t]})\Big| \\ &= \frac{1}{2}\sum_{\boldsymbol{x}_{[N]}} \Big| P_N^k(\boldsymbol{y},\boldsymbol{x}_{[0]}) - \pi_N(\boldsymbol{x}_{[0]})\Big| < \epsilon\end{aligned}$$