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

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

Taking $t \geq 1 + \frac{1}{\log(c_1 c_2)} \log\left(\frac{\min(\gamma_h^{(\text{max})}, c_2\gamma_v^{(\text{max})})}{\gamma_v^{(\text{min})}}\right)$ 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 $P^{(v)}(X(v_i) \mid x(h))$ and $P^{(v)}(Y(v_i) \mid y(h))$ guaranteed in Lemma 1. By doing this independently for all visible nodes, we have a valid coupling (X, Y) of $P^{(v)}(\cdot \mid x(h))$ and $P^{(v)}(\cdot \mid y(h))$. Then we can work out the expected Hamming distance of X and Y as

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

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

□

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*

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

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

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 $P^{(h)}(\cdot \mid X_s(v))$ and $P^{(h)}(\cdot \mid Y_s(v))$.
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 $P^{(v)}(\cdot \mid X_{s+1}(h))$ and $P^{(v)}(\cdot \mid Y_{s+1}(h))$. Otherwise, draw $(X_{s+1}(v), Y_{s+1}(v))$ according to the c_2-contractive coupling of $P^{(v)}(\cdot \mid X_{s+1}(h))$ and $P^{(v)}(\cdot \mid Y_{s+1}(h))$.

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

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

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

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

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

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

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 $Y = U$.

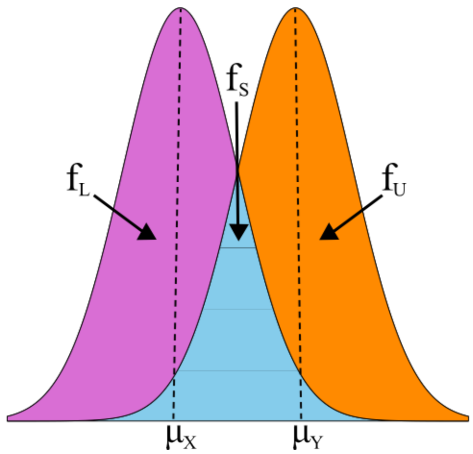

Figure 3. Illustration of the unnormalized densities f_S, f_U, f_L.

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

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

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

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

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 $Pr(X(h_j) \neq x(h_j) \mid x(v)) = \sigma\left(-\frac{nr}{2}\right)$ for all $j \in [m]$, where $\sigma(x) = 1/(1+\exp(-x))$ is the logistic sigmoid as before. Similarly, for any $i \in [n]$, $Pr(X(v_i) \neq x(v_i) \mid x(h)) = \sigma\left(-\frac{mr}{2}\right)$. 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. $\square$

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 $[0,x]^2$. $\square$

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,

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

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 $i' \in \mathcal{I}_-$,

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

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

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

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*

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

as the Ising partition function and

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

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

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

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

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

□