Supplementary material

In the following sections we provide additional material (proofs and figures) that supplement our main results. Section A outlines the preliminary facts and notations that we use for the proofs. The subsequent sections provide the detailed proofs for respective lemmas and theorems. Figure 5 compares the theoretical upper bound estimate with the actual simulated values for modes of two layer DBMs ($\mathcal{C}(n, m_1, m_2)$).

A. Preliminary Facts and Notations

In the proofs that follow we use the following facts and notations:

1. The probability density function (pdf) of standard normal distribution $\mathcal{N}(0,1)$

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

2. The cumulative distribution function (cdf) of standard normal distribution

$$\Phi(x) = \int_{-\infty}^{x} \phi(x)dx = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right] \text{ where } \text{erf}(x) = \frac{1}{\sqrt{\pi}}\int_{-x}^{x} e^{-t^2}dt$$

3. The pdf of a skew normal distribution $\hat{\mathcal{N}}$ with skew parameter α

$$f(x) = 2\phi(x)\Phi(\alpha x)$$

4. If $X \sim \mathcal{N}(\mu, \sigma^2)$, $a \in \mathbb{R}, \alpha = \frac{a-\mu}{\sigma}$, then X conditioned on $X > a$ follows a truncated normal distribution with moments

$$\begin{aligned}\mathbb{E}\left[X|X > a\right] &= \mu + \sigma\frac{\phi(\alpha)}{Z} \\ Var(X|X > a) &= \sigma^2\left[1 + \alpha\frac{\phi(\alpha)}{Z} - \left(\frac{\phi(\alpha)}{Z}\right)^2\right]\end{aligned}$$

where $Z = 1 - \Phi(\alpha)$.

5. *Squeeze Theorem*[8]: Let, $\{a_m\}, \{b_m\}, \{c_m\}$ be sequences such that $\forall m \geq m_0$ ($m_0 \in \mathbb{R}$)

$$a_m \leq b_m \leq c_m$$

Further, let $\lim_{m\to\infty} a_m = \lim_{m\to\infty} c_m = L$, then

$$\lim_{m\to\infty} b_m = L$$

B. Proof of Lemma 1 (See page 4)

Lemma 1. *A vector* $\mathbf{v}$ *is perfectly reconstructible for an* $RBM_{n,m}(\theta) \iff$ *the state* $\{\mathbf{v}, \text{up}(\mathbf{v})\}$ *is one-flip stable.*

Proof. Let $\mathbf{h}^* = \text{up}(\mathbf{v})$ (conditioning on θ is implicit). If $\mathbf{v}$ is perfectly reconstructible $\implies \mathbf{v} = \arg\max_{\mathbf{v}} P(\mathbf{v}|\mathbf{h}^*) \implies \forall \mathbf{v}' \neq \mathbf{v}, P(\mathbf{v}', \mathbf{h}^*) < P(\mathbf{v}, \mathbf{h}^*)$. Similarly since $\mathbf{h}^* = \arg\max_{\mathbf{h}} P(\mathbf{h}|\mathbf{v}), \forall \mathbf{h}' \neq \mathbf{h}^*, P(\mathbf{v}, \mathbf{h}') < P(\mathbf{v}, \mathbf{h}^*)$. Hence the state $\{\mathbf{v}, \mathbf{h}^*\}$ is stable against any number of flips of visible units and against any number of flips of hidden units, $\implies \{\mathbf{v}, \mathbf{h}^*\}$ is atleast one-flip stable.
Conversely let $\{\mathbf{v}^*, \mathbf{h}^*\}$ be one-flip stable. We shall prove by contradiction that $\text{up}(\mathbf{v}^*) = \mathbf{h}^*$ and $\text{down}(\mathbf{h}^*) = \mathbf{v}^*$. Assume $\text{up}(\mathbf{v}^*) = \mathbf{h}' \neq \mathbf{h}^*$. We use the fact that for an RBM the hidden units are conditionally independent of each other given the visible units. Thus $\mathbf{h}' = \arg\max_{\mathbf{h}} P(\mathbf{h}|\mathbf{v}^*) = \{\arg\max_{h_j} P(h_j|\mathbf{v}^*)\}_{j=1}^m$. Further $P(\mathbf{h}^*|\mathbf{v}^*) = \prod_{j=1}^m P(h_j^*|\mathbf{v}^*)$. Let k be an index such that $h_k' \neq h_k^*$. Since $h_k' = \arg\max_{h_k} P(h_k|\mathbf{v}^*), \implies P(h_k'|\mathbf{v}^*) > P(h_k^*|\mathbf{v}^*)$. Moreover, $P(\mathbf{v}^*, \mathbf{h}^*) = P(\mathbf{v}^*)P(\mathbf{h}^*|\mathbf{v}^*) = P(\mathbf{v}^*)\prod_{j=1}^m P(h_j^*|\mathbf{v}^*)$. Thus just by flipping h_k^* to h_k' we can increase the probability of the state $\{\mathbf{v}^*, \mathbf{h}^*\}$. This contradicts the one-flip stability hypothesis. Similarly using the conditional independence of visible units given the hidden units we can show that $\text{down}(\mathbf{h}^*) = \mathbf{v}^*$. □

[8] http://mathonline.wikidot.com/the-squeeze-theorem-for-convergent-sequences

C. Proof of Lemma 2 (See page 5)

Lemma 2. *For the set* $\textbf{RBM}_{n,m}$*, if a given vector* $\mathbf{v}$ *has* $r(\geq 1)$ *ones,* $\mathbf{h} = \text{up}(\mathbf{v})$ *has* l *ones and* $l \gg 1$*,then* [9] *for* $r > 1$*,*

$$\mathbb{E}\left[\mathbb{1}_{[\mathbf{v} \text{ is PR.}]}\right] \leq \left[\frac{1}{2} - \frac{1}{2}\operatorname{erf}\left(-\sqrt{\frac{l}{\pi r - 2}}\right)\right]^{r}\left(\frac{1}{2}\right)^{n-r}.$$

For $r = 1$*, the expression* $\mathbb{E}\left[\mathbb{1}_{[\mathbf{v} \text{ is PR.}]}\right]$ *equates to* $\left(\frac{1}{2}\right)^{n-1}$*. where* $\operatorname{erf}(x) = \frac{1}{\sqrt{\pi}}\int_{-x}^{x} e^{-t^2}dt$

Proof. We first note that given a visible vector $\mathbf{v} \in \{0,1\}^n$ the most likely configuration of the hidden vector

$$\left\{h_j = [\text{up}(\mathbf{v})]_j = \mathbb{1}_{\left[\sum_{i=1}^{n} w_{ij} v_i > 0\right]}\right\}_{j=1}^{m}$$

Likewise given a hidden vector $\mathbf{h}$, the most likely visible vector

$$\left\{v_i = [\text{down}(\mathbf{h})]_i = \mathbb{1}_{\left[\sum_{j=1}^{m} w_{ij} h_j > 0\right]}\right\}_{i=1}^{n}$$

Case 1: $r = 1$
By symmetry it can be assumed $v_1 = 1$, and $v_i = 0 (\forall i > 1)$. Then $\left\{h_j = \mathbb{1}_{[w_{1j} > 0]}\right\}_{j=1}^{m}$. Since each of w_{1j} is i.i.d. as per $\mathcal{N}(0, \sigma^2)$, h_j is a Bernoulli random variable with $P(h_j = 1) = \frac{1}{2}$. Again by symmetry it is assumed the first l units $\{h_j\}_{j=1}^{l}$ are one. Then the most likely *reconstructed* visible vector is given by $\left\{\hat{v}_i = \mathbb{1}_{\left[X_i = \sum_{j=1}^{l} w_{ij} > 0\right]}\right\}_{i=1}^{n}$. Since $w_{1j} > 0$ for all $1 \leq j \leq l \implies \hat{v}_1 = 1$. Also, for all $i > 1, w_{ij} \sim \mathcal{N}(0, \sigma^2) \implies X_i \sim \mathcal{N}(0, l\sigma^2) \implies \{\hat{v}_i\}_{i>1}$ is a Bernoulli random variable with $\left\{P[\hat{v}_i = 1] = \frac{1}{2}\right\}_{i=2}^{n}$. The result then follows by mutual independence of $\hat{v}_i$.

Case 2: $r > 1$
For $r(> 1)$ ones in $\mathbf{v}$ and l ones in $\mathbf{h} = \text{up}(\mathbf{v})$ the problem of computing $\{P[\hat{v}_i = 1]\}_{i=1}^{r}$ can be reformulated in terms of matrix row and column sums, viz, given $W \in \mathbb{R}^{r \times l}$ where all entries $w_{ij} \sim \mathcal{N}(0, \sigma^2)$ are i.i.d. and given that all the column sums $\{C_j = \sum_{i=1}^{r} w_{ij} > 0\}_{j=1}^{l}$, to compute the probability that all the row sums are positive, i.e., $\left\{R_i = \sum_{j=1}^{l} w_{ij} > 0\right\}_{i=1}^{r}$.

Using properties of normal distribution it can be shown that conditioned on the fact that $C_j > 0$, the posterior distribution of w_{ij} shall be *skew-normal* with mean $\mu_{ij} = \sigma\sqrt{\frac{2}{\pi r}}$ and variance $\sigma_{ij}^2 = \sigma^2\left(1 - \frac{2}{\pi r}\right)$. Since the random variables $\{w_{ij}|C_j > 0\}_{j=1}^{l}$ are independent the posterior mean of R_i shall be $\tilde{\mu}_i = l\sigma\sqrt{\frac{2}{\pi r}}$ and the posterior variance $\tilde{\sigma}_i^2 = l\sigma^2\left(1 - \frac{2}{\pi r}\right)$. Since $l \gg 1$ by *Central Limit Theorem* R_i follow a normal distribution. Since the R_i are negatively correlated (proof follows) and $\left\{P[\hat{v}_i = 1] = \frac{1}{2}\right\}_{i>r}$ by similar reasoning as in Case 1 we get our desired upper bound.

Negatively Correlated R_i's: Conditioned on the fact $\{C_j > 0\}_{j=1}^{l}$ the random variables $\{R_i\}_{i=1}^{r}$ are not independent. They are negatively correlated because for all $R_i, R_t (t \neq i)$,

$$P(R_i > 0 | \{C_j > 0\}_{j=1}^{l}, R_t > 0) < P(R_i > 0 | \{C_j > 0\}_{j=1}^{l})$$

Hence the expression given in Lemma 2 is an upper bound since we have neglected the negative correlation among the R_i and in the process over-estimated the probabilities.

□

[9] Here $l \gg 1$ means l is atleast 50 hidden units, which according to us is a reasonable assumption.

D. Proof of Lemma 3 (See page 5)

Lemma 3. *For the set $\boldsymbol{RBM}_{n,m}$, if $\boldsymbol{v}$ has $r(>1)$ ones, $\boldsymbol{h} = \mathrm{up}(\boldsymbol{v})$ has l ones, then $\exists \mu_c, \tilde{\mu}_c, \sigma_c, \tilde{\sigma}_c \in \mathbb{R}_+$ such that conditioned on $\{R_t > 0\}_{t=1}^{i-1}, C_j > 0$, the moments of posterior distribution of w_{ij} is given by*

$$\begin{aligned}\mathbb{E}\left[w_{ij}|\{R_t>0\}_{t=1}^{i-1}, C_j>0\right] &= (\tilde{\mu}_c - \mu_c)\frac{\sigma^2}{\sigma_c^2}\\ \mathit{Var}\left[w_{ij}|\{R_t>0\}_{t=1}^{i-1}, C_j>0\right] &= \tilde{\sigma}_c^2\left(\frac{\sigma^2}{\sigma_c^2}\right)^2 + \sigma^2\beta\end{aligned}$$

where $\beta = \left(1 - \frac{\sigma^2}{\sigma_c^2}\right)$

Proof. The conditional distribution for $R_1 = \sum_{j=1}^{l} w_{1j}$ is obtained from the proof of Lemma 2.

$$\left(R_1|\{C_j>0\}_{j=1}^{l}\right) \sim \mathcal{N}\left(\tilde{\mu}_1, \tilde{\sigma}_1^2\right)$$

where $\tilde{\mu}_1 = l\sigma\sqrt{\frac{2}{\pi r}}$, $\tilde{\sigma}_1^2 = l\sigma^2\left(1 - \frac{2}{\pi r}\right)$. Using similar arguments as in proof of Lemma 2, conditioned on $R_t > 0$ the posterior distribution of w_{tj} shall be skew normal $\hat{\mathcal{N}}\left[\sigma\sqrt{\frac{2}{\pi l}}, \sigma^2\left(1-\frac{2}{\pi l}\right)\right]$. Then conditioned on $\{R_t > 0\}_{t=1}^{i-1}$, C_j shall be distributed as per skew normal

$$\left(C_j|\{R_t>0\}_{t=1}^{i-1}\right) \sim \hat{\mathcal{N}}(\mu_c, \sigma_c^2)$$

where

$$\mu_c = (i-1)\sigma\sqrt{\frac{2}{\pi l}} \text{ and } \sigma_c^2 = (i-1)\sigma^2\left(1 - \frac{2}{\pi l}\right) + (r-i+1)\sigma^2$$

Here we approximate the above distribution to be Normal since if i is large then *Central Limit Theorem* would be applicable, otherwise the normally distributed variables $\{w_{kj}\}_{k=i}^{r}$ would dominate the sum. Then conditioned on $\{R_t > 0\}_{t=1}^{i-1}, C_j > 0$, C_j shall be distributed as per truncated normal distribution (Barr & Sherrill, 1999) with moments

$$\begin{aligned}\mathbb{E}\left[C_j|\{R_t>0\}_{t=1}^{i-1}, C_j>0\right] &= \tilde{\mu}_c = \quad \mu_c + \sigma_c\frac{\phi}{Z}\\ \mathrm{Var}\left[C_j|\{R_t>0\}_{t=1}^{i-1}, C_j>0\right] &= \tilde{\sigma}_c^2 = \quad \sigma_c^2\left[1 - \frac{\mu_c\phi}{\sigma_c Z} - \frac{\phi^2}{Z^2}\right]\end{aligned}$$

where $\sigma_c^2 = (i-1)\sigma^2\left(1-\frac{2}{\pi l}\right) + (r-i+1)\sigma^2$,
$\mu_c = (i-1)\sigma\sqrt{\frac{2}{\pi l}}$, $Z = \frac{1}{2} - \frac{1}{2}\operatorname{erf}\left(-\frac{\mu_c}{\sigma_c\sqrt{2}}\right)$ and $\phi = \frac{1}{\sqrt{2\pi}}e^{\left(-\frac{\mu_c^2}{2\sigma_c^2}\right)}$. Then $\mathbb{E}\left[w_{ij}|\{R_t>0\}_{t=1}^{i-1}, C_j = c\right] = (c-\mu_c)\frac{\sigma^2}{\sigma_c^2}$ and $\mathrm{Var}\left[w_{ij}|\{R_t>0\}_{t=1}^{i-1}, C_j = c\right] = \sigma^2\left(1 - \frac{\sigma^2}{\sigma_c^2}\right)$. The result then follows from Laws of total expectation and total variance respectively. □

Remark. The random variables $\{\tilde{w}_{ij} = w_{ij}|\{R_t>0\}_{t=1}^{i-1}, C_j > 0\}_{j=1}^{l}$ shall be negatively correlated with one another so we should subtract the covariance terms while determining the effective variance of $R_i = \sum_{j=1}^{l}\tilde{w}_{ij}$. Thus if we don't subtract the covariance terms from the variance we would get a lower bound on the posterior probability of R_i being positive. However it is close as can be seen in Figure 3.

E. Proof of Theorem 1 (See page 5)

Theorem 1. *(**ISC of RBM$_{n,m}$**) There exist non-trivial functions $L(n,m), U(n,m) : \mathbb{Z}\times\mathbb{Z} \to \mathbb{R}_+$ such that **ISC** of the set $\boldsymbol{RBM}_{n,m}$ obeys the following inequality.*

$$\frac{1}{n}\log_2(L(n,m)) \le \mathcal{C}(n,m) \le \frac{1}{n}\log_2(U(n,m))$$

Proof. The upper bound follows from Lemma 2 and applying linearity of expectation.

$$U_{n,m} = \sum_{r=1}^{n} \binom{n}{r} \sum_{l=1}^{m} \binom{m}{l} \left(\frac{1}{2}\right)^m \left[\frac{1}{2} - \frac{1}{2}\operatorname{erf}\left(-\sqrt{\frac{l}{\pi r - 2}}\right)\right]^r \left(\frac{1}{2}\right)^{n-r}$$

For lower bound, we use Lemma 3. We have $\mathbb{E}\left[w_{ij} | \{R_t > 0\}_{t=1}^{i-1}, C_j > 0\right] = \tilde{\mu}_i(r,l)$ and $\text{Var}\left[w_{ij} | \{R_t > 0\}_{t=1}^{i-1}, C_j > 0\right] = (\tilde{\sigma}_i(r,l))^2$. Thus posterior mean and variance of $\{R_i\}_{i=1}^{r}$ shall be $l\tilde{\mu}_i(r,l)$ and $l(\tilde{\sigma}_i(r,l))^2$ respectively. Then summing over all possibilities of l and applying linearity of expectation we get the lower bound.

$$L_{n,m} = \sum_{r=1}^{n} \binom{n}{r} \sum_{l=1}^{m} \binom{m}{l} \left(\frac{1}{2}\right)^m \left\{\prod_{i=1}^{r} \left[\frac{1}{2} - \frac{1}{2}\operatorname{erf}\left(-\frac{\tilde{\mu}_i(r,l)\sqrt{\frac{l}{2}}}{\tilde{\sigma}_i(r,l)}\right)\right]\right\} \left(\frac{1}{2}\right)^{n-r}$$

□

F. Proof of Corollary 1 (See page 5)

Corollary 1. (Large m limit) *For the set* ***RBM***$_{n,m}$*,* $\lim_{m\to\infty} \mathcal{C}(n,m) = \log_2 1.5 = 0.585$ *where* $C(n,m)$ *is defined in Theorem 1.*

Proof. We shall show that $\lim_{m\to\infty} U_{n,m} \le 1.5^n$ and $\lim_{m\to\infty} L_{n,m} \ge 1.5^n$. Then using *Squeeze Theorem* and the fact that limits preserve inequalities the result shall hold.

$$\lim_{m\to\infty} U_{n,m} = \lim_{m\to\infty} \left\{\sum_{r=1}^{n} \binom{n}{r} \sum_{l=1}^{m} \binom{m}{l} \left(\frac{1}{2}\right)^m \left[\frac{1}{2} - \frac{1}{2}\operatorname{erf}\left(-\sqrt{\frac{l}{\pi r - 2}}\right)\right]^r \left(\frac{1}{2}\right)^{n-r}\right\}$$

If we replace the l inside the erf function by m then we would be increasing the value of the expression since $m \ge l$. Thus

$$\begin{aligned}
\lim_{m\to\infty} U_{n,m} &\le \lim_{m\to\infty} \left\{\sum_{r=1}^{n} \binom{n}{r} \sum_{l=1}^{m} \binom{m}{l} \left(\frac{1}{2}\right)^m \left(\frac{1}{2}\right)^r \left[1 - \operatorname{erf}\left(-\sqrt{\frac{m}{\pi r - 2}}\right)\right]^r \left(\frac{1}{2}\right)^{n-r}\right\} \\
&= \lim_{m\to\infty} \sum_{r=1}^{n} \binom{n}{r} \sum_{l=1}^{m} \binom{m}{l} \left(\frac{1}{2}\right)^m \left(\frac{1}{2}\right)^r [2]^r \left(\frac{1}{2}\right)^{n-r} \\
&= 1.5^n
\end{aligned}$$

To get a lower bound on $L_{n,m}$ we choose a small fixed constant $\epsilon > 0$. Then

$$\begin{aligned}
\lim_{m\to\infty} L_{n,m} &= \lim_{m\to\infty} \sum_{r=1}^{n} \binom{n}{r} \sum_{l=1}^{m} \binom{m}{l} \left(\frac{1}{2}\right)^m \left\{\prod_{i=1}^{r} \left[\frac{1}{2} - \frac{1}{2}\operatorname{erf}\left(-\frac{\tilde{\mu}_i(r,l)\sqrt{\frac{l}{2}}}{\tilde{\sigma}_i(r,l)}\right)\right]\right\} \left(\frac{1}{2}\right)^{n-r} \\
&\ge \lim_{m\to\infty} \sum_{r=1}^{n} \binom{n}{r} \sum_{l=m\epsilon}^{m} \binom{m}{l} \left(\frac{1}{2}\right)^m \left\{\prod_{i=1}^{r} \left[\frac{1}{2} - \frac{1}{2}\operatorname{erf}\left(-\frac{\tilde{\mu}_i(r,l)\sqrt{\frac{l}{2}}}{\tilde{\sigma}_i(r,l)}\right)\right]\right\} \left(\frac{1}{2}\right)^{n-r} \\
&\ge \lim_{m\to\infty} \sum_{r=1}^{n} \binom{n}{r} \sum_{l=m\epsilon}^{m} \binom{m}{l} \left(\frac{1}{2}\right)^m \left\{\prod_{i=1}^{r} \left[\frac{1}{2} - \frac{1}{2}\operatorname{erf}\left(-\frac{\tilde{\mu}_i(r,l)\sqrt{\frac{m\epsilon}{2}}}{\tilde{\sigma}_i(r,l)}\right)\right]\right\} \left(\frac{1}{2}\right)^{n-r}
\end{aligned}$$

Since $\tilde{\mu}_i(r,l)$ and $\tilde{\sigma}_i(r,l)$ are non-zero finite quantities regardless of the value of l amd m and ϵ is a fixed non-zero constant,

$$\begin{aligned}\lim_{m\to\infty} L_{n,m} &\geq \lim_{m\to\infty}\sum_{r=1}^{n}\binom{n}{r}\sum_{l=m\epsilon}^{m}\binom{m}{l}\left(\frac{1}{2}\right)^m\left\{\prod_{i=1}^{r}\left[\frac{1}{2}-\frac{1}{2}\operatorname{erf}\underbrace{\left(-\frac{\tilde{\mu}_i(r,l)\sqrt{\frac{m\epsilon}{2}}}{\tilde{\sigma}_i(r,l)}\right)}_{\to-\infty}\right]\right\}\left(\frac{1}{2}\right)^{n-r}\\ &= \lim_{m\to\infty}\sum_{r=1}^{n}\binom{n}{r}\underbrace{\sum_{l=m\epsilon}^{m}\binom{m}{l}\left(\frac{1}{2}\right)^m}_{\text{Prob}(l>m\epsilon)}\left\{\left(\frac{1}{2}\right)^r[2]^r\right\}\left(\frac{1}{2}\right)^{n-r}\end{aligned}$$

Since ϵ is an arbitrarily small number that we have chosen and l denotes the number of successes in m Bernoulli trials, $\text{Prob}(l > m\epsilon) = 1$.

$$\implies \lim_{m\to\infty} L_{n,m} \geq 1.5^n$$

$$\implies 1.5^n \leq \lim_{m\to\infty} L_{n,m} \leq \lim_{m\to\infty}\mathcal{C}(n,m) \leq \lim_{m\to\infty} U_{n,m} \leq 1.5^n$$

□

G. Proof of Theorem 2 (See page 6)

Theorem 2. **(ISC of RBM$_{n,m_1,m_2}$)** *For an **RBM**$_{n,m_1,m_2}$ ($n, m_1 > 0$ and $m_2 \geq 0$), if we denote $u = \max(m_1, n + m_2), l = \min(m_1, n + m_2)$, then*

$$\mathcal{C}(n,m_1,m_2) \leq \frac{1}{n}\log_2 S$$

whenever $S < \gamma 2^n,\ S = \left[1 - \frac{1}{2}\operatorname{erf}\left(-\sqrt{\frac{u}{\pi l - 4}}\right)\right]^l$

Proof. As shown in Figure 2 we construct a single layer RBM$_{n+m_2,m_1}$ that has the same bipartite connections as **RBM**$_{n,m_1,m_2}$. The expected number of perfectly reconstructible vectors for the single layer RBM can then be obtained from Equation 10.

$$\begin{aligned}\mathcal{C}(n+m_2,m_1) &\leq \frac{1}{n}\log_2 U_{n+m_2,m_1} = \frac{1}{n}\log_2 S\\ &= \frac{1}{n}\log_2\left[1-\frac{1}{2}\operatorname{erf}\left(-\sqrt{\frac{u}{\pi l-4}}\right)\right]^l\end{aligned}$$

However this quantity is an overestimate. This counts the number of pairs of vectors $\{\mathbf{v}, \mathbf{h}_2\}$ such that $\begin{pmatrix}\mathbf{v}\\ \mathbf{h}_2\end{pmatrix}$ is perfectly reconstructible for RBM$_{n+m_2,m_1}$. Among these, there can be vectors like $\begin{pmatrix}\mathbf{v}^{(1)}\\ \mathbf{h}_2^{(1)}\end{pmatrix}$ and $\begin{pmatrix}\mathbf{v}^{(2)}\\ \mathbf{h}_2^{(2)}\end{pmatrix}$ where $\mathbf{v}^{(1)} = \mathbf{v}^{(2)}$ resulting in repetitions. Assuming such vectors $\mathbf{v}^{(i)}$ are uniformly distributed among the 2^n possibilities, we approximate the problem to the following. Given 2^n distinct vectors, we make S draws from them uniformly randomly with replacement. The expected number of distinct vectors that result is given by $2^n\left[1 - \left(1 - \frac{1}{2^n}\right)^S\right]$. If $S < \gamma 2^n$ then binomial approximation an be applied and we get the desired result. □

H. Proof of Corollary 2 (See page 6)

Corollary 2. *(***Layer 1 Wide, Layer 2 Narrow***)* *For an **RBM**$_{n,m_1,m_2}$ ($n, m_1 > 0$ and $m_2 \geq 0$), if $\alpha_1 = \frac{m_1}{n} > \frac{1}{\gamma}$ and $\alpha_2 = \frac{m_2}{n} < \gamma$ then*

$$\mathcal{C}(n,m_1,m_2) \leq (1+\alpha_2)\log_2(1.5)$$

Proof. For $\alpha_1 > \frac{1}{\gamma}$, $S = \left[1 - \frac{1}{2}\operatorname{erf}\left(-\sqrt{\frac{n\alpha_1}{\pi n(1+\alpha_2)-4}}\right)\right]^{n(1+\alpha_2)} = 1.5^{n(1+\alpha_2)}$.

Moreover for $\alpha_2 < \gamma$, since $S = 1.5^{n(1+\alpha_2)} < 1.5^{n(1+\gamma)} = 2^{n(1+\gamma)\log_2(1.5)} = 2^{0.614n}(< \gamma 2^n$ for reasonable choices of n), we can apply binomial approximation and the result follows. □

I. Proof of Corollary 3 (See page 6)

Corollary 3. (**Fixed budget on parameters**) *For an* $\textbf{RBM}_{n,m_1,m_2}$ *($n, m_1 > 0$ and $m_2 \geq 0$), if there is a budget of cn^2 on the total number of parameters, i.e, $\alpha_1(1+\alpha_2) = c$ then the maximum possible* ***ISC****,* $\max_{\alpha_1,\alpha_2}\mathcal{C}(n,\alpha_1,\alpha_2) \leq \tilde{U}(n,\alpha_1^*,\alpha_2^*)$ *where*

$$\tilde{U}(n,\alpha_1^*,\alpha_2^*) = \begin{cases} \min(1, \sqrt{c}\log_2(1.29)) & \text{if } c \geq 1 \\ c\log_2\left[1 - \frac{1}{2}\operatorname{erf}\left(-\sqrt{\frac{1}{\pi c}}\right)\right] & \text{if } c < 1 \end{cases}$$

Proof. We consider two regimes.

Regime 1 ($\alpha_1 \leq 1 + \alpha_2$)

In this regime using Theorem 2, $\mathcal{C}(n,m_1,m_2) \leq \frac{1}{n}\log_2 S$ where

$$S = \left[1 - \frac{1}{2}\operatorname{erf}\left(-\sqrt{\frac{u}{\pi l - \underbrace{4}_{=o(1)}}}\right)\right]^l = \left[1 - \frac{1}{2}\operatorname{erf}\left(-\sqrt{\frac{\frac{nc}{\alpha_1}}{\pi n\alpha_1}}\right)\right]^{n\alpha_1} = \left[1 - \frac{1}{2}\operatorname{erf}\left(-\sqrt{\frac{c}{\pi\alpha_1^2}}\right)\right]^{n\alpha_1}$$

We will prove that $\frac{\partial S}{\partial\alpha_1} > 0$. Taking natural logarithm on both sides,

$$\ln S = n\alpha_1 \ln\left[1 - \frac{1}{2}\operatorname{erf}\left(-\sqrt{\frac{c}{\pi\alpha_1^2}}\right)\right]$$

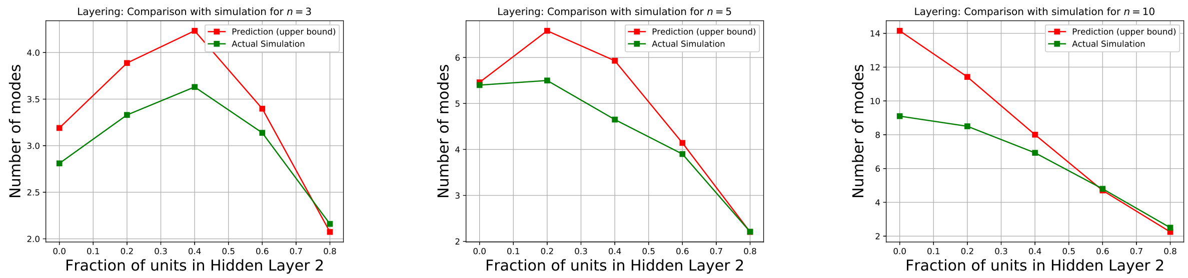
$$\begin{aligned} \frac{1}{S}\frac{\partial S}{\partial\alpha_1} &= n\ln\left[1 - \frac{1}{2}\operatorname{erf}\left(-\sqrt{\frac{c}{\pi\alpha_1^2}}\right)\right] + \frac{n\alpha_1}{1 - \frac{1}{2}\operatorname{erf}\left(-\sqrt{\frac{c}{\pi\alpha_1^2}}\right)}\left[-\frac{1}{\sqrt{(\pi)}}\exp\left(-\frac{c}{\pi\alpha_1^2}\right)\right]\left(\frac{1}{\alpha_1^2}\sqrt{\frac{c}{\pi}}\right) \\ &= n\ln\left[1 - \frac{1}{2}\operatorname{erf}\left(-\sqrt{\frac{c}{\pi\alpha_1^2}}\right)\right] - \frac{n\alpha_1}{1 - \frac{1}{2}\operatorname{erf}\left(-\sqrt{\frac{c}{\pi\alpha_1^2}}\right)}\left[\frac{1}{\sqrt{(\pi)}}\exp\left(-\frac{c}{\pi\alpha_1^2}\right)\right]\left(\frac{1}{\alpha_1^2}\sqrt{\frac{c}{\pi}}\right) \end{aligned}$$

Now since $c = \alpha_1(1+\alpha_2)$ and we are in the regime $\alpha_1 \leq 1 + \alpha_2$, $\implies \frac{c}{\alpha_1^2} \geq 1$. Hence

$$\begin{aligned} \frac{1}{S}\frac{\partial S}{\partial\alpha_1} &\geq n\ln\left[1 - \frac{1}{2}\operatorname{erf}\left(-\sqrt{\frac{1}{\pi}}\right)\right] - \frac{\frac{n}{\sqrt{\pi}}}{1 - \frac{1}{2}\operatorname{erf}\left(-\sqrt{\frac{1}{\pi}}\right)}\underbrace{\left[\left(\sqrt{\frac{c}{\pi\alpha_1^2}}\right)\exp\left(-\frac{c}{\pi\alpha_1^2}\right)\right]}_{x\exp(-x^2)\leq 0.428} \\ &= 0.252n - 0.187n \\ \implies \frac{\partial S}{\partial\alpha_1} &> 0 \end{aligned}$$

Similarly we can show that in the **Regime** $\alpha_1 > 1 + \alpha_2$, $\frac{\partial S}{\partial\alpha_2} > 0$ which would imply $\frac{\partial S}{\partial\alpha_1} < 0$.

Hence the maximum occurs when either $\alpha_1 = 1 + \alpha_2 = \sqrt{c}\ (c \geq 1)$ or $\alpha_1 = c\ (c < 1)$. □



(a) $2^{n\mathcal{C}(n,m_1,m_2)}$ for $n = 3$ (b) $2^{n\mathcal{C}(n,m_1,m_2)}$ for $n = 5$ (c) $2^{n\mathcal{C}(n,m_1,m_2)}$ for $n = 10$

Figure 5. Comparison chart of the upper estimates with the actual simulation value for two layered RBM with $m_1 + m_2 = 10$. The values are plotted for various values of $\beta = \frac{m_2}{m_1}$.

J. Relationship between modes of joint and marginal distribution

Proposition

Let $\{\mathbf{v}_r\}_{r=1}^k$ be visible vectors such that for each pair of vectors $\{\mathbf{v}_i, \mathbf{v}_j\}$ in $\{\mathbf{v}_r\}_{r=1}^k$, $d_H(\mathbf{v}_i, \mathbf{v}_j) \geq 2$. For an RBM$_{n,m_1,\ldots,m_L}(\theta)$ that fits the input distribution $p(\mathbf{v}) = \frac{1}{k}\sum_{i=1}^k \delta(\mathbf{v} - \mathbf{v}_i)$, if a vector $\mathbf{v}$ is a mode of marginal distribution, then there exist vectors $\{\mathbf{h}_l^*\}_{l=1}^L$ such that $(\mathbf{v}, \{\mathbf{h}_l^*\}_{l=1}^L)$ is a mode of joint distribution $p(\mathbf{v}, \{\mathbf{h}_l\}_{l=1}^L)$.

Proof. Since v is a mode, $\implies p(\mathbf{v}) = \frac{1}{k} > 0$.
Further, let $\{\mathbf{h}_l^*\}_{l=1}^L = \arg\max_{\{\mathbf{h}_l\}} P(\mathbf{v}, \{\mathbf{h}_l\}_{l=1}^L)$, that is, the state $(\mathbf{v}, \{\mathbf{h}_l^*\}_{l=1}^L)$ is stable against flip of any **hidden** unit[10]. Moreover, since for all neighbours $\mathbf{v}'$ of $\mathbf{v}$, $p(\mathbf{v}') = 0 \implies p(\mathbf{v}', \{\mathbf{h}_l^*\}_{l=1}^L) = 0$, it implies that $(\mathbf{v}, \{\mathbf{h}_l^*\}_{l=1}^L)$ is stable against flip of any visible unit also.

Thus $(\mathbf{v}, \{\mathbf{h}_l^*\}_{l=1}^L)$ is one-flip stable and hence a mode of the joint distribution. □

[10]Here we assume that energy function values of any two distinct configurations are different.