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 ($C(n, m_1, m_2)$).

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 $N(0, 1)$
   \[ \phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \]

2. The cumulative distribution function (cdf) of standard normal distribution
   \[ \Phi(x) = \int_{-\infty}^{x} \phi(x)dx = \frac{1}{2} \left[ 1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right) \right] \]
   where $\text{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-t^2} dt$

3. The pdf of a skew normal distribution $\tilde{N}$ with skew parameter $\alpha$
   \[ f(x) = 2\phi(x)\Phi(\alpha x) \]

4. If $X \sim N(\mu, \sigma^2)$, $a \in \mathbb{R}$, $\alpha = \frac{a-\mu}{\sigma}$, then $X$ conditioned on $X > a$ follows a truncated normal distribution with moments
   \[ E[X|X > a] = \mu + \sigma \frac{\phi(\alpha)}{Z} \]
   \[ Var(X|X > a) = \sigma^2 \left[ 1 + \frac{\phi(\alpha)}{Z} - \left(\frac{\phi(\alpha)}{Z}\right)^2 \right] \]
   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 $v$ is perfectly reconstructible for an RBM$_{n,m}(\theta)$ $\iff$ the state $\{v, up(v)\}$ is one-flip stable.

Proof. Let $h^* = up(v)$ (conditioning on $\theta$ is implicit). If $v$ is perfectly reconstructible $\implies v = \arg \max_h P(v|h^*) \implies \forall v \neq v, P(v', h^*) < P(v, h^*)$. Similarly since $h^* = \arg \max_h P(h|v), v' \neq h^*, P(v, h') < P(v, h^*)$. Hence the state $\{v, h^*\}$ is stable against any number of flips of visible units and against any number of flips of hidden units, $\implies \{v, h^*\}$ is atleast one-flip stable.

Conversely let $\{v^*, h^*\}$ be one-flip stable. We shall prove by contradiction that $up(v^*) = h^*$ and $down(h^*) = v^*$. Assume $up(v^*) = h' \neq h^*$. We use the fact that for an RBM the hidden units are conditionally independent of each other given the visible units. Thus $h' = \arg \max_{h_k} P(h|v^*) = \{\arg \max_{h_k} P(h_k|v^*)\}_{j=1}^m$. Further $P(h^*|v^*) = \prod_{j=1}^m P(h^*_j|v^*)$. Let $k$ be an index such that $h'_k \neq h^*_k$. Since $h'_k = \arg \max_{h_k} P(h_k|v^*)$, $\implies P(h'_k|v^*) > P(h^*_k|v^*)$. Moreover, $P(v^*, h^*) = P(v^*)P(h^*|v^*) = P(v^*)\prod_{j=1}^m P(h^*_j|v^*)$. Thus just by flipping $h'_k$ to $h^*_k$ we can increase the probability of the state $\{v^*, h^*\}$. This contradicts the one-flip stability hypothesis. Similarly using the conditional independence of visible units given the hidden units we can show that $down(h^*) = v^*$. \hfill \qed

\(^8\)http://mathonline.wikidot.com/the-squeeze-theorem-for-convergent-sequences
C. Proof of Lemma 2 (See page 5)

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

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

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

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

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

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

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

Case 1: $r = 1$

By symmetry it can be assumed $v_1 = 1$, and $v_i = 0 (\forall i > 1)$. Then $\{ h_j = \mathbb{1}_{w_{ij} > 0} \}_{j=1}^m$. 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

$$\{ \hat{v}_i = \mathbb{1}_{\{x_i = \sum_{j=1}^{m} w_{ij}h_j > 0\}} \}_{i=1}^n.$$ 

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 $\{ P[\hat{v}_i = 1] = \frac{1}{2} \}_{i=2}^n$. 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 $\mathbf{h} = \up(\mathbf{v})$ 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.,

$$\{ R_i = \sum_{j=1}^{l} w_{ij} > 0 \}_{i=1}^r.$$ 

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 $\mu_{ij} = \sigma \sqrt{\frac{1}{\pi r}}$ and variance $\sigma_{ij}^2 = \sigma^2 \left( 1 - \frac{1}{2\pi} \right)$. Since the random variables $\{ w_{ij}\mid C_j > 0 \}_{j=1}^r$ are independent the posterior mean of $R_i$ shall be $\bar{\mu}_i = l\sigma \sqrt{\frac{1}{\pi r}}$, and the posterior variance $\bar{\sigma}_i^2 = l\sigma^2 \left( 1 - \frac{1}{2\pi} \right)$. Since $l \gg 1$ by Central Limit Theorem $R_i$ follow a normal distribution. Since the $R_i$ are negatively correlated (proof follows) and $\{ P[\hat{v}_i = 1] = \frac{1}{2} \}_{i>1}$ 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 $R_i, R_t (t \neq i)$,

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

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.
D. Proof of Lemma 3 (See page 5)

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

$$E [w_{ij}| \{R_t > 0\}_{i=1}^{n-1}, C_j > 0] = (\tilde{\mu}_c - \mu_c) \frac{\sigma^2_c}{\sigma^2_c}$$

$$\text{Var} [w_{ij}| \{R_t > 0\}_{i=1}^{n-1}, C_j > 0] = \tilde{\sigma}^2_c \left( \frac{\sigma^2_c}{\sigma^2_c} \right)^2 + \sigma^2 \beta$$

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

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

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

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

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

where

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

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 $\{w_{kj}\}_{k=1}^n$ would dominate the sum. Then conditioned on $\{R_t > 0\}_{i=1}^{n-1}, C_j > 0, C_j$ shall be distributed as per truncated normal distribution (Barr & Sherrill, 1999) with moments

$$E [C_j|\{R_t > 0\}_{i=1}^{n-1}, C_j > 0] = \tilde{\mu}_c = \mu_c + \sigma \frac{\phi}{Z}$$

$$\text{Var} [C_j|\{R_t > 0\}_{i=1}^{n-1}, C_j > 0] = \tilde{\sigma}^2_c = \sigma^2_c \left[1 - \frac{\mu_c \phi}{\sigma_c Z} - \frac{\phi^2}{Z^2}\right]$$

where $\sigma^2_c = (i-1)\sigma^2 \left(1 - \frac{2}{\pi l}\right) + (r-i+1)\sigma^2$,

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

**Remark.** The random variables $\{\tilde{w}_{ij} = w_{ij}|\{R_t > 0\}_{i=1}^{n-1}, C_j > 0\}_{j=1}^l$ shall be negatively correlated with one another so we should subtract the covariance terms while determining the effective variance of $\tilde{R}_i = \sum_{j=1}^l \tilde{w}_{ij}$. Thus if we don’t subtract the covariance terms from the variance we would get a lower bound on the posterior probability of $R_t$ 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 $RBM_{n,m}$ obeys the following inequality.

$$\frac{1}{n} \log_2(L(n,m)) \leq C(n,m) \leq \frac{1}{n} \log_2(U(n,m))$$
Proof. The upper bound follows from Lemma 2 and applying linearity of expectation.

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

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

\[
L_{n,m} = \sum_{r=1}^{n} \left( \frac{n}{r} \right) \sum_{l=1}^{m} \left( \frac{m}{l} \right) \left( \frac{1}{2} \right)^m \left\{ \prod_{i=1}^{r} \left[ \frac{1}{2} - \frac{1}{2} \text{erf} \left( -\frac{\mu_i(r,l)\sqrt{2}}{\hat{\sigma}_i(r,l)} \right) \right] \right\} \left( \frac{1}{2} \right)^{n-r}
\]

\[ \blacksquare \]

F. Proof of Corollary 1 (See page 5)

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

Proof. We shall show that \( \lim_{m \to \infty} U_{n,m} \leq 1.5^n \) and \( \lim_{m \to \infty} L_{n,m} \geq 1.5^n \). Then using Squeeze Theorem and the fact that limits preserve inequalities the result shall hold.

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

If we replace the \( l \) inside the \( \text{erf} \) function by \( m \) then we would be increasing the value of the expression since \( m \geq l \). Thus

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

\[
= \lim_{m \to \infty} \sum_{r=1}^{n} \left( \frac{n}{r} \right) \sum_{l=1}^{m} \left( \frac{m}{l} \right) \left( \frac{1}{2} \right)^m \left[ 2 \right]^r \left( \frac{1}{2} \right)^{n-r}
\]

\[
= 1.5^n
\]

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

\[
\lim_{m \to \infty} L_{n,m} = \lim_{m \to \infty} \sum_{r=1}^{n} \left( \frac{n}{r} \right) \sum_{l=1}^{m} \left( \frac{m}{l} \right) \left( \frac{1}{2} \right)^m \left\{ \prod_{i=1}^{r} \left[ \frac{1}{2} - \frac{1}{2} \text{erf} \left( -\frac{\mu_i(r,l)\sqrt{2}}{\hat{\sigma}_i(r,l)} \right) \right] \right\} \left( \frac{1}{2} \right)^{n-r}
\]

\[
\geq \lim_{m \to \infty} \sum_{r=1}^{n} \left( \frac{n}{r} \right) \sum_{l=m}^{m} \left( \frac{m}{l} \right) \left( \frac{1}{2} \right)^m \left\{ \prod_{i=1}^{r} \left[ \frac{1}{2} - \frac{1}{2} \text{erf} \left( -\frac{\mu_i(r,l)\sqrt{2}}{\hat{\sigma}_i(r,l)} \right) \right] \right\} \left( \frac{1}{2} \right)^{n-r}
\]

\[
\geq \lim_{m \to \infty} \sum_{r=1}^{n} \left( \frac{n}{r} \right) \sum_{l=m}^{m} \left( \frac{m}{l} \right) \left( \frac{1}{2} \right)^m \left\{ \prod_{i=1}^{r} \left[ \frac{1}{2} - \frac{1}{2} \text{erf} \left( -\frac{\mu_i(r,l)\sqrt{2}}{\hat{\sigma}_i(r,l)} \right) \right] \right\} \left( \frac{1}{2} \right)^{n-r}
\]
Since $\tilde{\mu}_i(r, l)$ and $\tilde{\sigma}_i(r, l)$ are non-zero finite quantities regardless of the value of $l$ and $m$ and $\epsilon$ is a fixed non-zero constant,

$$
\lim_{m \to \infty} L_{n,m} \geq \lim_{m \to \infty} \sum_{r=1}^{n} \binom{n}{r} \sum_{l \leq m} \binom{m}{l} \left( \frac{1}{2} \right)^m \left\{ \prod_{i=1}^{r} \left[ 1 - \frac{1}{2} \text{erf} \left( \frac{\tilde{\mu}_i(r, l) \sqrt{\frac{2}{\pi \epsilon}}} {\tilde{\sigma}_i(r, l)} \right) \right] \right\} \left( \frac{1}{2} \right)^{n-r}
$$

$$
= \lim_{m \to \infty} \sum_{r=1}^{n} \binom{n}{r} \sum_{l \leq m} \binom{m}{l} \left( \frac{1}{2} \right)^m \left\{ \left( \frac{1}{2} \right)^r \right\} \left( \frac{1}{2} \right)^{n-r}
$$

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

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

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

\[ \Box \]

**G. Proof of Theorem 2 (See page 6)**

**Theorem 2. (ISC of RBM) For an RBM $n_1, m_2 (n_1 > 0$ and $m_2 \geq 0$), if we denote $u = \max(m_1, n + m_2), l = \min(m_1, n + m_2)$, then**

$$
C(n_1, m_2) \leq \frac{1}{n} \log_2 S
$$

whenever $S < 2^n$, $S = \left[ 1 - \frac{1}{2} \text{erf} \left( -\sqrt{\frac{u}{\pi \epsilon}} \right) \right]^{l_i}$

**Proof.** As shown in Figure 2 we construct a single layer RBM $n_1, m_2$, that has the same bipartite connections as RBM $n_1, m_2$. The expected number of perfectly reconstructible vectors for the single layer RBM can then be obtained from Equation 10.

$$
C(n_1, m_2) \leq \frac{1}{n} \log_2 U_{n_1, m_2, \epsilon} = \frac{1}{n} \log_2 S
$$

$$
= \frac{1}{n} \log_2 \left[ 1 - \frac{1}{2} \text{erf} \left( -\sqrt{\frac{u}{\pi \epsilon \epsilon}} \right) \right]^{l_i}
$$

However this quantity is an overestimate. This counts the number of pairs of vectors $\{v, h_2\}$ such that $\left( \begin{array}{c} \hat{v} \\ h_2 \end{array} \right)$ is perfectly reconstructible for RBM $n_1, m_2, \epsilon$. Among these, there can be vectors like $\left( \begin{array}{c} v^{(1)} \\ h_2^{(1)} \end{array} \right)$ and $\left( \begin{array}{c} v^{(2)} \\ h_2^{(2)} \end{array} \right)$ where $v^{(1)} = v^{(2)}$ resulting in repetitions. Assuming such vectors $v^{(i)}$ 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 $S < 2^n$ then binomial approximation can be applied and we get the desired result. \[ \Box \]

**H. Proof of Corollary 2 (See page 6)**

**Corollary 2. (Layer 1 Wide, Layer 2 Narrow) For an RBM $n_1, m_2 (n_1 > 0$ and $m_2 \geq 0$), if $\alpha_1 = \frac{m_1}{n_1} > \frac{1}{\gamma}$ and $\alpha_2 = \frac{m_2}{n_1} < \gamma$ then**

$$
C(n_1, m_2) \leq (1 + \alpha_2) \log_2(1.5)
$$
Proof. For $\alpha_1 > \frac{1}{\gamma}$, $S = \left[ 1 - \frac{1}{2} \operatorname{erf} \left( -\sqrt{\frac{\alpha_1}{\pi n(1+\alpha_2)}} \right) \right]^{n(1+\alpha_2)} = 1.5^{\alpha(1+\alpha_2)}$.

Moreover for $\alpha_2 < \gamma$, since $S = 1.5^{n(1+\alpha_2)} < 1.5^{n(1+\gamma)} = 2^{n(1+\gamma)\log_2(1.5)} = 2^{0.614n}(< \gamma^2n$ 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 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, $\max_{\alpha_1, \alpha_2} C(n, \alpha_1, \alpha_2) \leq \hat{U}(n, \alpha_1^*, \alpha_2^*)$ where

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

Proof. We consider two regimes.

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

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

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

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

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

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

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

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

$$\frac{1}{S} \frac{\partial S}{\partial \alpha_1} \geq n \ln \left[ 1 - \frac{1}{2} \operatorname{erf} \left( -\sqrt{\frac{1}{\pi}} \right) \right] - \frac{n}{\sqrt{\pi}} \left[ \frac{1}{\sqrt{\pi}} \exp \left( -\frac{1}{\pi} \right) \right] \left\{ \prod_{x \exp(-x^2) \leq 0.428} \right\}$$

$$= 0.252n - 0.187n$$

$$\Rightarrow \frac{\partial S}{\partial \alpha_1} > 0$$

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 $\alpha_1 = c (c < 1)$.
Using Inherent Structures to design Lean 2-layer RBMs

Layering: Comparison with simulation for $n = 3$

Prediction (upper bound) | Actual Simulation
---|---

Layering: Comparison with simulation for $n = 5$

Prediction (upper bound) | Actual Simulation
---|---

Layering: Comparison with simulation for $n = 10$

Prediction (upper bound) | Actual Simulation
---|---

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

Proof. Since $v$ is a mode, $\implies p(v) = \frac{1}{Z} > 0$.

Further, let $\{h_l^*\}_{l=1}^L = \max_{h_l} P(v, \{h_l\}_{l=1}^L)$, that is, the state $(v, \{h_l^*\}_{l=1}^L)$ is stable against flip of any hidden unit. Moreover, since for all neighbours $v'$ of $v$, $p(v') = 0 \implies p(v', \{h_l^*\}_{l=1}^L) = 0$, it implies that $(v, \{h_l^*\}_{l=1}^L)$ is stable against flip of any visible unit also.

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

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