Mixing Rates for the Alternating Gibbs Sampler over Restricted Boltzmann Machines and Friends

Christopher Tosh CTOSH@CS.UCSD.EDU
Department of Computer Science and Engineering, UC San Diego, 9500 Gilman Dr., La Jolla, CA 92093

Abstract

Alternating Gibbs sampling is a modification of classical Gibbs sampling where several variables are simultaneously sampled from their joint conditional distribution. In this work, we investigate the mixing rate of alternating Gibbs sampling with a particular emphasis on Restricted Boltzmann Machines (RBMs) and variants.

1. Introduction

Markov Random Fields (MRFs) are a popular class of graphical models which have found uses from image restoration (Geman & Geman, 1984), to modeling in statistical physics (Ising, 1925; Potts, 1952), to pretraining deep neural networks (Hinton et al., 2006; Bengio, 2009). Formally, a Markov Random Field consists of an underlying graph $G = (V, E)$ and a set of random variables $X = (X_v)_{v \in V}$ indexed by the vertices V satisfying

$$P\left(X_v \,|\, X_{V \setminus \{v\}}\right) = P\left(X_v \,|\, X_{N(v)}\right)$$

where $N(v)$ is the set of vertices adjacent to v in G, also known as the *Markov blanket* of v.

A fundamental problem in the setting of MRFs is to sample from the joint distribution $P(X)$. When the state space of X is finite and each state has positive probability, the Hammersley-Clifford theorem (Hammersley & Clifford, 1971; Besag, 1974) tells us that we can decompose the probability density function as

$$P(X = x) = \frac{1}{Z} \prod_{c \in \mathrm{cl}(G)} \psi_c(x_c)$$

where $\mathrm{cl}(G)$ is the set of maximal cliques of G, $\psi_c(\cdot)$ are positive functions, and Z is the normalizing constant to make the density sum to one. In general, computing Z is a hard problem (Bulatov & Grohe, 2005), which makes exactly sampling from $P(X)$ challenging. The solution to this problem is to approximately sample from $P(X)$.

The Gibbs sampler is a generic Markov chain method to approximately sample from a joint distribution using only the conditional distributions to do so. In the case of MRFs, the Gibbs sampler maintains a current state $(X_v = x_v)_{v \in V}$, and it takes a single step by choosing an index $v \in V$ and updating the value of X_v according to the conditional distribution $P(X_v \,|\, X_{N(v)} = x_{N(v)})$. If we can efficiently sample from these conditional distributions then each step of the Gibbs sampler is also efficient. For many MRFs of interest, this is indeed the case.

It is well known that the state of the Gibbs sampler converges to the joint distribution $P(X)$ (Levin et al., 2008, Chapter 3). Unfortunately, this convergence is only guaranteed in the limit. Thus, a central object of study in Markov chain literature is the rate at which a Markov chain converges to its stationary distribution, and this quantity is known as its *mixing rate*.

In some cases, it is possible to efficiently sample more than a single random variable at a time. Consider an MRF whose underlying graph is k-colorable, i.e. there is a partition $B_1, \ldots, B_k$ of V such that for all $i \in \{1, \ldots, k\}$ and all $u, v \in B_i$, the edge (u, v) does not appear in the graph. Then conditioning on $V \setminus B_i$, the elements of B_i are independent and the joint conditional distribution factorizes:

$$P(X_{B_i} \,|\, X_{V \setminus B_i}) = \prod_{v \in B_i} P(X_v \,|\, X_{N(v)}).$$

If we can efficiently sample from the individual conditional distributions then we can also do so for these joint conditional distributions. Moreover, we can modify the Gibbs sampler so that at each step it updates an entire block B_i, and it will still converge to the correct distribution. This Markov chain is the *alternating Gibbs sampler*.

1.1. Restricted Boltzmann Machines

An important special case of a Markov Random Field is the Restricted Boltzmann Machine (RBM). The underly-

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

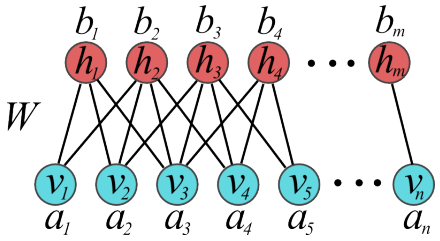

Figure 1. The structure of a Restricted Boltzmann Machine.

ing graph of an RBM is a fully connected bipartite graph with visible nodes $v = (v_1, \ldots, v_n)$ and hidden nodes $h = (h_1, \ldots, h_n)$. A *configuration* $x = (x(h), x(v))$ is an assignment of each node to a value in $\{0, 1\}$. The *energy* of a configuration x is

$$E(x) = -\sum_{i=1}^{n} a_i x(v_i) - \sum_{j=1}^{m} b_j x(h_j) - \sum_{i,j} x(v_i) W_{ij} x(h_j)$$

where the a_i's and b_j's are biases and the W_{ij}'s are interaction strengths or weights. This induces the Gibbs distribution over configurations: for a random configuration X, $P(X = x) = \frac{1}{Z} e^{-E(x)}$, where Z is the normalizing constant to make the distribution integrate to one. Because the underlying graph is bipartite, the conditional distribution of a visible node v_i is

$$\begin{aligned} P(X(v_i) = 1 \,|\, x(N(v_i))) &= P(X(v_i) = 1 \,|\, x(h)) \\ &= \sigma(a_i + \sum_{j=1}^{m} W_{ij} x(h_j)) \end{aligned}$$

where $\sigma(t) = 1/(1+e^{-t})$ is the logistic sigmoid. Similarly, the conditional distribution of a hidden node h_j is

$$P(X(h_j) = 1 \,|\, x(v)) = \sigma(b_j + \sum_{i=1}^{n} W_{ij} x(v_i)).$$

Because these conditional distributions are easy to sample, the Gibbs sampler can be implemented efficiently.

Alternating Gibbs sampling is particularly simple in the case of RBMs. Since RBMs are built on bipartite graphs, the alternating Gibbs sampler first independently samples all of the hidden nodes conditioned on the visible nodes and then independently samples all of the visible nodes conditioned on the hidden nodes. Further, this simplicity is not restricted to RBMs themselves; it only requires that the MRF in question have an underlying graph that is bipartite.

In this paper, we consider the mixing rates for the alternating Gibbs sampler for a wide variety of bipartite MRFs.

1.2. High-level Overview

In Section 3 we give conditions which are sufficient to guarantee that the alternating Gibbs sampler over discrete-valued bipartite MRFs rapidly converges to its stationary distribution. As corollaries we establish conditions for rapid mixing for the alternating Gibbs sampler over RBMs as well as two important variants: Deep Boltzmann Machines and Softmax RBMs.

In Section 4 we give conditions which guarantee rapid convergence for the alternating Gibbs sampler over general, continuously-valued bipartite MRFs. As a consequence we are also able to establish conditions for rapid convergence for the alternating Gibbs sampler over two continuously-valued variants of RBMs: Gaussian-NReLU RBMs and Gaussian-Gaussian RBMs.

In Section 5, we establish lower bounds on the mixing rate for the alternating Gibbs sampler in the cases of RBMs and Gaussian-Gaussian RBMs.

In Section 6 we discuss computational complexity issues surrounding RBMs and remaining open questions.

1.3. Related Work

There has been some recent work on proving mixing rates for the Gibbs sampler on a wide range of models. Notably, Liu and Domkey (2014), De Sa et al. (2015), Gotovos et al. (2015) gave upper bounds for the mixing rate for the single-site update Gibbs sampler over a wide class of models which include certain discrete-valued MRFs.

De Sa et al. and Gotovos et al. both introduced quantities for the models that they consider for which the mixing rate of the Gibbs sampler is polynomial in the size of the model and exponential in these special quantites. More closely related to our work, Gotovos et al. and Liu and Domke also showed that if the model meets a certain 'bounded influence' criterion, then the single-site update Gibbs sampler mixes in time $O(n \log n)$.

There has also been some recent work on single-site Gibbs sampling in general state spaces. Notably, Wang and Wu (2014) gave general convergence rates for the single-site update Gibbs sampler on general state spaces.

In this work, we also give general convergence results in both discrete and continuous spaces, but for the alternating Gibbs sampler as opposed to the single-site update Gibbs sampler. We then apply these results to a variety of models closely related to the standard RBM, such as the Gaussian-NReLU RBM, for which mixing rate bounds were previously unknown.

2. Preliminaries and Notation

A Markov chain is a stochastic process $(X_t)_{t=0}^{\infty}$ taking values in some space Ω and satisfying the *Markov property*: $Pr(X_t \in A \,|\, X_{t-1}, \ldots, X_0) = Pr(X_t \in A \,|\, X_{t-1})$. For

ease of exposition, we will assume that Ω is a finite set. For an overview of the case when Ω is a general space, see (Roberts & Rosenthal, 2004).

The transition probabilities can be viewed as a matrix Q indexed by elements of Ω such that

$$Q(x, y) = Pr(X_t = y \,|\, X_{t-1} = x).$$

Q is *irreducible* if, for all $x, y \in \Omega$, there exists a $t > 0$ such that $Q^t(x, y) > 0$. It is *aperiodic* if

$$\gcd(\{t : Q^t(x, y) > 0\}) = 1 \text{ for all } x, y \in \Omega.$$

A distribution π over Ω is a *stationary distribution* of Q if, when π and Q are viewed as matrices indexed by Ω, then $\pi = \pi Q$. A fundamental result of Markov chain theory says that if a Markov chain Q is irreducible and aperiodic, then it has a unique stationary distribution. Furthermore, the distribution of X_t converges to π, regardless of initial distribution (Levin et al., 2008, Theorem 4.9).

Convergence is only guaranteed in the limit. Thus, we need to define when the distribution of the Markov chain is a good approximation of its stationary distribution. Given probability measures μ, ν over Ω, the *total variation distance* has two equivalent characterizations:

$$\|\mu - \nu\|_{TV} := \sup_{A \subset \Omega} |\mu(A) - \nu(A)| = \frac{1}{2} \sum_{\omega \in \Omega} |\mu(\omega) - \nu(\omega)|$$

where the supremum is taken over all measurable subsets of Ω. The *mixing rate* of a Markov chain Q with unique stationary distribution π is the function $\tau : (0, 1) \to \mathbb{N}$

$$\tau(\epsilon) := \min\{t : \max_{x \in \Omega} \|Q^t(x, \cdot) - \pi\|_{TV} < \epsilon\}.$$

2.1. Couplings

In this section, we introduce our main technique for bounding the mixing rate of a Markov chain: the coupling.

Let μ and ν be probability measures over a space Ω. A pair of random variables (X, Y) is a *coupling* of μ and ν if for all measurable sets A and B,

$$Pr(X \in A) = \mu(A) \text{ and } Pr(Y \in B) = \nu(B).$$

Couplings are convenient probabilistic tools for bounding distances between measures. The following lemma, whose proof can be found in (Aldous, 1983), tells us that not only does any coupling provide an upper bound on the total variation distance between measures but also that there exists a coupling that achieves this bound.

Lemma 1. *Let μ and ν be probability measures.*

(a) For any coupling (X, Y) of μ and ν, $\|\mu - \nu\|_{TV} \leq Pr(X \neq Y)$.

(b) There exists a coupling (X, Y) satisfying $\|\mu - \nu\|_{TV} = Pr(X \neq Y)$.

Couplings involving entire stochastic processes can be quite cumbersome to work with. A convenient restricted class of couplings for Markov chains are the Markovian couplings. A *Markovian coupling* of a Markov chain over Ω with transition matrix Q is a Markov chain (X_t, Y_t) over $\Omega \times \Omega$ whose transitions satisfy

$$Pr(X_{t+1} = x' \,|\, X_t = x, Y_t = y) = Q(x, x'),$$
$$Pr(Y_{t+1} = y' \,|\, X_t = x, Y_t = y) = Q(y, y').$$

The following lemma relates Markovian couplings to the mixing time. It dates back at least to Aldous (1983) and can be found in the form we present, for example, in (Jerrum, 2003, Lemma 4.7).

Lemma 2. *Let (X_t, Y_t) be a Markovian coupling for Markov chain Z_t such that there exists a function $\tau_{couple} : (0, 1) \to \mathbb{N}$ satisfying that for all $x, y \in \Omega$ and $\epsilon > 0$, $Pr(X_{\tau_{couple}(\epsilon)} \neq Y_{\tau_{couple}(\epsilon)} \,|\, X_0 = x, Y_0 = y) \leq \epsilon$. Then the mixing rate for Z_t satisfies $\tau(\epsilon) \leq \tau_{couple}(\epsilon)$.*

2.2. Semimetrics and Matrix Norms

A bivariate function $d(\cdot, \cdot)$ is a *semimetric* over a space $\mathcal{X}$ if for all $x, y \in \mathcal{X}$ if it satisfies all the properties of a metric except for the triangle inequality, i.e. non-negativity, identity iff equality, and symmetry. Any metric is trivially a semimetric. In addition, distances such as ℓ_2^2-distance are also semimetrics.

Given a matrix $A \in \mathbb{R}^{n \times m}$ and positive reals $p, q > 0$, the $L_{p,q}$ norm of A is defined as

$$\|A\|_{p,q} = \left(\sum_{j=1}^{m} \left[\sum_{i=1}^{n} |A_{ij}|^p \right]^{q/p} \right)^{1/q}.$$

Special cases of the $L_{p,q}$ include the Frobenius norm,

$$\|A\|_F = \|A\|_{2,2} = \sqrt{\sum_{j=1}^{m} \sum_{i=1}^{n} A_{ij}^2},$$

the L_1-norm, $\|A\|_1 = \|A\|_{1,\infty} = \max_{1 \leq j \leq m} \sum_{i=1}^{n} |A_{ij}|$, and the max-norm, $\|A\|_{\max} = \|A\|_{\infty,\infty} = \max_{i,j} |A_{ij}|$.

3. The Discrete Case

Suppose that we have two vectors of nodes: visible nodes $v = (v_1, \ldots, v_n)$ and hidden nodes $h = (h_1, \ldots, h_m)$. Let $\mathcal{X}$ be some finite space, and let Ω_v denote the set of configurations x which assign to each visible node a value in $\mathcal{X}$. We can also define Ω_h to be the same except for

hidden nodes and $\Omega = \Omega_v \times \Omega_h$ to be the configurations which assign to every node a value in $\mathcal{X}$.

For $x \in \Omega_h$, let $P^{(v)}(\cdot \,|\, x(h))$ denote the conditional distribution of the visible nodes given an assignment to the hidden nodes. We can symmetrically define $P^{(h)}(\cdot \,|\, x(v))$. For two configurations $x, y \in \Omega$, let $d_v(x, y)$ denote a semimetric over the assignments to the visible nodes. Similarly, let $d_h(x, y)$ denote a semimetric over the hidden nodes. Define

$$\gamma_v^{(\min)} = \min_{x \neq y} d_v(x, y) \text{ and } \gamma_v^{(\max)} = \max_{x \neq y} d_v(x, y).$$

Similarly, define $\gamma_h^{(\min)}$ and $\gamma_h^{(\max)}$ as the corresponding extremal hidden distances.

The alternating Gibbs sampler is the Markov chain $(X_t)_{t=0}^{\infty}$ taking values in Ω, which starts at some initial configuration $X_0 = x_0$, and performs the following for $t = 1, 2, \ldots$

1. Draw $X_t(h) \sim P^{(h)}(\cdot \,|\, X_{t-1}(v))$
2. Draw $X_t(v) \sim P^{(v)}(\cdot \,|\, X_t(h))$

We say that the distribution $P^{(v)}$ is *c-contractive* if for any assignments $x, y \in \Omega$ there exists a coupling (X, Y) of $P^{(v)}(\cdot \,|\, x(h))$ and $P^{(v)}(\cdot \,|\, y(h))$ satisfying

$$\mathbb{E}\left[d_v(X, Y)\right] \leq c\, d_h(x, y).$$

Contractivity for $P^{(h)}$ is defined symmetrically. With these notions in hand, we are ready to state our first theorem.

Theorem 3. *Let $c_1, c_2 \geq 0$ such that $c_1 c_2 < 1$, $P^{(v)}$ is c_1-contractive, and $P^{(h)}$ is c_2-contractive. Then the mixing rate of the Gibbs sampler is bounded as*

$$\tau(\epsilon) \leq 1 + \frac{1}{\log(1/c_1 c_2)} \log\left(\frac{C}{\epsilon}\right)$$

where $C = \min\left(\frac{\gamma_v^{(max)}}{\gamma_v^{(min)}}, \frac{\gamma_h^{(max)}}{\gamma_h^{(min)}}, \frac{c_2 \gamma_v^{(max)}}{\gamma_h^{(min)}}\right)$.

Proof. We will prove $\tau(\epsilon) \leq 1 + \frac{1}{\log(1/c_1 c_2)} \log\left(\frac{\gamma_v^{(\max)}}{\epsilon \gamma_v^{(\min)}}\right)$. The other inequalities are left to the appendix. Our strategy is to glue together the two contractive couplings for the conditional distributions in order to make a Markovian coupling for the Gibbs sampler. Formally, if we are at time step t, then we will first sample $(X_{t+1}(h), Y_{t+1}(h))$ according to the c_1-contractive coupling of $P^{(h)}(\cdot \,|\, X_t(v))$ and $P^{(h)}(\cdot \,|\, Y_t(v))$. Then we will sample $(X_{t+1}(v), Y_{t+1}(v))$ according to the c_2-contractive coupling of $P^{(v)}(\cdot \,|\, X_{t+1}(h))$ and $P^{(v)}(\cdot \,|\, Y_{t+1}(h))$. By construction, this is a valid Markovian coupling for the alternating Gibbs sampler. For $t \geq 1$ and any initial distribution of X_0 and Y_0, we have

$$Pr(X_t \neq Y_t) \leq Pr(d_v(X_{t-1}, Y_{t-1}) \geq \gamma_v^{(\min)}).$$

By Markov's inequality and the law of total expectation, we have

$$\begin{aligned} Pr(d_v(X_{t-1}, Y_{t-1}) \geq \gamma_v^{(\min)}) &\leq \frac{\mathbb{E}[d_v(X_{t-1}, Y_{t-1})]}{\gamma_v^{(\min)}} \\ &\leq \frac{(c_1 c_2)^{t-1} \mathbb{E}[d_v(X_0, Y_0)]}{\gamma_v^{(\min)}} \\ &\leq \frac{(c_1 c_2)^{t-1} \gamma_v^{(\max)}}{\gamma_v^{(\min)}} \end{aligned}$$

For $t \geq 1 + \frac{1}{\log(c_1 c_2)} \log\left(\frac{\gamma_v^{(\max)}}{\epsilon \gamma_v^{(\min)}}\right)$, the above is less than ϵ. Applying Lemma 2 completes the proof. □

3.1. Restricted Boltzmann Machines

Returning to the case of RBMs, recall that a configuration assigns values in $\{0, 1\}$ and the conditional distributions are product distributions whose components are of the form

$$P_{RBM}^{(v)}(X(v_i) = 1 \,|\, x(h)) = \sigma(a_i + \sum_{j=1}^{m} W_{ij} x(h_j))$$

$$P_{RBM}^{(h)}(X(h_j) = 1 \,|\, x(v)) = \sigma(b_j + \sum_{i=1}^{n} W_{ij} x(v_i))$$

where $\sigma(t) = 1/(1 + \exp(-t))$ is the logistic sigmoid, and $a \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, and $W \in \mathbb{R}^{n \times m}$ are parameters of the model. We will use Hamming distance as our semimetric for both hidden and visible distances, i.e.

$$\begin{aligned} d_v(x, y) &= |\{i : x(v_i) \neq y(v_i)\}| \text{ and} \\ d_h(x, y) &= |\{j : x(h_j) \neq y(h_j)\}| \end{aligned}$$

The following lemma, which is proven in the appendix, establishes the contractivity of the RBM conditional distributions with respect to Hamming distance.

Lemma 4. *$P_{RBM}^{(v)}$ and $P_{RBM}^{(h)}$ are $\frac{\|W\|_1}{2}$- and $\frac{\|W^T\|_1}{2}$-contractive, respectively.*

Combining this with the simple observations that $\gamma_v^{(\min)} = \gamma_h^{(\min)} = 1$, $\gamma_v^{(\max)} = n$, and $\gamma_h^{(\max)} = m$ we have the following corollary of Theorem 3.

Corollary 5. *The mixing rate for the alternating Gibbs sampler over an RBM whose weight matrix W satisfies $\|W\|_1 \|W^T\|_1 < 4$ is upper bounded as*

$$\tau(\epsilon) \leq \frac{1}{\log(4) - \log(\|W\|_1 \|W^T\|_1)} \log\left(\frac{\min(n, m)}{\epsilon}\right).$$

3.2. Deep Boltzmann Machines

A natural way to generalize an RBM is to consider several stacked layers of nodes $v^{(1)}, \ldots, v^{(K)}$ of sizes $n_1, \ldots, n_K$ with interaction matrices $W^{(i)} \in \mathbb{R}^{n_i \times n_{i+1}}$ connecting

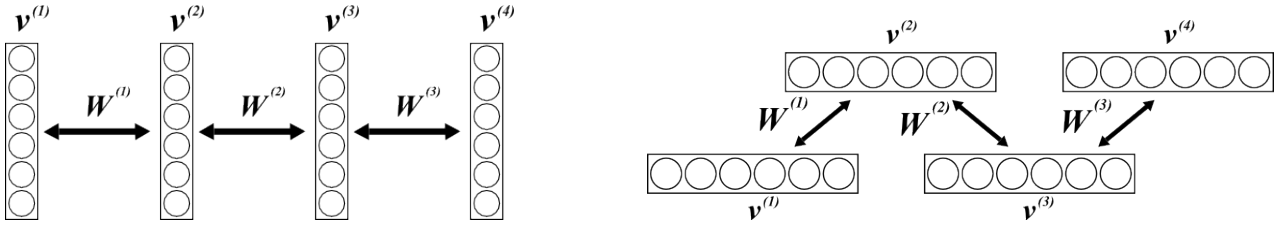

Figure 2. Unrolling a DBM. *Left*: the standard stacked view of a DBM. *Right*: unrolling a DBM into a bipartite graph.

them. This Markov field is known as a *Deep Boltzmann Machine* (DBM) (Salakhutdinov & Hinton, 2009). Figure 2 gives two visualizations of a 4-layer DBM.

As one can see from the 'unrolled' view in Figure 2, DBMs are also bipartite MRFs. Indeed, they are a special case of RBMs in which the visible nodes correspond to the odd layer nodes and the hidden nodes correspond to the even layer nodes and the weight matrix is given by

$$W := \begin{bmatrix} W^{(1)} & 0 & 0 & \\ {W^{(2)}}^T & W^{(3)} & 0 & \\ 0 & {W^{(4)}}^T & W^{(5)} & \\ 0 & 0 & & \ddots \end{bmatrix}$$

Thus the alternating Gibbs sampler can be applied to DBMs where we sample first the even layers and then the odd layers. Corollary 5 then immediately implies the following.

Corollary 6. *Let $W^{(1)}, \dots, W^{(K)}$ be the weight matrices of a DBM and let W be defined as above. Then if $\|W\|_1\|W^T\|_1 < 4$ the mixing rate of the alternating Gibbs sampler is bounded above as*

$$\tau(\epsilon) \leq \frac{1}{\log(4) - \log(\|W\|_1\|W^T\|_1)} \log\left(\frac{\min(n,m)}{\epsilon}\right).$$

where $n = n_1 + n_3 + \cdots$ is the total number of nodes in the odd layers and $m = n_2 + n_4 + \cdots$ is the total number of nodes in the even layers.

The matrix W is far more structured in the setting of DBMs than in the setting of general RBMs, with most of its entries take the value 0. For example, if $K = 2M$, then

$$\|W\|_1 = \max_{1\leq k\leq M} \max_{t\in n_{2k}} \sum_{i=1}^{n_{2k-1}} |W_{it}^{(2k-1)}| + \sum_{j=1}^{n_{2k}} |W_{tj}^{(2k)}|$$

$$\|W^T\|_1 = \max_{0\leq k\leq M} \max_{t\in n_{2k+1}} \sum_{i=1}^{n_{2k}} |W_{it}^{(2k)}| + \sum_{j=1}^{n_{2k+1}} |W_{tj}^{(2k+1)}|$$

where $W^{(0)}$ and $W^{(2M+1)}$ are taken to be zero matrices of the appropriate dimensions. Thus $\|W\|_1\|W^T\|_1 < 4$ is a much less restrictive requirement in the case of DBMs than it is for general RBMs.

3.3. Softmax RBMs

Another generalization of RBMs is to replace the binary logistic sigmoid units with K-ary softmax units. In this setting, the n visible units take values in $[K] = \{1, \dots, K\}$ and the m hidden units take values in $\{0, 1\}$. Further, there are K weight matrices $W^{(1)}, \dots, W^{(K)}$, K visible bias vectors $a^{(1)}, \dots, a^{(K)}$, and a hidden bias vector b. Given $x \in \Omega$, the conditional distribution of a hidden node h_j is

$$P_{\mathrm{S}}^{(h)}(X(h_j) = 1 \,|\, x(v)) = \sigma(b_j + \sum_{i,k} W_{ij}^{(k)} \mathbb{1}[x(v_i) = k]).$$

For a visible node v_i, the conditional distribution is

$$P_{\mathrm{S}}^{(v)}(X(v_i) = k \,|\, x(h)) = \frac{e^{a_i^{(k)} + \sum_j x(h_j) W_{ij}^{(k)}}}{\sum_{k'=1}^{K} e^{a_i^{(k')} + \sum_j x(h_j) W_{ij}^{(k')}}}.$$

Finally, the full conditional distributions of the hidden and visible nodes are simply the product distributions. Define $W \in \mathbb{R}^{n\times m}$ as the matrix with entries

$$W_{ij} = \max_{k,k'} \left| W_{ij}^{(k)} - W_{ij}^{(k')} \right|.$$

Because there is no a priori relationship between the values in $[K]$ or in $\{0, 1\}$, we will again use Hamming distance for both visible and hidden distances. The following lemma, which is proven in the appendix, establishes the contractivity of our conditional distributions.

Lemma 7. *$P_S^{(h)}$ and $P_S^{(v)}$ are $\frac{1}{2}\|W^T\|_1$- and $\frac{1}{2}\binom{K}{2}\|W\|_1$-contractive, respectively.*

We then have the following corollary.

Corollary 8. *The mixing rate for the Gibbs sampler over a softmax RBM whose matrices satisfies $\binom{K}{2}\|W\|_1\|W^T\|_1 < 4$ is upper bounded as*

$$\tau(\epsilon) \leq \frac{1}{\log(4) - \log\left(\binom{K}{2}\|W\|_1\|W^T\|_1\right)} \log\left(\frac{\min(n,m)}{\epsilon}\right).$$

In the case where $K = 2$, the Softmax RBM is the original RBM in disguise. Identifying the state 1 with 0 and the state 2 with 1, taking $W = W^{(2)} - W^{(1)}$, and taking $a = a^{(2)} - a^{(1)}$ gives us the RBM conditional distributions. Thus, Corollary 8 is a generalization of Corollary 5.

4. The General Case

We now turn our attention to a more general setting. Suppose that our vectors v and h take values in spaces Ω_v and Ω_h equipped with semimetrics $d_v(\cdot,\cdot)$ and $d_h(\cdot,\cdot)$ that do not have a minimum distance for distinct elements, i.e.

$$\inf_{x\neq y} d_v(x,y) = 0 = \inf_{x\neq y} d_h(x,y).$$

In this case, we cannot hope to apply Theorem 3 even if we could bound the diameter of Ω_v and Ω_h. Contractivity of the conditional distributions alone is not sufficient to guarantee rapid convergence in total variation distance.

To guarantee rapid mixing, we will require another property of one of our conditional distributions. For convenience, we will use the visible conditional distribution.

Definition 9. *We say that $P^{(v)}$ is (ϵ,δ,M)-*gamble admissible *if for any $x,y\in\Omega$, there exists a coupling (X,Y) of $P^{(v)}(\cdot\,|\,x(h))$ and $P^{(v)}(\cdot\,|\,y(h))$ such that*

(i) $Pr(X\neq Y\,|\,d_h(x,y)\leq\epsilon)\leq\delta$.

(ii) $\mathbb{E}\left[d_v(X,Y)\,|\,d_h(x,y)\leq\epsilon, X(v)\neq Y(v)\right]\leq M$.

(iii) $Pr(X\neq Y\,|\,x(h)=y(h))=0$.

We call a coupling (X,Y) that satisfies conditions (i)-(iii) a (ϵ,δ,M)-*gamble coupling*. As opposed to the contractive couplings given in Section 3, a gamble coupling aims to set $X=Y$ instead of simply shrinking $d_v(X,Y)$. In particular, if $d_h(x,y)$ is small enough (less than ϵ), then condition (i) guarantees that $X=Y$ with probability at least $1-\delta$. On the other hand, in the event that $X\neq Y$, condition (ii) guarantees that the expected distance between X and Y is not too large. Finally, condition (iii) guarantees that if $P^{(v)}(\cdot\,|\,x(h))=P^{(v)}(\cdot\,|\,y(h))$, then X and Y will be the same with probability one.

The following lemma says that if both conditional distributions are contractive and one is gamble admissible, then these couplings can be interleaved in such a way to produce a Markovian coupling whose time to couple is small.

Lemma 10. *Let $c_1,c_2,\epsilon_0,\delta_0,M>0$ such that $c_1c_2<1$, $P^{(h)}$ is c_1-contractive, $P^{(v)}$ is c_2-contractive and (ϵ_0,δ_0,M)-gamble admissible, then there exists a Markovian coupling (X_t,Y_t) s.t. if $\mathbb{E}[d_v(X_0,Y_0)]\leq M$, then for any $\delta>0$, if*

$$t\geq\frac{\log(2/\delta)}{\log(1/c_1c_2)\log(1/\delta_0)}\log\left(\frac{2c_1M}{\delta\epsilon_0}\cdot\frac{\log(2/\delta)}{\log(1/\delta_0)}\right)$$

we have $Pr(X_t(v)\neq Y_t(v))\leq\delta$.

The strategy for proving Lemma 10 is to use our contractive coupling until $d_h(X_s,Y_s)\leq\epsilon_0$ and then apply our gamble coupling. We will succeed with probability $1-\delta_0$, but even if we fail we are no worse off than when we started in expectation. Therefore, we can repeat this process until we achieve convergence, roughly $\frac{\log(1/\delta)}{\log(1/\delta_0)}$ times. We present the full proof in the appendix.

Unfortunately, we can not simply use Lemma 10 along with Lemma 2 to get upper bounds on the mixing rate due to the unbounded nature of our state space. That is, so long as our conditional distributions have contractivity greater than 0, for any $T\in\mathbb{N}$ and $\delta\in(0,1)$, there may exist an initial pair of states x,y such that $Pr(X_T\neq Y_T\,|\,X_0=x,Y_0=y)>1-\delta$ under the coupling (X_t,Y_t) in Lemma 10.

Therefore, to get bounds on the rate of convergence, we assume that the initial state of the alternating Gibbs sampler is close enough to a random state drawn from the stationary distribution in expectation. When this assumption is made, the following theorem tells us how quickly we converge to the stationary distribution.

Theorem 11. *Let c_1, c_2, ϵ_0, δ_0, M, $P^{(h)}$, and $P^{(v)}$ satisfy the conditions of Lemma 10. If X_t is the Gibbs sampler whose initial state X_0 satifies $\mathbb{E}[d_v(X_0,Y)]\leq M$ where Y is drawn independently from the stationary distribution π, then for $\delta>0$ and any t satisfying*

$$t\geq 1+\frac{\log(2/\delta)}{\log(1/c_1c_2)\log(1/\delta_0)}\log\left(\frac{2c_1M}{\delta\epsilon_0}\cdot\frac{\log(2/\delta)}{\log(1/\delta_0)}\right)$$

we have $\|X_t-\pi\|_{TV}\leq\delta$.

Proof. Let (X_s,Y_s) be the Markovian coupling from Lemma 10. Say $Y_0\sim\pi$ independently from X_0, then $Y_0,Y_1,\ldots\sim\pi$. Further, if at time $S\geq 0$ we have $X_S(v)=Y_S(v)$, then for any time $s\geq S+1$ we have $X_s=Y_s$. Therefore for t satisfying our lower bound and for any measurable subset $A\subset\Omega$,

$$\begin{aligned}Pr(X_t\in A)&\geq Pr(X_t=Y_t,Y_t\in A)\\&\geq 1-(Pr(X_t\neq Y_t)+Pr(Y_t\notin A))\\&\geq Pr(Y_t\in A)-Pr(X_{t-1}(v)\neq Y_{t-1}(v))\\&\geq\pi(A)-\delta.\end{aligned}$$

Where we used Lemma 10 to bound $Pr(X_{t-1}(v)\neq Y_{t-1}(v))$. Since the above holds for any measurable subset A, we can conclude $\|X_t-\pi\|_{TV}\leq\delta$. □

4.1. Gaussian RBMs

We now turn our attention to two special cases of continuous-valued RBMs: Gaussian-Gaussian RBMs and Gaussian-NReLU RBMs. In both cases, our configurations take values in $\mathbb{R}$.

For the Gaussian-Gaussian RBM, we have a weight matrix $W\in\mathbb{R}^{n\times m}$, bias vectors $a\in\mathbb{R}^n$ and $b\in\mathbb{R}^m$, variance

vectors $\sigma^2 \in \mathbb{R}^n$ and $s^2 \in \mathbb{R}^m$, and the conditional distributions are all independent normal:

$$P_{GG}^{(v)}(X(v_i) \,|\, x(h)) = \mathcal{N}(a_i + \sum_{j=1}^{m} W_{ij}\, x(h_j),\, \sigma_i^2)$$

$$P_{GG}^{(h)}(X(h_j) \,|\, x(v)) = \mathcal{N}(b_j + \sum_{i=1}^{n} W_{ij}\, x(v_i),\, s_j^2)$$

For the Gaussian-NReLU RBM, the parameters W, a, and σ^2, and the conditional distribution for the visible nodes $P_{GN}^{(v)}$ are the same as in the Gaussian-Gaussian RBM. However, the hidden conditional distribution $P_{GN}^{(h)}$ has changed so that all $X(h_j)$ are independently distributed according to the noisy rectified linear distribution $\mathcal{R}(\sum_{i=1}^n W_{ij}\, x(v_i))$ (Nair & Hinton, 2010), where if $Z \sim \mathcal{N}(z, \sigma(z))$, then $\max(0, Z)$ is distributed according to $\mathcal{R}(z)$.

For both cases, the visible and hidden semimetrics that we will use will be ℓ_2^2-distance, i.e. for configurations x, y, $d_v(x,y) = \sum_{i=1}^n (x(v_i) - y(v_i))^2$. Similarly for $d_h(x,y)$. The following lemma, whose proof appears in the appendix, establishes contractivity and gamble-admissibility for the conditional distributions we have defined.

Lemma 12. *The following holds.*

(a) $P_{GG}^{(v)}$, $P_{GG}^{(h)}$, $P_{GN}^{(v)}$ *are* $\|W\|_F^2$*-contractive.*

(b) $P_{GN}^{(h)}$ *is* $\frac{5}{4}\|W\|_F^2$*-contractive.*

(c) $P_{GG}^{(v)}$ *and* $P_{GN}^{(v)}$ *are* $(\epsilon_0, \delta_0, M)$*-gamble admissible for* $\epsilon_0 = \frac{1}{4\|(W/\sigma)^T\|_{2,1}^2}$, $\delta_0 = 1/4$, *and*

$$M = 4\|\sigma\|_2^2 + \sqrt{\frac{2}{\pi}}\frac{\left\|(W\sigma)^T\right\|_{2,1}}{\|(W/\sigma)^T\|_{2,1}} + \left(\frac{\|W\|_F}{2\|(W/\sigma)^T\|_{2,1}}\right)^2$$

where W/σ *and* $W\sigma$ *denote* $n \times m$ *matrices whose entries are* W_{ij}/σ_i *and* $W_{ij}\sigma_i$, *respectively*

Lemma 12 and Theorem 11 imply the following corollary on the mixing rate for the alternating Gibbs sampler over Gaussian-Gaussian RBMs and Gaussian-NReLU RBMs.

Corollary 13. *Let* M *be the quantity given in Lemma 12. Let* X_t *denote the Gibbs sampler for the Gaussian-Gaussian RBM with stationary distribution* π_X *and* Y_t *denote the Gibbs sampler for the Gaussian-NReLU RBM with stationary distribution* π_Y. *If there exists* $M^* > 0$ *s.t.*

$$\max\left(\mathbb{E}_{X\sim\pi_X}\left[d_v(X_0, X)\right], \mathbb{E}_{Y\sim\pi_Y}\left[d_v(Y_0, Y)\right], M\right) \leq M^*$$

then for $\delta > 0$ *and* $C = \frac{M^*\|(W/\sigma)^T\|_{2,1}^2\|W\|_F^2\log(\frac{2}{\delta})}{\delta\log(4)}$,

(a) if $\|W\|_F \leq 1$ *and*

$$t \geq 1 + \frac{\log(2/\delta)\log(8C)}{\log\left(\frac{1}{\|W\|_F^4}\right)\log(4)}$$

then $\|X_t - \pi_X\|_{TV} \leq \delta$, *and*

(b) if $\|W\|_F^4 \leq 4/5$ *and*

$$t \geq 1 + \frac{\log(2/\delta)\log(10C)}{\log\left(\frac{4}{5\|W\|_F^4}\right)\log(4)}$$

then $\|Y_t - \pi_Y\|_{TV} \leq \delta$.

5. Lower Bounds

We now turn our attention towards providing lower bounds for the mixing rate of the alternating Gibbs sampler. To do so, we will use the the method of conductance. Given a Markov chain Q over a state space Ω and its stationary distribution π, the *conductance* of $S \subset \Omega$ is

$$\Phi(S) := \frac{1}{\pi(S)} \sum_{x\in S, y\in S^c} \pi(x)Q(x,y)$$

and the conductance of Q, denoted by Φ^*, is the minimum conductance of any set S with $\pi(S) \leq 1/2$. The following theorem, due to Sinclair (1988), relates the mixing rate and conductance of a Markov chain.

Theorem 14 (Sinclair (1988))**.** *For any aperiodic, irreducible Markov chain with conductance* Φ^*, $\tau_{mix} \geq \frac{1}{4\Phi^*}$.

Our first lower bound is for the case of RBMs.

Theorem 15. *Pick any* $T > 0$ *and* $n, m \in \mathbb{N}$ *even positive integers. Then there is a weight matrix* $W \in \mathbb{R}^{n\times m}$ *satisfying* $\|W\|_{max} \leq \frac{2}{\min(n,m)}\ln(4T(n+m))$ *such that the Gibbs sampler over the RBM with zero bias and weight matrix* W *has mixing rate bounded as* $\tau_{mix} \geq T$.

The proof of Theorem 15 appears in the appendix, but we present the main idea here. We construct a weight matrix W such that the energy function associated with W has two antipodal global minima. Because there are two minima, the singleton set consisting of one minima has probability mass less than 1/2 under the Gibbs distribution. Escaping from one of these minima is a very unlikely event, which implies that the conductance is small.

Our second lower bound is for the case of Gaussian-Gaussian RBMs. The state space of a Gaussian-Gaussian RBM is unbounded, but any implementation of the Gibbs sampler is necessarily in a bounded state space. Therefore, lower bounds that exploit the unbounded nature of the state space may not be particularly meaningful. To compensate for this, we work with a restricted version of the alternating Gibbs sampler. Given $B > 0$, consider the following B*-thresholded* alternating Gibbs sampler $(Y_t)_{t=0}^\infty$. At time step t, it performs the following.

1. For each hidden node h_j, draw $X_t(h_j) \sim \mathcal{N}(b_j + \sum_{i=1}^n W_{ij}\, Y_{t-1}(v_i),\, s_j^2)$. Set $Y_t(h_j)$ to be the closest point in $[-B, B]$ to $X_t(h_j)$.

2. For each hidden node h_j, draw $X_t(v_i) \sim \mathcal{N}(a_i + \sum_{j=1}^m W_{ij}\, Y_t(h_j), \sigma_i^2)$. Set $Y_t(v_i)$ to be the closest point in $[-B, B]$ to $X_t(v_i)$.

The following theorem, whose proof appears in the appendix, gives a lower bound on the mixing rate for this restricted Markov chain.

Theorem 16. *Let $T, B > 0$ and $n, m \in \mathbb{N}$ be even positive integers. Then there exists weight matrix $W \in \mathbb{R}^{n \times m}$ s.t.*

$$\|W\|_{max} \leq \frac{1}{\min(n,m)} \left(1 + \frac{1}{B}\sqrt{8\log(4T\max(n,m))}\right)$$

such that the B-truncated chain of the Gibbs sampler for the Gaussian-Gaussian RBM with no biases and unit variances mixes in time $\tau_{mix} \geq T$.

In the case where $n = m$, the restriction on W translates to a $1 + \frac{1}{B}\sqrt{8\log(4Tn)}$ upper bound on the Frobenius norm of W. This implies that for any $\epsilon, T > 0$, there exists a $B > 0$ and a weight matrix W such that $\|W\|_F \leq 1 + \epsilon$, but the alternating Gibbs sampler mixes in time bounded below by T. In this sense, the condition on the Frobenius norm of W given in Corollary 13(a) is tight for establishing finite convergence rates on the alternating Gibbs sampler over Gaussian-Gaussian RBMs.

6. Complexity of RBMs

The results in the previous sections give conditions under which a particular algorithm, the alternating Gibbs sampler, can efficiently sample from the Gibbs distributions of RBMs and several of its variants. It is natural to ask how much better can we hope to do with either a better analysis of the Gibbs sampler or a different algorithm altogether.

The complexity of approximately sampling from a distribution is often closely tied to the complexity of approximately computing its normalizing constant or partition function (Jerrum et al., 1986; Long & Servedio, 2010). Therefore, to help understand the complexity of sampling from the Gibbs distribution over RBMs, we will focus on the complexity of computing approximate solutions to the following problem.

Name: #RBM
Instance: Parameters $W \in \mathbb{R}^{n \times m}$, $a \in \mathbb{R}^n$, and $b \in \mathbb{R}^m$.
Output: The partition function

$$Z = \sum_{x\,:\,(v,h) \to \{0,1\}^{n+m}} e^{a^T x(v) + b^T x(h) + x(v)^T W x(h)}.$$

In the complexity literature, there are three well-documented categories that an approximate counting problem can be placed in. The first category consists of problems for which we have an efficient algorithm to approximately count or compute a partition function. The second category consists of problems for which an efficient approximate counting algorithm would imply the equivalence of two complexity classes widely viewed to be distinct, such as P and NP. Finally, problems in the third category do not belong to either of the above categories but often are placed in well-defined classes of possibly intermediate computational complexity. As we shall see, #RBM exhibits flavors of all three of these categories.

Jerrum and Sinclair (1993) showed that when all weights are positive and all biases are consistent, there is an efficient algorithm to approximate #RBM. Moreover, we can combine our results from Section 3 with annealing techniques (Štefankovič et al., 2009) to get an efficient algorithm for the general case when $\|W\|_1\|W^T\|_1 < 4$. Putting this all together, we have the following result which place certain instances of #RBM into the first category.

Theorem 17. (Jerrum and Sinclair (1993), this paper, Štefankovič et al. (2009)) *#RBM admits an efficient algorithm in both of the following cases.*

(i) $\forall\,(i,j) \in [n] \times [m]$, $W_{ij} \geq 0$ *and* $sign(a_i) = sign(b_j)$.

(ii) $\|W\|_1\|W^T\|_1 < 4$.

On the other hand, Long and Servedio (2010) showed that when the max-norm of the weight matrix grows quickly enough, #RBM falls into the second category.

Theorem 18 (Long and Servedio (2010))**.** *There is a universal constant $\alpha > 0$ such that if $P \neq NP$, then there is no polynomial-time algorithm such that given an $n \times n$ matrix W such that $\|W\|_{max} \leq \psi(n) = \omega(n)$[1], the algorithm approximates #RBM with weight matrix W and no bias to within a multiplicative factor of $e^{\alpha\psi(n)}$.*

Finally, Goldberg and Jerrum (2007) showed that when the weights are constrained to be positive but the biases may be arbitrary, #RBM falls into the third category. Formally, they showed that it belongs to a class of problems introduced by Dyer et al. (2004) that are approximation-preserving interreducible[2] with the problem of counting independent sets in bipartite graphs (#BIS).

Theorem 19. (Goldberg and Jerrum (2007)) *#BIS $\equiv_{AP}$ #RBM when $W_{ij} \geq 0$ and $\|W\|_1\|W^T\|_1 = \Omega(n^2)$.*

Theorems 18 and 19 both imply that for large values of $\|W\|_1\|W^T\|_1$, it seems unlikely that we will be able to sample from the Gibbs distribution over RBMs, even when all the weights are constrained to be positive. On the other hand, Theorem 17 gives hope that there are cases when we can succeed. However, there are large gaps in the cases that we know can be efficiently solved and those in which we believe that they cannot. Closing these gaps remains an interesting direction for future research.

[1] Two functions $f(n), g(n)$ satisfy the relationship $g(n) = \omega(f(n))$ if $\lim_{n\to\infty} \frac{g(n)}{f(n)} = \infty$.

[2] For a precise definition of approximation-preserving reducibility, see (Dyer et al., 2004).

Acknowledgments

The author would like to thank Sanjoy Dasgupta and the anonymous reviewers for insightful feedback as well as the NSF for support under grant DGE-1144086.

References

Aldous, David. Random walks on finite groups and rapidly mixing markov chains. In *Séminaire de Probabilités XVII 1981/82*, pp. 243–297. Springer, 1983.

Bengio, Yoshua. Learning deep architectures for ai. *Foundations and trends in Machine Learning*, 2(1):1–127, 2009.

Besag, Julian. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 192–236, 1974.

Bulatov, Andrei and Grohe, Martin. The complexity of partition functions. *Theoretical Computer Science*, 348(2): 148–186, 2005.

De Sa, Christopher, Zhang, Ce, Olukotun, Kunle, and Ré, Christopher. Rapidly mixing gibbs sampling for a class of factor graphs using hierarchy width. In *Advances in Neural Information Processing Systems*, pp. 3079–3087, 2015.

Dyer, Martin, Goldberg, Leslie Ann, Greenhill, Catherine, and Jerrum, Mark. The relative complexity of approximate counting problems. *Algorithmica*, 38(3):471–500, 2004.

Geman, Stuart and Geman, Donald. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 721–741, 1984.

Goldberg, Leslie Ann and Jerrum, Mark. The complexity of ferromagnetic ising with local fields. *Combinatorics, Probability and Computing*, 16(01):43–61, 2007.

Gotovos, Alkis, Hassani, Hamed, and Krause, Andreas. Sampling from probabilistic submodular models. In *Advances in Neural Information Processing Systems*, pp. 1936–1944, 2015.

Hammersley, John and Clifford, Peter. Markov fields on finite graphs and lattices. 1971.

Hinton, Geoffrey E, Osindero, Simon, and Teh, Yee-Whye. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

Ising, Ernst. A contribution to the theory of ferromagnetism. *z. Phys*, 31(1):253–258, 1925.

Jawitz, James W. Moments of truncated continuous univariate distributions. *Advances in water resources*, 27 (3):269–281, 2004.

Jerrum, Mark. *Counting, sampling and integrating: algorithms and complexity*. Springer Science & Business Media, 2003.

Jerrum, Mark and Sinclair, Alistair. Polynomial-time approximation algorithms for the ising model. *SIAM Journal on computing*, 22(5):1087–1116, 1993.

Jerrum, Mark, Valiant, Leslie, and Vazirani, Vijay. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43: 169–188, 1986.

Levin, David A., Peres, Yuval, and Wilmer, Elizabeth L. *Markov Chains and Mixing Times*. American Mathematical Society, 2008.

Liu, Xianghang and Domke, Justin. Projecting markov random field parameters for fast mixing. In *Advances in Neural Information Processing Systems*, pp. 1377–1385, 2014.

Long, Philip M. and Servedio, Rocco A. Restricted boltzmann machines are hard to approximately evaluate or simulate. In *ICML*, pp. 703–710. Omnipress, 2010.

Nair, Vinod and Hinton, Geoffrey. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 807–814, 2010.

Potts, Renfrey Burnard. Some generalized order-disorder transformations. In *Mathematical proceedings of the cambridge philosophical society*, volume 48, pp. 106–109. Cambridge Univ Press, 1952.

Roberts, Gareth and Rosenthal, Jeffrey. General state space markov chains and mcmc algorithms. *Probability Surveys*, 1:20–71, 2004.

Rüschendorf, L and Rachev, S. T. A characterization of random variables with minimum l2-distance. *Journal of Multivariate Analysis*, 32(1):48–54, 1990.

Salakhutdinov, Ruslan and Hinton, Geoffrey. Deep boltzmann machines. In *International conference on artificial intelligence and statistics*, pp. 448–455, 2009.

Sinclair, Alistair. *Randomised Algorithms for Counting and Generating Combinatorial Structures*. PhD thesis, University of Edinburgh, 1988.

Štefankovič, Daniel, Vempala, Santosh, and Vigoda, Eric. Adaptive simulated annealing: A near-optimal connection between sampling and counting. *Journal of the ACM (JACM)*, 56(3):18, 2009.

Wang, Neng-Yi and Wu, Liming. Convergence rate and concentration inequalities for gibbs sampling in high dimension. *Bernoulli*, 20(4):1698–1716, 2014.

Appendix: Proof Details

Proofs from Section 3

Theorem 3. *Let $c_1, c_2 \geq 0$ such that $c_1 c_2 < 1$, $P^{(v)}$ is c_1-contractive, and $P^{(h)}$ is c_2-contractive. Then the mixing rate of the Gibbs sampler is bounded as*

$$\tau(\epsilon) \leq 1 + \frac{1}{\log(1/c_1 c_2)} \log\left(\frac{C}{\epsilon}\right)$$

where $C = \min\left(\frac{\gamma_v^{(max)}}{\gamma_v^{(min)}}, \frac{\gamma_h^{(max)}}{\gamma_h^{(min)}}, \frac{c_2 \gamma_v^{(max)}}{\gamma_h^{(min)}}\right)$.

Proof. To see $\tau(\epsilon) \leq 1 + \frac{1}{\log(1/c_1 c_2)} \log\left(\frac{1}{\epsilon} \min\left(\frac{\gamma_h^{(\max)}}{\gamma_h^{(\min)}}, \frac{c_2 \gamma_v^{(\max)}}{\gamma_h^{(\min)}}\right)\right)$, we will use the same coupling (X_t, Y_t) as given in the first part of the proof. Then by similar arguments,

$$\begin{aligned}
Pr(X_t \neq Y_t) &\leq Pr(d_h(X_t, Y_t) \geq \gamma_h^{(\min)}) \\
&\leq \frac{\mathbb{E}[d_h(X_t, Y_t)]}{\gamma_h^{(\min)}} \\
&\leq \frac{(c_1 c_2)^{t-1} \mathbb{E}[d_h(X_1, Y_1)]}{\gamma_h^{(\min)}} \\
&\leq \frac{(c_1 c_2)^{t-1} \min(\gamma_h^{(\max)}, c_2 \gamma_v^{(\max)})}{\gamma_h^{(\min)}}
\end{aligned}$$

Taking $t \geq 1 + \frac{1}{\log(c_1 c_2)} \log\left(\frac{\min(\gamma_h^{(\max)}, c_2 \gamma_v^{(\max)})}{\gamma_v^{(\min)}}\right)$ makes the above less than ϵ. Applying Lemma 2 completes the proof. □

Lemma 4. *$P_{RBM}^{(v)}$ and $P_{RBM}^{(h)}$ are $\frac{\|W\|_1}{2}$- and $\frac{\|W^T\|_1}{2}$-contractive, respectively.*

Proof. Let $x, y \in \Omega$ be two configurations. We will prove the claim for the visible conditional distributions. The proof for the hidden conditional distributions will follow symmetrically.

For each visible node v_i, let $(X(v_i), Y(v_i))$ be the maximal coupling of $P^{(v)}(X(v_i) \,|\, x(h))$ and $P^{(v)}(Y(v_i) \,|\, y(h))$ guaranteed in Lemma 1. By doing this independently for all visible nodes, we have a valid coupling (X, Y) of $P^{(v)}(\cdot \,|\, x(h))$ and $P^{(v)}(\cdot \,|\, y(h))$. Then we can work out the expected Hamming distance of X and Y as

$$\begin{aligned}
\mathbb{E}[d_v(X, Y)] &= \sum_{i=1}^{n} \left\| P^{(v)}(X(v_i) \,|\, x(h)) - P^{(v)}(Y(v_i) \,|\, y(h)) \right\|_{TV} \\
&= \sum_{i=1}^{n} \left| P^{(v)}(X(v_i) = 1 \,|\, x(h)) - P^{(v)}(Y(v_i) = 1 \,|\, y(h)) \right| \\
&= \sum_{i=1}^{n} \left| \frac{1}{1 + \exp\left(-a_i - \sum_{j=1}^{m} W_{ij} x(h_j)\right)} - \frac{1}{1 + \exp\left(-a_i - \sum_{j=1}^{m} W_{ij} y(h_j)\right)} \right| \\
&\leq \sum_{i=1}^{n} \left| \frac{1 - \exp\left(\sum_{j=1}^{m} W_{ij} \left(y(h_j) - x(h_j)\right)\right)}{1 + \exp\left(\sum_{j=1}^{m} W_{ij} \left(y(h_j) - x(h_j)\right)\right)} \right| \\
&= \sum_{i=1}^{n} \left| \tanh\left(\frac{\sum_{j=1}^{m} W_{ij} \left(Y_{t+1/2}(h_j) - X_{t+1/2}(h_j)\right)}{2} \right) \right| \\
&\leq \sum_{i=1}^{n} \frac{1}{2} \left| \sum_{j=1}^{m} W_{ij} \left(y(h_j) - x(h_j)\right) \right|
\end{aligned}$$

$$\leq \frac{1}{2} \sum_{j \,:\, y(h_j) \neq x(h_j)} \sum_{i=1}^{n} |W_{ij}|$$
$$\leq \frac{1}{2} \|W\|_1 d_h(x, y).$$

□

Lemma 7. *$P_S^{(h)}$ and $P_S^{(v)}$ are $\frac{1}{2}\|W^T\|_1$- and $\frac{1}{2}\binom{K}{2}\|W\|_1$-contractive, respectively.*

Proof. We will first show that $P_S^{(h)}$ is $\frac{1}{2}\|W^T\|_1$-contractive. To do so, let $x, y \in \Omega$ be two configurations. Our coupling (X, Y) of $P_S^{(h)}(\cdot \,|\, x(v))$ and $P_S^{(h)}(\cdot \,|\, y(v))$ is exactly the same as the coupling given in the proof of Lemma 4. Then, from the proof of Lemma 4, we have

$$\begin{aligned}
\mathbb{E}[d_h(X, Y) \,|\, x(v), y(v)] &= \sum_{j=1}^{m} \left| P_S^{(h)}(X(h_j) = 1 \,|\, x(v)) - P_S^{(h)}(Y(h_j) = 1 \,|\, y(v)) \right| \\
&\leq \left| \frac{1 - \exp\left(\sum_{i=1}^{n} \sum_{k=1}^{K} W_{ij}^{(k)} (\mathbb{1}[y(v_i) = k] - \mathbb{1}[x(v_i) = k]) \right)}{1 + \exp\left(\sum_{i=1}^{n} \sum_{k=1}^{K} W_{ij}^{(k)} (\mathbb{1}[y(v_i) = k] - \mathbb{1}[x(v_i) = k]) \right)} \right| \\
&= \left| \tanh\left(\frac{\sum_{i=1}^{n} \sum_{k=1}^{K} W_{ij}^{(k)} (\mathbb{1}[y(v_i) = k] - \mathbb{1}[x(v_i) = k])}{2} \right) \right| \\
&\leq \frac{1}{2} \left| \sum_{i=1}^{n} \sum_{k=1}^{K} W_{ij}^{(k)} (\mathbb{1}[y(v_i) = k] - \mathbb{1}[x(v_i) = k]) \right| \\
&\leq \frac{1}{2} \sum_{i \,:\, x(v_i) \neq y(v_i)} \sum_{j=1}^{m} W_{ij} \\
&\leq \frac{1}{2} \|W^T\|_1 d_v(x, y).
\end{aligned}$$

To prove $P_S^{(v)}$ is $\frac{1}{2}\binom{K}{2}\|W\|_1$-contractive, we will again use Lemma 1 to construct independent couplings $(X(v_i), Y(v_i))$ of $P_S^{(v)}(v_i \,|\, x(h))$ and $P_S^{(v)}(v_i \,|\, y(h))$ for each visible node v_i. Then by Lemma 1, we have

$$\begin{aligned}
\mathbb{E}[d_v(X, Y) \,|\, x(h), y(h)] &= \sum_{i=1}^{n} \|P_S^{(v)}(X(v_i) \,|\, x(h)) - P_S^{(v)}(Y(v_i) \,|\, y(h))\|_{TV} \\
&= \sum_{i=1}^{n} \frac{1}{2} \sum_{k=1}^{K} |P_S^{(v)}(X(v_i) = k \,|\, x(h)) - P_S^{(v)}(Y(v_i) = k \,|\, y(h))| \\
&\leq \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} \left| \frac{1 - \sum_{k' \neq k} \exp\left(b^{(k')} - b^{(k)} + \sum_{j=1}^{m} (W_{ij}^{(k')} - W_{ij}^{(k)})(y(h_j) - x(h_j)) \right)}{1 + \sum_{k' \neq k} \exp\left(b^{(k')} - b^{(k)} + \sum_{j=1}^{m} (W_{ij}^{(k')} - W_{ij}^{(k)})(y(h_j) - x(h_j)) \right)} \right| \\
&= \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{k' \neq k} \left| \tanh\left(\frac{\sum_{j=1}^{m} (W_{ij}^{(k')} - W_{ij}^{(k)})(y(h_j) - x(h_j))}{2} \right) \right| \\
&\leq \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{k' \neq k} \frac{1}{2} \left| \sum_{j=1}^{m} (W_{ij}^{(k')} - W_{ij}^{(k)})(y(h_j) - x(h_j) \right|
\end{aligned}$$

$$\leq \frac{1}{2} \sum_{j\,:\,x(h_j)\neq y(h_j)} \frac{K(K-1)}{2} \sum_{i=1}^{n} W_{ij}$$
$$= \frac{1}{2}\binom{K}{2}\|W\|_1 d_h(x,y).$$

□

Proofs from Section 4

Lemma 10. *Let $c_1, c_2, \epsilon_0, \delta_0, M > 0$ such that $c_1 c_2 < 1$, $P^{(h)}$ is c_1-contractive, $P^{(v)}$ is c_2-contractive and $(\epsilon_0, \delta_0, M)$-gamble admissible, then there exists a Markovian coupling (X_t, Y_t) s.t. if $\mathbb{E}[d_v(X_0, Y_0)] \leq M$, then for any $\delta > 0$, if*

$$t \geq \frac{\log(2/\delta)}{\log(1/c_1c_2)\log(1/\delta_0)} \log\left(\frac{2c_1 M}{\delta\epsilon_0} \cdot \frac{\log(2/\delta)}{\log(1/\delta_0)}\right)$$

we have $Pr(X_t(v) \neq Y_t(v)) \leq \delta$.

Proof. Let (X_s, Y_s) be the *interleaved coupling* whose initial state is (X_0, Y_0) and is evolved according to the following rule.

1. Draw $(X_{s+1}(h), Y_{s+1}(h))$ according to the c_1-contractive coupling of $P^{(h)}(\cdot \,|\, X_s(v))$ and $P^{(h)}(\cdot \,|\, Y_s(v))$.
2. If $d_h(X_{s+1}, Y_{s+1}) \leq \epsilon_0$, draw $(X_{s+1}(v), Y_{s+1}(v))$ according to the $(\epsilon_0, \delta_0, M)$-gamble coupling of $P^{(v)}(\cdot \,|\, X_{s+1}(h))$ and $P^{(v)}(\cdot \,|\, Y_{s+1}(h))$. Otherwise, draw $(X_{s+1}(v), Y_{s+1}(v))$ according to the c_2-contractive coupling of $P^{(v)}(\cdot \,|\, X_{s+1}(h))$ and $P^{(v)}(\cdot \,|\, Y_{s+1}(h))$.

It is not too hard to see that (X_s, Y_s) is a Markovian coupling of the alternating Gibbs sampler.

Let us define two stochastic processes $Z_s = d_h(X_{s+1}, Y_{s+1})$, and $S_i = \inf\{s > S_{i-1} \,:\, Z_s \leq \epsilon_0\}$ where $S_0 = 0$. Due to the definition of the interleaved coupling, it is not hard to see that for any finite $i \geq 1$, $S_i < \infty$ with probability one. Moreover, because of the Markovian nature of S_i, we know that given S_{i-1}, S_i is independent of $S_0, S_1, \ldots, S_{i-2}$.

Now let $T, K \geq 1$ be given. Then we can work out the following

$$\begin{aligned}
Pr(X_{KT}(v) \neq Y_{KT}(v)) &= Pr(X_{KT}(v) \neq Y_{KT}(v) \,|\, S_1 \geq T) Pr(S_1 \geq T) + Pr(X_{KT}(v) \neq Y_{KT}(v) \,|\, S_1 \leq T-1) Pr(S_1 \leq T) \\
&\leq Pr(S_1 \geq T) + Pr(X_{KT} \neq Y_{KT} \,|\, S_1 \leq T-1) \\
&\leq Pr(S_1 \geq T) + Pr(S_2 \geq 2T \,|\, S_1 \leq T-1) + Pr(X_{KT} \neq Y_{KT} \,|\, S_1 \leq T-1, S_2 \leq 2T-1) \\
&\vdots \\
&= \overbrace{\sum_{k=1}^{K} Pr(S_k \geq kT \,|\, S_{k-1} \leq (k-1)T-1)}^{a} + \overbrace{Pr(X_{KT} \neq Y_{KT} \,|\, S_1 \leq T-1, \ldots S_K \leq KT-1)}^{b}
\end{aligned}$$

We can bound the above two terms separately. To bound (a), note that for any $1 \leq k \leq K$,

$$\begin{aligned}
Pr(S_k \geq kT \,|\, S_{k-1} \leq (k-1)T-1) &= Pr(d_h(X_{kT+1}, Y_{kT+1}) \geq \epsilon_0 \,|\, S_{k-1} \leq (k-1)T-1) \\
&\leq \frac{\mathbb{E}\left[d_h(X_{kT+1}, Y_{kT+1}) \,|\, S_{k-1} \leq (k-1)T-1, X_{S_{k-1}}(v) \neq Y_{S_{k-1}}(v)\right]}{\epsilon_0} \\
&\leq \frac{c_1}{\epsilon_0}\mathbb{E}\left[d_v(X_{kT}, Y_{kT}) \,|\, S_{k-1} \leq (k-1)T-1, X_{S_{k-1}}(v) \neq Y_{S_{k-1}}(v)\right] \\
&\leq \frac{c_1}{\epsilon_0}\mathbb{E}\left[(c_1c_2)^{kT-S_{k-1}-1} d_v(X_{S_{k-1}}, Y_{S_{k-1}}) \,|\, S_{k-1} \leq (k-1)T-1, X_{S_{k-1}}(v) \neq Y_{S_{k-1}}(v)\right] \\
&\leq \frac{c_1(c_1c_2)^T M}{\epsilon_0}
\end{aligned}$$

To bound (b) we make use of the fact that at each random time S_k we have at least a $1-\delta_0$ chance of setting $X_{S_k}(v) = Y_{S_k}(v)$. Therefore,

$$Pr(X_{KT}(v) \neq Y_{KT}(v) \,|\, S_1 \leq T-1, \ldots S_K \leq KT-1) \leq \delta_0^K.$$

Then for $K = \frac{\log(2/\delta)}{\log(1/\delta_0)}$ and $T = \frac{1}{\log(1/c_1c_2)} \log\left(\frac{2c_1KM}{\delta\epsilon_0}\right)$,

$$Pr(X_{KT}(v) \neq Y_{KT}(v)) \leq \frac{c_1(c_1c_2)^T KM}{\epsilon_0} + \delta_0^K \leq \delta.$$

The lemma follows by our choice of K and T. □

Lemma 20. *(a) There exists a coupling (X,Y) of $\mathcal{N}(\mu_X, \sigma_X^2)$ and $\mathcal{N}(\mu_Y, \sigma_Y^2)$ such that*

$$\mathbb{E}\left[(X-Y)^2\right] = (\mu_X - \mu_Y)^2 + (\sigma_X - \sigma_Y)^2.$$

(b) There exists a coupling (X,Y) of $\mathcal{N}(\mu_X, \sigma^2)$ and $\mathcal{N}(\mu_Y, \sigma^2)$ such that

$$Pr(X \neq Y) \leq \frac{|\mu_X - \mu_Y|}{2\sigma} \text{ and}$$
$$\mathbb{E}\left[(X-Y)^2 \,|\, X \neq Y\right] \leq 4\sigma^2\left[1 + \frac{|\mu_X-\mu_Y|}{\sqrt{2\pi}\sigma} + \left(\frac{|\mu_X-\mu_Y|}{2\sigma}\right)^2\right].$$

Proof. Part (a) follows from a more general result Ruschendorf and Rachev (1990). To prove part (b), we introduce some notation. Let $\bar{\mu} = \frac{\mu_X+\mu_Y}{2}$. Assume w.l.o.g. $\bar{\mu} = 0$, $\mu_X = -\mu$, and $\mu_Y = \mu$ for some $\mu \geq 0$. Let f_X and f_Y denote the p.d.f.'s of $\mathcal{N}(\mu_X, \sigma^2)$ and $\mathcal{N}(\mu_Y, \sigma^2)$, respectively. Now define three more p.d.f.'s:

$$f_S(x) = \frac{\min\left(f_X(x), f_Y(x)\right)}{Z_S} \quad \text{for } x \in \mathbb{R}$$
$$f_U(x) = \frac{f_Y(x) - f_X(x)}{Z_U} \quad \text{for } x \geq 0$$
$$f_L(x) = \frac{f_X(x) - f_Y(x)}{Z_L} \quad \text{for } x \leq 0$$

Here Z_S, Z_U, and Z_L are chosen so that their respective distributions integrate to 1. It is not too hard to work out that

$$Z_S = 2\left(1 - \Phi\left(\frac{\mu}{\sigma}\right)\right) = 1 - \text{erf}\left(\frac{\mu}{\sigma\sqrt{2}}\right) \quad \text{and} \quad Z_U = Z_L = \Phi\left(\frac{\mu}{\sigma}\right) - \Phi\left(-\frac{\mu}{\sigma}\right) = \text{erf}\left(\frac{\mu}{\sigma\sqrt{2}}\right).$$

Here $\Phi(\cdot)$ denotes cumulative distribution function for the standard normal distribution and $\text{erf}(\cdot)$ denotes the error function. Figure 3 helps explain the picture.

Then our coupling is the following.

1. Draw $S \sim f_S$, $U \sim f_U$, $L = -U$.
2. With probability Z_S, set $X = S = Y$.
3. With probability $1 - Z_S$, set $X = L$ and $Y = U$.

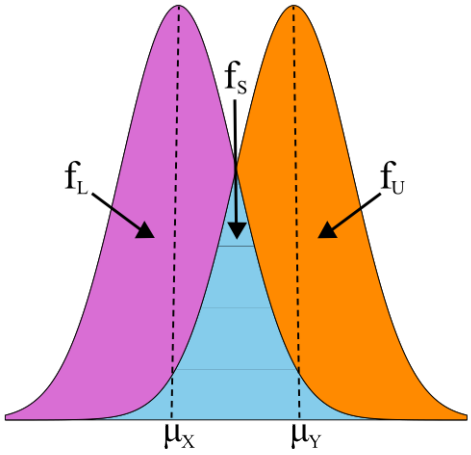

Figure 3. Illustration of the unnormalized densities f_S, f_U, f_L.

It is not hard to see that (X, Y) is a valid coupling of f_X and f_Y. We now turn to the two claims of this coupling. The first is easy:

$$Pr(X \neq Y) = 1 - Z_S = \text{erf}\left(\frac{\mu}{\sigma\sqrt{2}}\right) = \text{erf}\left(\frac{|\mu_X - \mu_Y|}{2\sigma\sqrt{2}}\right) \leq \frac{|\mu_X - \mu_Y|}{2\sigma}.$$

Now we turn to the second claim. To handle this, we will first introduce two more random variables.

- Let U_X be distributed according to $f_{U_X}(x) = \frac{f_X(x)}{1-\Phi(\frac{\mu}{\sigma})}$.
- Let U_Y be distributed according to $f_{U_Y}(x) = \frac{f_Y(x)}{1-\Phi(-\frac{\mu}{\sigma})}$.

Then we can rewrite our objective to bound as

$$\mathbb{E}[(X-Y)^2 \,|\, X \neq Y] = \mathbb{E}[(U-L)^2] = 4\mathbb{E}[U^2] = \frac{4}{Z_U}\left[\left(1 - \Phi\left(-\frac{\mu}{\sigma}\right)\right)\mathbb{E}[U_Y^2] - \left(1 - \Phi\left(\frac{\mu}{\sigma}\right)\right)\mathbb{E}[U_X^2]\right].$$

It is easy to see that U_X and U_Y follow truncated normal distributions. Their moments can be worked out according to formulas given by Jawitz (2004), in particular we have for $\epsilon = \frac{|\mu_X - \mu_Y|}{2\sigma}$

$$\begin{aligned}
\mathbb{E}[(X-Y)^2 \,|\, X \neq Y] &= \frac{4}{Z_U}\left[(\mu^2 + \sigma^2)Z_U + \frac{2\mu}{\sqrt{2\pi}}\exp\left(-\frac{\mu^2}{2\sigma^2}\right)\right] \\
&= 4\sigma^2\left[1 + \epsilon^2 + \frac{2\epsilon}{\sqrt{2\pi}\text{erf}(\epsilon/\sqrt{2})}\right] \\
&\leq 4\sigma^2\left[1 + \epsilon^2 + \sqrt{\frac{2}{\pi}}\left(\epsilon + \sqrt{\frac{\pi}{2}}\right)\right] \\
&= 4\sigma^2\left[2 + \epsilon\sqrt{\frac{2}{\pi}} + \epsilon^2\right]
\end{aligned}$$

where the inequality in the third line comes from the inequality $\frac{x}{\text{erf}(x/\sqrt{2})} \leq x + \sqrt{\frac{\pi}{2}}$. □

Lemma 12. *The following holds.*

(a) $P_{GG}^{(v)}$, $P_{GG}^{(h)}$, $P_{GN}^{(v)}$ *are* $\|W\|_F^2$*-contractive.*

(b) $P_{GN}^{(h)}$ *is* $\frac{5}{4}\|W\|_F^2$*-contractive.*

(c) $P_{GG}^{(v)}$ *and* $P_{GN}^{(v)}$ *are* $(\epsilon_0, \delta_0, M)$*-gamble admissible for* $\epsilon_0 = \frac{1}{4\|(W/\sigma)^T\|_{2,1}^2}$, $\delta_0 = 1/4$, *and*

$$M = 4\|\sigma\|_2^2 + \sqrt{\frac{2}{\pi}}\frac{\left\|(W\sigma)^T\right\|_{2,1}}{\|(W/\sigma)^T\|_{2,1}} + \left(\frac{\|W\|_F}{2\|(W/\sigma)^T\|_{2,1}}\right)^2$$

where W/σ *and* $W\sigma$ *denote* $n \times m$ *matrices whose entries are* W_{ij}/σ_i *and* $W_{ij}\sigma_i$*, respectively*

Proof. To prove part (a), we need only show the result for $P_{GG}^{(v)}$; the bounds for $P_{GG}^{(h)}$ and $P_{GN}^{(v)}$ will follow symmetrically.

Recall that our distance is ℓ_2^2-distance $d_v(x,y) := \sum_{i=1}^n (x(v_i) - y(v_i))^2$. To see that $P_{GG}^{(v)}$ is contractive, let $x, y \in \Omega$ be given. We will construct our contractive coupling (X, Y) by coupling each visible node $X(v_i)$ independently. In particular, we will use the coupling from Lemma 20(a) to couple together the marginal distributions $\mathcal{N}(a_i + \sum_{j=1}^m W_{ij}x(h_j),\, \sigma_i^2)$ and $\mathcal{N}(a_i + \sum_{j=1}^m W_{ij}y(h_j),\, \sigma_i^2)$. By Lemma 20(a), we have

$$\begin{aligned}
\mathbb{E}[d_v(X,Y)] &= \sum_{i=1}^n \mathbb{E}\left[(X(v_i) - Y(v_i))^2\right] \\
&\leq \sum_{i=1}^n \left(\sum_{j=1}^m W_{ij}\left(x(h_j) - y(h_j)\right)\right)^2 \\
&\leq \left(\sum_{i=1}^n \sum_{j=1}^m W_{ij}^2\right) \sum_{j=1}^m (x(h_j) - y(h_j))^2 \\
&= \|W\|_F^2\, d_h\left(x, y\right).
\end{aligned}$$

To prove part (b), we will couple each unit h_j independently as follows.

1. Let (Z_j, Z_j') be the coupling from Lemma 20(a) of $\mathcal{N}(\sum_{i=1}^n W_{ij}x(v_i), \sigma(\sum_{i=1}^n W_{ij}x(v_i)))$ and $\mathcal{N}(\sum_{i=1}^n W_{ij}y(v_i), \sigma(\sum_{i=1}^n W_{ij}y(v_i)))$.
2. Let $X(h_j) = \max(0, Z_j)$ and $Y(h_j) = \max(0, Z_j')$.

Then by the definition of NReLU, X and Y have the correct marginal distributions. To see that they are contractive, note first that for each h_j,

$$\mathbb{E}\left[(X(h_j) - Y(h_j))^2\right] \leq \mathbb{E}\left[\left(Z_j - Z_j'\right)^2\right] = \left(\sum_{i=1}^n W_{ij}\left(x(v_i) - y(v_i)\right)\right)^2 + \left(\sigma\left(\sum_{i=1}^n W_{ij}x(v_i)\right) - \sigma\left(\sum_{i=1}^n W_{ij}y(v_i)\right)\right)^2$$

where the second equality comes from Lemma 20(a). Thus,

$$\begin{aligned}
\mathbb{E}[d_h(X,Y)] &\leq \sum_{j=1}^m \left(\sum_{i=1}^n W_{ij}\left(x(v_i) - y(v_i)\right)\right)^2 + \sum_{j=1}^m \left(\sigma\left(\sum_{i=1}^n W_{ij}x(v_i)\right) - \sigma\left(\sum_{i=1}^n W_{ij}y(v_i)\right)\right)^2 \\
&\leq \|W\|_F^2\, d_v\left(x, y\right) + \sum_{j=1}^m \left(\frac{1}{2}\sum_{i=1}^n W_{ij}(x(v_i) - y(v_i))\right)^2 \\
&\leq \|W\|_F^2\, d_v\left(x, y\right) + \frac{1}{4}\|W\|_F^2\, d_v\left(x, y\right) \\
&= \frac{5}{4}\|W\|_F^2\, d_v\left(x, y\right).
\end{aligned}$$

To prove part (c), it will suffice to prove that $P_{GG}^{(v)}$ is gamble admissible; the gamble admissibility of $P_{GN}^{(v)}$ will follow by symmetry. To do this, we will construct a gamble coupling (X, Y) by independently coupling the visible nodes v_i according to the coupling from Lemma 20(b). The probability that we set $X(v) \neq Y(v)$ is bounded as

$$\begin{aligned} Pr(X(v) \neq Y(v)) &= Pr\left(\exists v_i \text{ s.t. } X(v_i) \neq Y(v_i)\right) \\ &\leq \sum_{i=1}^{n} Pr(X(v_i) \neq Y(v_i)) \\ &\leq \sum_{i=1}^{n} \frac{\left|\sum_{j=1}^{m} W_{ij}\left(x(h_j) - y(h_j)\right)\right|}{2\sigma_i} \\ &\leq \frac{1}{2}\sum_{i=1}^{n} \sqrt{\sum_{j=1}^{m}\left(\frac{W_{ij}}{\sigma_i}\right)^2}\sqrt{\sum_{j=1}^{m}\left(x(h_j) - y(h_j)\right)^2} \\ &= \frac{\left\|(W/\sigma)^T\right\|_{2,1}}{2}\sqrt{d_h\left(x, y\right)} \end{aligned}$$

Similarly we can bound the expected visible distance given $X(v) \neq Y(v)$ as

$$\begin{aligned} \mathbb{E}[d_v(X,Y) \,|\, X(v) \neq Y(v)] &\leq \sum_{i=1}^{n} \mathbb{E}\left[\left(X(v_i) - Y(v_i)\right)^2 \,|\, X(v_i) \neq Y(v_i)\right] \\ &\leq \sum_{i=1}^{n} 4\sigma_i^2\left[1 + \frac{\left|\sum_{j=1}^{m} W_{ij}\left(x(h_j) - y(h_j)\right)\right|}{\sqrt{2\pi}\sigma_i} + \left(\frac{\left|\sum_{j=1}^{m} W_{ij}\left(x(h_j) - y(h_j)\right)\right|}{2\sigma_i}\right)^2\right] \\ &= \sum_{i=1}^{n} 4\sigma_i^2 + 2\sqrt{\frac{2}{\pi}}\sum_{i=1}^{n}\left|\sum_{j=1}^{m}\sigma_i W_{ij}\left(x(h_j) - y(h_j)\right)\right| + \sum_{i=1}^{n}\left(\sum_{j=1}^{m} W_{ij}\left(x(h_j) - y(h_j)\right)\right)^2 \\ &\leq \sum_{i=1}^{n} 4\sigma_i^2 + 2\sqrt{\frac{2}{\pi}}\left\|(W\sigma)^T\right\|_{2,1}\sqrt{d_h\left(x, y\right)} + \|W\|_F^2\, d_h\left(x, y\right). \end{aligned}$$

Plugging in $d_h(x, y) \leq \epsilon_0 = \frac{1}{4\|(W/\sigma)^T\|_{2,1}^2}$ finishes the proof.

□

Proofs from Section 5

Theorem 15. *Pick any $T > 0$ and $n, m \in \mathbb{N}$ even positive integers. Then there is a weight matrix $W \in \mathbb{R}^{n\times m}$ satisfying $\|W\|_{max} \leq \frac{2}{\min(n,m)}\ln\left(4T(n+m)\right)$ such that the Gibbs sampler over the RBM with zero bias and weight matrix W has mixing rate bounded as $\tau_{mix} \geq T$.*

Proof. Let $r = \frac{2}{\min(n,m)}\ln\left(4T(n+m)\right)$. Choose a canonical configuration x such that exactly half of the $x(v_i)$'s are 1 and exactly half of the $x(h_j)$'s are 1. Now let $W \in \mathbb{R}^{n\times m}$ such that $W_{ij} = r$ if $x(v_i) = x(h_j)$ and $-r$ otherwise. Let $\pi(\cdot)$ denote the Gibbs distribution for the RBM with weight matrix W and zero bias and let $S = \{x\}$ be the singleton set containing only the canonical configuration. Note that if $\bar{x}$ satisfies that $\bar{x}(v_i) = 1$ iff $x(v_i) = 0$ and $\bar{x}(h_j) = 1$ iff $x(h_j) = 0$, then $\pi(x) = \pi(\bar{x})$. Thus, $\pi(S) \leq 1/2$.

It is not hard to see $Pr(X(h_j) \neq x(h_j) \,|\, x(v)) = \sigma\left(-\frac{nr}{2}\right)$ for all $j \in [m]$, where $\sigma(x) = 1/(1+\exp(-x))$ is the logistic sigmoid as before. Similarly, for any $i \in [n]$, $Pr(X(v_i) \neq x(v_i) \,|\, x(h)) = \sigma\left(-\frac{mr}{2}\right)$. Thus,

$$Pr(\text{leave state } x) \leq \frac{m}{1+\exp\left(\frac{nr}{2}\right)} + \frac{n}{1+\exp\left(\frac{mr}{2}\right)} \leq \frac{1}{4T}$$

Thus the conductance of S (and therefore Φ^*) is upper bounded as

$$\Phi(S) = \frac{1}{\pi(S)} \sum_{x \in S, y \in S^c} \pi(x) Pr(\text{we transition from } x \text{ to } y) = Pr(\text{leave state } x) \leq \frac{1}{4T}$$

Theorem 14 completes the proof. □

Lemma 21. $\Phi(x) \leq 1 - \sqrt{1 - \exp\left(-\frac{x^2}{2}\right)}$ *for* $x \leq 0$.

Proof. We begin by writing $\Phi(\cdot)$ in terms of the error function:

$$\Phi(x) = \frac{1}{2}\left(1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right).$$

Thus it suffices to prove

$$\text{erf}\,(x)^2 \geq 1 - e^{-x^2}.$$

By calculus, we have

$$\begin{aligned}
\text{erf}(x)^2 &= \frac{4}{\pi}\int_0^x \int_0^x e^{-(s^2+t^2)}\, ds\, dt \\
&\geq \frac{4}{\pi}\int_0^{\pi/2} \int_0^x r e^{-r^2}\, dr\, d\theta \\
&= \frac{4}{\pi}\int_0^{\pi/2} \left[-\frac{1}{2}e^{-r^2}\Big|_{r=0}^{x}\right] d\theta \\
&= \frac{4}{\pi}\int_0^{\pi/2} \frac{1}{2}\left(1 - e^{-x^2}\right) d\theta \\
&= 1 - e^{-x^2}
\end{aligned}$$

where the inequality comes from the fact that $e^{-(s^2+t^2)} \geq 0$ and the quarter circle of radius x centered at the origin and lying in the first quadrant is a subset of the square $[0, x]^2$. □

Theorem 16. *Let* $T, B > 0$ *and* $n, m \in \mathbb{N}$ *be even positive integers. Then there exists weight matrix* $W \in \mathbb{R}^{n \times m}$ *s.t.*

$$\|W\|_{max} \leq \frac{1}{\min(n,m)}\left(1 + \frac{1}{B}\sqrt{8\log(4T\max(n,m))}\right)$$

such that the B*-truncated chain of the Gibbs sampler for the Gaussian-Gaussian RBM with no biases and unit variances mixes in time* $\tau_{mix} \geq T$.

Proof. Let $r = \frac{1}{\min(n,m)}\left(1 + \frac{1}{B}\sqrt{8\log(4T\max(n,m))}\right)$. Let $\mathcal{I}_-, \mathcal{I}_+$ be an even partition of $[n]$, i.e. $|\mathcal{I}_-| = n/2 = |\mathcal{I}_+|$. Similarly, let $\mathcal{J}_-, \mathcal{J}_+$ be an even partition of $[m]$. Define

$$W_{ij} = \begin{cases} r & \text{if } (i,j) \in \mathcal{I}_- \times \mathcal{J}_- \cup \mathcal{I}_+ \times \mathcal{J}_+ \\ -r & \text{else} \end{cases}$$

$$S_v = \{x(v) \in [-B, B]^n \,:\, x(v_i) \geq B/2 \text{ if } i \in \mathcal{I}_+ \text{ and } x(v_i) \leq -B/2 \text{ else}\}$$

$$S_h = \{x(h) \in [-B, B]^m \,:\, x(h_j) \geq B/2 \text{ if } j \in \mathcal{J}_+ \text{ and } x(h_j) \leq -B/2 \text{ else}\}$$

Then our low conductance set of configurations is $S = S_v \times S_h$. Note that the c.d.f.'s of the conditional distributions for the B-thresholded chain are exactly the same as the regular normal distribution for points within $[-B, B]$. That is, given $x \in \Omega$ and $p \in (-B, B)$, for any hidden node h_j and visible node v_i

$$P(X(h_j) < p \,|\, x(v)) = \Phi\left(p - \sum_{i=1}^{n} W_{ij}\, x(v_i)\right) \quad \text{and} \quad P(X(v_i) < p \,|\, x(h)) = \Phi\left(p - \sum_{j=1}^{m} W_{ij}\, x(h_j)\right).$$

For $x \in S$ and $j \in \mathcal{J}_+$, we have by Lemma 21,

$$\begin{aligned} P(X(h_j) < B/2 \,|\, x(v)) &= \Phi\left(\frac{B}{2} - r\left(\sum_{i \in \mathcal{I}+} x(v_i) - \sum_{i \in \mathcal{I}_-} x(v_i)\right)\right) \\ &\leq \Phi\left(\frac{B}{2}(1 - rn)\right) \\ &\leq 1 - \sqrt{1 - \exp\left(-\frac{B^2}{8}(1 - rn)^2\right)}. \end{aligned}$$

Symmetric inequalities also hold for $P(X(h_j) > -B/2 \,|\, x(v))$ when $j \in \mathcal{J}_-$. Additionally, for $i \in \mathcal{I}_+$ and $i' \in \mathcal{I}_-$,

$$P(X(v_i) < B/2 \,|\, x(h)), P(X(v_{i'}) > -B/2 \,|\, x(h)) \leq 1 - \sqrt{1 - \exp\left(-\frac{B^2}{8}(1 - rm)^2\right)}.$$

Therefore, given that the current state of our chain Y_t is in S, we can bound the probability that we transition out of S in the next step as

$$P(Y_{t+1} \notin S \,|\, Y_t \in S) \leq m\left(1 - \sqrt{1 - \exp\left(-\frac{B^2}{8}(1 - rn)^2\right)}\right) + n\left(1 - \sqrt{1 - \exp\left(-\frac{B^2}{8}(1 - rm)^2\right)}\right).$$

Plugging in our value for r gives us an upperbound of $\frac{1}{4T}$. Theorem 14 completes the proof. □

Proofs from Section 6

The works of Jerrum and Sinclair (1993), Long and Servedio (2010), and Goldberg and Jerrum (2007) technically deal with Ising (or spin glass) models as opposed to Boltzmann machines. As the following lemma demonstrates, however, the partition functions of these models differs only by an easily computable constant. Thus, they are approximation-preserving interreducible in the sense of Dyer et al. (Dyer et al., 2004).

Lemma 22. *Let $G = (V, E)$ be a graph, $W_{ij} \in \mathbb{R}$ for all $(i, j) \in E$, $b_i \in \mathbb{R}$ for all $i \in V$, and define*

$$Z_{Ising}(G, W, b) = \sum_{x\,:\,V \to \{-1,1\}^V} \exp\left(\sum_{(i,j) \in E} W_{ij} x(i) x(j) + \sum_{i \in V} b_i x(i)\right)$$

as the Ising partition function and

$$Z_{Boltzmann}(G, W, b) = \sum_{x\,:\,V \to \{0,1\}^V} \exp\left(\sum_{(i,j) \in E} W_{ij} x(i) x(j) + \sum_{i \in V} b_i x(i)\right)$$

as the Boltzmann partition function then $C Z_{Ising}(G, W, b) = Z_{Boltzmann}(G, W', b')$ where

$$W' = 4W \qquad \text{and} \qquad b'_i = 2b_i - 2 \sum_{j \text{ s.t.} (i,j) \in E} W_{ij} \qquad \text{and} \qquad C = \exp\left(\sum_{i \in V} b_i - \sum_{(i,j) \in E} W_{ij}\right).$$

Proof. The key idea is to identify every Ising configuration $x : V \to \{-1, 1\}^V$ with a Boltzmann configurations $y : V \to \{0, 1\}^V$. The convention we will take is $y(i) = \frac{1}{2}(x(i) + 1)$, which has the effect of identifying the spin -1 with 0

and 1 with 1. Then for any Ising/Boltzmann corresponding pair x, y, we have

$$\begin{aligned}
\exp\left(\sum_{(i,j)\in E} W'_{ij}y(i)y(j) + \sum_{i\in V} b'_i y(i)\right) &= \exp\left(\sum_{(i,j)\in E} 4W_{ij}y(i)y(j) + \sum_{i\in V} y(i)\left(2b_i - 2\sum_{j \text{ s.t.}(i,j)\in E} W_{ij}\right)\right) \\
&= \exp\left(\sum_{(i,j)\in E} W_{ij}(x(i)+1)(x(j)+1) + \sum_{i\in V}(x(i)+1)\left(b_i - \sum_{j \text{ s.t.}(i,j)\in E} W_{ij}\right)\right) \\
&= \exp\left(\sum_{(i,j)\in E} W_{ij}(x(i)+x(j)+1) - \sum_{i\in V}(x(i)+1)\left(\sum_{j \text{ s.t.}(i,j)\in E} W_{ij}\right) + \sum_{i\in V} b_i\right) \cdot \\
&\quad \exp\left(\sum_{(i,j)\in E} W_{ij}x(i)x(j) + \sum_{i\in V} b_i x(i)\right) \\
&= C\exp\left(\sum_{(i,j)\in E} W_{ij}x(i)x(j) + \sum_{i\in V} b_i x(i)\right).
\end{aligned}$$

Because the mapping from Ising to Boltzmann configurations is bijective, it then holds that

$$\begin{aligned}
Z_{\text{Boltzmann}}(G, W', b') &= \sum_{y\,:\,V\to\{0,1\}^V} \exp\left(\sum_{(i,j)\in E} W'_{ij}y(i)y(j) + \sum_{i\in V} b'_i y(i)\right) \\
&= \sum_{x\,:\,V\to\{-1,1\}^V} C\exp\left(\sum_{(i,j)\in E} W_{ij}x(i)x(j) + \sum_{i\in V} b_i x(i)\right) \\
&= C Z_{\text{Ising}}(G, W, b).
\end{aligned}$$

□