Using Inherent Structures to design
Lean 2-layer RBMs

Abhishek Bansal [1]   **Abhinav Anand** [2]   **Chiranjib Bhattacharyya** [2]

Abstract

Understanding the representational power of Restricted Boltzmann Machines (RBMs) with multiple layers is an ill-understood problem and is an area of active research. Motivated from the approach of *Inherent Structure formalism* (Stillinger & Weber, 1982), extensively used in analysing Spin Glasses, we propose a novel measure called *Inherent Structure Capacity* (**ISC**), which characterizes the representation capacity of a fixed architecture RBM by the expected number of modes of distributions emanating from the RBM with parameters drawn from a prior distribution. Though **ISC** is intractable, we show that for a single layer RBM architecture **ISC** approaches a finite constant as number of hidden units are increased and to further improve the **ISC**, one needs to add a second layer. Furthermore, we introduce *Lean* RBMs, which are multi-layer RBMs where each layer can have at-most $O(n)$ units with the number of visible units being n. We show that for every single layer RBM with $\Omega(n^{2+r}), r \geq 0$, hidden units there exists a two-layered *lean* RBM with $\Theta(n^2)$ parameters with the same **ISC**, establishing that 2 layer RBMs can achieve the same representational power as single-layer RBMs but using far fewer number of parameters. To the best of our knowledge, this is the first result which quantitatively establishes the need for layering.

[1]IBM Research [2]Dept of CSA, IISc, Bengaluru, India. Correspondence to: Abhishek Bansal <abbansal@in.ibm.com>, Abhinav Anand <abhinava@iisc.ac.in>, Chiranjib Bhattacharyya <chiru@iisc.ac.in>.

1. Introduction

Deep Boltzmann Machines (DBMs[1]) are largely tuned using empirical methods based on trial and error. Despite much effort, there is still very little theoretical understanding about why a particular neural network architecture works better than another for any given application. Furthermore there is no well defined metric to compare different network architectures.

It is known that given any input distribution on the set of binary vectors of length n, there exists an RBM with $\alpha 2^n - 1$ ($\alpha < 1$) hidden units that can approximate that distribution to an arbitrary precision (Montúfar & Rauh, 2017). However with these many hidden units the number of parameters increase exponentially. We call a network *lean* if for each layer, the number of hidden units $m = \mathcal{O}(n)$ where n is the number of visible units. The deep narrow Boltzmann Machines whose universal approximation properties were studied in (Montúfar, 2014) are a special case of lean networks. In this paper we study lean 2-layer deep RBMs.

We ask the questions, is there a measure that can relate DBM architectures to their representational power? Once we have such a measure then can we gain insights into the capabilities of different DBM architectures?

For example, given a *wide* single layer RBM, an RBM with many hidden nodes, can we find a *lean* multilayer RBM with equivalent representational power but with far lesser parameters? Despite much effort these questions are not satisfactorily answered and may provide important insights to the area of Deep Learning.

Our main contributions are as follows:

1. We study the *Inherent structures* formalism, first introduced in Statistical Mechanics(Stillinger & Weber, 1982), to understand the configuration space of RBMs. We introduce a capacity measure Inherent Structure Capacity (**ISC**) (Definition 5) and discuss its relation with the expected number of perfectly reconstructible vectors (Montúfar & Morton, 2015), one-flip stable states and the modes of the input distribu-

---

[1]We shall use the terms RBM and DBM interchangebly

tion. We use this as a measure of representation power of an RBM.

2. Existing methods for computing expected number of inherent structures are rooted in Statistical Mechanics. They use the *replica* approach (Bray & Moore, 1980) which does not extend well to DBMs since it is not straightforward to incorporate the bipartite nature and layering in the calculations. We use a first principles approach to devise a method that yields upper and lower bounds for single layered and two-layered DBMs (Theorems 1,2). We show that the bounds become tight as we increase the number of hidden units.

3. Previous results have shown that a sufficiently large single layer RBM can represent any distribution on the 2^n input visible vectors. However we show that if we continue adding units to hidden layer then the **ISC** tapers to 0.585 as opposed to the maximum limit of 1.0 (Corollary 2). This implies that although an RBM is a universal approximator, if the input distribution contains large number of modes multi-layering should be considered. We have empirically verified that when the number of units in a single hidden layer RBM, $m \geq 20n$, the **ISC** saturates (Figure 3).

4. By analyzing the **ISC** for two layer RBM we obtain an interesting result that for any such RBM with $m = \Omega(n^2)$ hidden units (number of parameters $\Omega(n^3)$) one can construct a two layered DBM with $1.6n$ units in hidden layer 1 and $0.6n$ units in layer 2 (Corollary 4) and with number of parameters $\Theta(n^2)$, resulting in an order of magnitude saving in parameters. To the best of our knowledge this is the first such result which establishes the superiority of 2 layer DBMs over wide single layer RBMs in terms of representational efficiency. We conduct extensive experiments on synthetic datasets to verify our claim.

2. Model Definition and Notations

An RBM with n visible and m hidden units, denoted by RBM$_{n,m}(\theta)$, is a probability distribution on $\{0,1\}^n$ of the form

$$P(\mathbf{v}, \mathbf{h}|\theta) = \frac{e^{-E(\mathbf{v},\mathbf{h}|\theta)}}{Z(\theta)} \qquad (1)$$

$$E(\mathbf{v}, \mathbf{h}|\theta) = -\mathbf{a}^T\mathbf{v} - \mathbf{b}^T\mathbf{h} - \mathbf{v}^T\mathbf{W}\mathbf{h} \qquad (2)$$

where $\mathbf{v} \in \{0,1\}^n$ denotes the visible vector, hidden vector is denoted by $\mathbf{h} \in \{0,1\}^m$, the parameter $\theta = \{\mathbf{a}, \mathbf{b}, \mathbf{W}\}$ denotes the set of biases $\mathbf{a} \in \mathbb{R}^n, \mathbf{b} \in \mathbb{R}^m$ and coupling matrix $\mathbf{W} \in \mathbb{R}^{n \times m}$ and $Z(\theta) = \sum_{\mathbf{v},\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}$ is the normalization constant. The log-likelihood of a given visible vector $\mathbf{v}$ for an RBM$_{n,m}(\theta)$ is given by

$$\mathcal{L}(\mathbf{v}|\theta) = \ln P(\mathbf{v}|\theta) = \ln \sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})} - \ln Z(\theta) \qquad (3)$$

In the sequel **RBM**$_{n,m}$ will denote the family of distributions parameterized by RBM$_{n,m}(\theta)$.

Definition 1 (Modes). Given a distribution p on vectors $\{0,1\}^n$, a vector $\mathbf{v}$ is said to be a mode of that distribution if for all $\mathbf{v}'$ such that $d_H(\mathbf{v}, \mathbf{v}') = 1$, $p(\mathbf{v}) > p(\mathbf{v}')$. Here $d_H(\mathbf{v}, \mathbf{v}') = \sum_{i=1}^{N}[1-\delta(\mathbf{v}_i, \mathbf{v}'_i)]$ is the Hamming distance[2].

Definition 2 (Perfectly Reconstructible Vectors). For an RBM$_{n,m}(\theta)$ we define the function up $: \{0,1\}^n \to \{0,1\}^m$ that takes a visible vector $\mathbf{v}$ as input and outputs the most likely hidden units vector $\mathbf{h}$ conditioned on $\mathbf{v}$, i.e., up$(\mathbf{v}) \triangleq \arg\max_{\mathbf{h}} P(\mathbf{h}|\mathbf{v}, \theta)$. Similarly down$(\mathbf{h}) \triangleq \arg\max_{\mathbf{v}} P(\mathbf{v}|\mathbf{h}, \theta)$. A visible units vector $\mathbf{v}$ is said to be *perfectly reconstructible* (**PR**) if down$($up$(\mathbf{v})) = \mathbf{v}$.

For any set C the cardinality will be denoted by $|C|$. For an RBM$_{n,m}(\theta)$ we define

$$\text{prv}(n, m, \theta) \triangleq |\{\mathbf{v} : \mathbf{v} \text{ is } \mathbf{PR} \text{ for RBM}_{n,m}(\theta)\}|$$

3. Problem Statement

We consider fitting an RBM$_{n,m}(\theta)$ to a distribution $p(\mathbf{v}) = \frac{1}{k}\sum_{i=1}^{k}\delta(\mathbf{v} - \mathbf{v}_i)$ where δ denotes the Dirac Delta function and where for each pair of vectors $\{\mathbf{v}_i, \mathbf{v}_j\}$ in $\{\mathbf{v}_r\}_{r=1}^{k}$, $d_H(\mathbf{v}_i, \mathbf{v}_j) \geq 2$. We need to find the smallest $m^* \in \mathbb{N}$ such that the set **RBM**$_{n,m^*}$ contains an RBM$_{n,m^*}(\theta)$ that represents p. We also study the case of a DBM with 2 hidddden layers. We denote a DBM with n visible units, L hidden layers with m_k hidden units in layer k by RBM$_{n,m_1,...,m_L}(\theta)$. We denote the respective set of DBMs by **RBM**$_{n,m_1,...,m_L}$. We would like to ask the following question. Are there *lean* two layer architectures, **RBM**$_{n,m_1,m_2}$ which can model distributions with the same number of modes as that of distributions generated by a one layer architecture **RBM**$_{n,m}$ where $m \gg m_1, m_2$.

3.1. Related Work

The representational power of Restricted Boltzmann Machines (RBMs) is an ongoing area of study (Le Roux & Bengio, 2008; Montúfar et al., 2011; van der Maaten, 2011; Martens et al., 2013; Cueto et al., 2010).It is well known that an RBM with one hidden layer is a universal approximator (Le Roux & Bengio, 2008; Montufar & Ay, 2011; Montúfar & Rauh, 2017). (Le Roux & Bengio, 2008) showed that the set **RBM**$_{n,m}$ can approximate any input

---

[2]We shall use d_H to denote Hamming distance between two vectors. $\delta(x,y)$ is the kronecker distance and is defined as $\delta(x,y) = 1$ whenever $x = y$, and 0 otherwise.
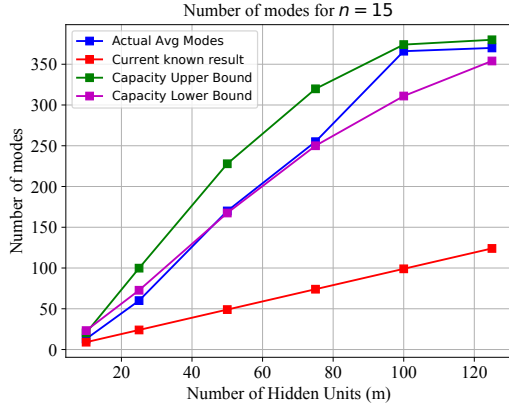
Figure 1. Number of modes attained for different choices of hidden units for $n = 15$. Can be seen that the current known result for the number of hidden units required (red graph) is a large over-estimate. The green and purple graphs are estimates given by Theorem 1. These are closer to the actual number of enumerated modes, given by the blue graph.

distribution with support set size k arbitrarily well if following inequality is satisfied.

$$m \geq k + 1 \quad (4)$$

If we know the number of modes of our input distribution, then we could design our RBM as per Eqn (4). Unfortunately the number of modes could be large resulting in a large RBM.

To test the bound in Eqn (4), we conducted simulation experiments. We kept $n = 15$, $m \in \{10, 25, 50, 75, 100, 125\}$ and generated random coupling weight matrix whose entries were i.i.d. $\mathcal{N}(0, 1)$ and enumerated all the modes of the generated distribution. We averaged our readings over 100 different weight matrices. The results are shown in Figure 1. The results show that the bound gives a highly conservative estimate for k. *For example on average the set $\mathbf{RBM}_{15,50}$ has the capability to represent distributions with 170 modes, instead of only 49 modes.* Thus although the number of modes is an important design criteria, a more practical metric is desirable.

4. Inherent structures of RBM

To understand the complex structure in Spin glasses the notion of *Inherent Structures*(IS) was introduced in (Stillinger & Weber, 1982). The IS approach consists of partitioning the configuration space into *valleys*, where each valley consists of configurations in the vicinity of a local minimum. The number of such valleys can thus be indicative of *Complexity* of the system.

In this section we recall the IS approach in a general setting

to motivate a suitable capacity measure. Consider a system governed by the probability model

$$P(\mathbf{s}|W) = \frac{e^{-\frac{1}{T}E(\mathbf{s}|W)}}{Z(W)}, Z(W) = \sum_{\mathbf{s}} e^{-\frac{1}{T}E(\mathbf{s})} \quad (5)$$

where $E : \{0,1\}^N \to \mathbb{R}$ is an energy function defined over N dimensional binary vectors with parameter W.

Definition 3 (One-flip Stable States). (Stein & Newman, 2013) For an Energy function E a configuration, $\mathbf{s}^*$ is called a local minimum, also called One flip stable state, if $\forall \mathbf{s} \in \{s : d_H(\mathbf{s}, \mathbf{s}^*) = 1\}, E(\mathbf{s}) - E(\mathbf{s}^*) > 0$ (equivalently $P(\mathbf{s}) < P(\mathbf{s}^*)$).

For every one-flip stable state $\mathbf{s}^*$ we define the set $OF(\mathbf{s}^*) = \{\mathbf{s}| \ d_H(\mathbf{s}, \mathbf{s}^*) \leq 1\}$. Let $\{\mathcal{P}_1, \ldots, \mathcal{P}_K\}$ form a partition of the configuration space where each $\mathcal{P}_a = OF(\mathbf{s}_a)$ corresponds to the local minimum $\mathbf{s}_a$ and K is the total number of valleys [3]. The logarithm of the partition function

$$\log Z(W) = \log \sum_{\mathbf{s}} e^{-\frac{1}{T}E(\mathbf{s})} = \log \sum_{a=1}^{K} Z_a(W)$$

where $Z_a(W) = \sum_{\mathbf{s} \in P_a} e^{-\frac{1}{T}E(\mathbf{s})}$. Now, for any p in a K dimensional probability simplex, using the non-negativity of KL divergence, it is straightforward to show that

$$H(p) + \sum_{a=1}^{K} p_a \log Z_a(W) \leq \log Z(W) \quad (6)$$

where $H(p) = -\sum_{a=1}^{K} p_a \log p_a$ is the entropy of p. Equality holds whenever $p_a^* = \frac{Z_a}{Z}, \forall a \in \{1, \ldots, K\}$. One could construct $\log Z$ from $\log Z_a$ if one had access to p^*, and knew K^* which is defined as $H(p^*) = \log K^*$.

$$\log K^* + \sum_{a=1}^{K} p_a^* \log Z_a(W) = \log Z(W) \quad (7)$$

From the properties of entropy function one could write

$$1 \leq K^* \leq K \leq 2^N \quad (8)$$

where the lower bound on K^* is attained at $H(p^*) = 0$ and is realized when the Energy surface has only one local minimum, a very un-interesting case. The upper bound on K^* is attained at $p^* = \frac{1}{K}$, which happens only when all valleys are considered similar. Since Number of states can be at most 2^N, the last upper bound holds and Thus $\frac{1}{N} \log_2 K$, can be viewed as a measure of Complexity, of

---

[3]Here we assume that temperature parameter T is small so that all states with Hamming distance > 1 from a one-flip stable state will have negligible contribution to partition function.

the energy surface. One could put a suitable prior distribution over the parameters W and evaluate the complexity averaged over the prior, motivating the following definition.

Definition 4. (**Complexity**) The *Complexity* of the model described in Eqn (5) is given by

$$\frac{1}{N}\mathbb{E}_W \log_2 K, \quad W \sim \mathbf{P},$$

where K is the number of One-Flip stable states for Energy function defined with parameter W and $\mathbf{P}$ is a prior distribution over W.

For Ising models, *Complexity* has been estimated in the large N limit (Bray & Moore, 1980; Tanaka & Edwards, 1980) by methods such as *Replica technique*. However, extending their methods to RBMs for a finite size N is not straightforward.

It has been shown (see e.g. (Parisi & Potters, 1995)) that IS decomposition gives a very accurate picture of energy landscape of Ising models at Temperature, $T = 0$. But, for $T > 0$, one needs to take into account both the Valley structure and the energy landscape of the free energy (Biroli & Monasson, 2000). Obtaining accurate estimates of Complexity is an active area of study, for a recent review see (Auffinger et al., 2013).

Our goal is to apply the aforementioned IS decomposition to RBMs. We now show the equivalence between these perfectly reconstructible vectors and one-flip stable states for an RBM. The IS decomposition then allows us to define the measure of capacity in terms of the modes of input distribution.

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

Proof. See Supplementary material. □

Thus we see that there is a one-one equivalence between perfectly reconstructible vectors and the one-flip stable states for a single layer $\mathrm{RBM}_{n,m}(\theta)$.

Relationship between the modes of $p(\mathbf{v})$ and $p(\mathbf{v}, \mathbf{h})$ In this section we discuss the relationship between the modes of the marginal distribution, $p(\mathbf{v})$ and the joint distributuon $p(\mathbf{v}, \mathbf{h})$. We make a mild assumption on one-flip stable state.

  A1 For a single layer RBM, given a visible vector $\mathbf{v}$, vector $\mathbf{h}^* = \mathrm{up}(\mathbf{v})$ is unique.

If the weights are given small random perturbation, then Assumption 1 holds with probability one. However it does not hold true for an $L \geq 2$ layer $\mathrm{RBM}_{n,m_1,\dots,m_L}(\theta)$. We

denote $\{\mathbf{h}_l \in \{0,1\}^{m_l}\}_{l=1}^L$ to be hidden vectors, $\mathbf{v} \in \{0,1\}^n$ to be visible vector and define the set

$$\mathcal{H}(\mathbf{v}) \triangleq \{\{\mathbf{h}_l\}_{l=1}^L | (\mathbf{v}, \{\mathbf{h}_l\}_{l=1}^L) \text{ is one-flip stable state}\}$$

It can be seen that $|\mathcal{H}(\mathbf{v})|$ can be more than one. For input distributions considered in Section 3, the modes of joint distribution $p(\mathbf{v}, \mathbf{h}_1, \dots, \mathbf{h}_L)$ with distinct $\mathbf{v}$ are atleast as many as modes of marginal distribution p($\mathbf{v}$). A formal statement with proof is given in supplementary material.

As discussed, for $L \geq 1$, the modes of the marginal distribution could be smaller than modes of the joint distribution. However, (Montúfar & Morton, 2015) [Theorem 1.6] gave precise conditions under which the number of modes for marginal and joint distributions are same for a single layer network. We suspect that a similar argument holds for $L > 1$. For the rest of the paper we will assume that the modes of joint distribution are same as those of $p(\mathbf{v})$.

Armed with these observations we are now ready to define a measure which relates the architecture of a DBM and the expected number of such modes under a prior distribution on the model parameters. More formally,

Definition 5 (**Inherent Structure Capacity**). For an L layered DBM with $m_1, \dots, m_L$ hidden units and n visible units we define the *Inherent Structure Capacity* (**ISC**), denoted by $C(n, m_1, \dots, m_L)$, to be the logarithm (divided by n) of the expected number of modes of all possible distributions generated over the visible units by the DBM.

$$\mathcal{C}(n, m_1, \dots, m_L) = \frac{1}{n} \log_2 \mathbb{E}_\theta \left[|\{\mathbf{v} : |\mathcal{H}(\mathbf{v})| \geq 1\}| \right]$$

We note that for the single layer case this definition reduces to $\frac{1}{n}\log_2 \mathbb{E}_\theta [\mathrm{prv}(n, m, \theta)]$. **ISC** as a measure would be useful in identifying DBM architectures which can model modes of an input distribution defined over the visible units.

This measure serves as a recipe for fitting DBMs. Suppose we know that the input distribution has k modes then one could find a suitable DBM architecture, i.e. $m_1, \dots, m_L$ by the following criterion

$$\frac{1}{n} \log_2 k \leq \mathcal{C}(n, m_1, \dots, m_L) \quad (9)$$

Once the architecture has been identified one can then use a standard learning algorithm to learn parameters to fit a given distribution.

In the following sections we investigate the computation of **ISC** and their applications to single and two layer networks, i.e. $L = 1$ and $L = 2$. To keep the exposition simple we assume the bias parameters to be zero [4]. We

---

[4]Analysis can be extended to non-zero biases in straightforward manner.

also assume that the coupling weights are distributed as per mean zero Gaussian, i.e., $\forall i, j, w_{ij} \sim \mathcal{N}(0, \sigma^2)$.

5. Computing capacity of $\mathbf{RBM}_{n,m}$ and need for more layers

In this section we discuss the computation of **ISC** for a single layer RBM. In absence of a definitive proof we conjecture that **ISC** is intractable just like the Complexity measure in Spin glasses. The problem of computing Complexity has been addressed in the Statistical Mechanics community using the Replica method (Roberts, 1981; De Dominicis et al., 1980) which yields reasonable estimates. However the applicability of Replica trick to Multi-layer DBMs is not clear. In this section we develop an alternative method for estimating **ISC**.

5.1. Computing ISC of $\mathbf{RBM}_{n,m}$

For any arbitrary vector $\mathbf{v} \in \{0, 1\}^n$ we compute $\mathbb{E}\left[\mathbb{1}_{[\mathbf{v} \text{ is PR.}]}\right]$ where $\mathbb{1}$ is the indicator random variable and expectation is over the model parameters θ with prior as stated in Section 4. We then sum this over all 2^n vectors, i.e., $\sum_{\mathbf{v}} \mathbb{E}\left[\mathbb{1}_{[\mathbf{v} \text{ is PR.}]}\right]$. Before stating our main theorem we state a few Lemmas.

Lemma 2. *For the set $\mathbf{RBM}_{n,m}$, if a given vector $\mathbf{v}$ has $r(\geq 1)$ ones, $\mathbf{h} = \mathrm{up}(\mathbf{v})$ has l ones and $l \gg 1$, then[5] 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. See Supplementary Material. □

For $r(> 1)$ ones in $\mathbf{v}$ and l ones in $\mathbf{h} = \mathrm{up}(\mathbf{v})$ the problem of computing $\{P[[\mathrm{down}(\mathbf{h})]_i = 1]\}_{i=1}^r$ 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., $\left\{R_i = \sum_{j=1}^l w_{ij} > 0\right\}_{i=1}^r$. Conditioned on the fact $\{C_j > 0\}_{j=1}^l$ the random variables $\{R_i\}_{i=1}^r$ are negatively correlated. This gives us an upper bound mentioned in Lemma 2. We now get a lower bound for the estimate.

---

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

Lemma 3. *For the set $\mathbf{RBM}_{n,m}$, if $\mathbf{v}$ has $r(> 1)$ ones, $\mathbf{h} = \mathrm{up}(\mathbf{v})$ 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*

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

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

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

Proof. See Supplementary Material. □

Lemma 2 gives an upper bound $U(n, m)$ on expected number of *PR* vectors while Lemma 3 gives us a posterior distribution on w_{ij} after taking into account the conditional correlation between $\{R_i\}_{i=1}^r$. This eventually results in a lower bound $L(n, m)$. Thus even though a closed-form expression for **ISC** is difficult, we obtain bounds on it as the following theorem states.

Theorem 1. (ISC of $\mathbf{RBM}_{n,m}$) *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 $\mathbf{RBM}_{n,m}$ obeys the following inequality.*

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

Proof. See Supplementary material. □

5.2. Need for more hidden layers

Theorem 1 establishes the lower and upper bounds for **ISC**. A direct corollary of the theorem establishes that $C(n, m)$ approaches a limit as m increases.

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

Proof. In the Supplementary material we show that $\lim_{m \to \infty} \frac{1}{n}\log_2 L(n, m) = \lim_{m \to \infty} \frac{1}{n}\log_2 U(n, m) = \log_2 1.5$. Then claim follows from squeeze theorem[6]. □

Empirically we observe that this saturation limit is achieved when $m > 20n$ (see Figure 3). Here we discuss the implications of the results derived in the previous subsection.

1. We plotted the actual expected modes attained and the **ISC** estimates derived from Theorem 1 for $n = 15$ and varying number of hidden units (Figure 1). We

---

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

can see that even a small number of hidden units admits a large **ISC** and the current known bound given in Equation 4 is not necessary. This shows that for a large class of distributions we give a more practical estimate of number of hidden units required than the current state of the art.

2. The upper bound on the **ISC** estimated above seems surprising at first sight since it seems to contradict the well established fact that RBMs are universal approximators (Freund & Haussler, 1992; Le Roux & Bengio, 2008). However, one should note that the bound is in expected sense which means that in the family **RBM**$_{n,(m\to\infty)}$ many RBMs shall have modes close to or less than $(1.5)^n$. For the class of input distributions for which number of modes $k \gg (1.5)^n$ training an RBM$_{n,m}(\theta)$ to represent these might be difficult. *The need for multi-layering arises in such conditions.*

3. Corollary 1 shows for a large enough m the bounds become tight and the expression is exact. We also show this through simulations in Section 7.

Remark. When $n, m \to \infty$ we can approximate $U(n, m)$ by the following relatively simple expression that we can use to conduct further analysis.

$$U(n,m) \approx \left[1 - \frac{1}{2}\operatorname{erf}\left(-\sqrt{\frac{m}{\pi n - 4}} \right) \right]^n \qquad (10)$$

6. ISC of two-layer RBM architecture

To study the effect of adding layers, we consider the family **RBM**$_{n,m_1,m_2}$. As stated in Section 4, adapting analysis for single layer RBMs to multi-layer RBMs is not straightforward. In this section we discuss the computation of **ISC** and study its application to design RBMs.

6.1. Computing the capacity of 2 layer RBM

We observe that an **RBM**$_{n,m_1,m_2}$ shares the same bipartite structure as a single layer **RBM**$_{n+m_2,m_1}$ (Figure 2). This enables us to extend our single layer result to two layers. We introduce a threshold quantity $\gamma = 0.05$. This value was obtained by simulating the asymtotics of $f(x) = 1 - 0.5\operatorname{erf}(-\sqrt{\frac{x}{\pi}})$.

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. See Supplementary material. □

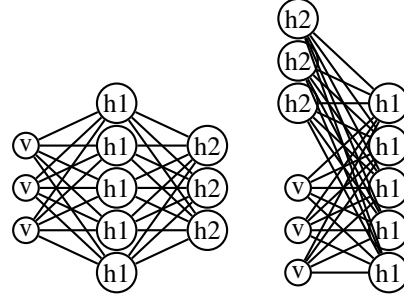*Figure 2.* Two Layer RBM$_{n,m_1,m_2}$ shares same bipartite graph structure as single layer RBM$_{n+m_2,m_1}$

Theorem 2 gives a general formula from which different regimes can be derived by varying m_1, m_2. We will use this theorem to understand the design of multi-layer RBMs.

In the previous section we saw that in a single layer RBM, irrespective of number of hidden units, **ISC**, achieves a limiting value of 0.585. The theorem will be useful to quantitatively show that **ISC** can indeed be improved if we consider layering. For an **RBM**$_{n,m_1,m_2}$ ($n, m_1 > 0$ and $m_2 \geq 0$), we denote $\alpha_1 = \frac{m_1}{n}, \alpha_2 = \frac{m_2}{n}$. We say that a layer with m hidden units is *narrow* if $m < \gamma$ and it is *wide* if $m > \frac{1}{\gamma}$.

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. See Supplementary material. □

The Corollary shows that for a RBM with a wide first layer and a narrow second layer, the upper bound on **ISC** increases linearly with the number of units in second layer.

6.2. RBM$_{n,m_1,m_2}$ design under budget on parameters

We extend the result obtained in previous section to consider a real scenario wherein we have a budget on the maximum number of parameters that we can use and we have to design a two-layered DBM given this constraint. For a given input distribution with k modes, the DBM should have $\mathcal{C}(n, m_1, m_2) > \frac{1}{n}\log_2 k$.

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**, $\max_{\alpha_1,\alpha_2} \mathcal{C}(n, \alpha_1, \alpha_2) \leq \tilde{U}(n, \alpha_1^*, \alpha_2^*)$ where*

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

Table 1. **ISC** Values for different $\alpha_1 = \frac{m_1}{n}, \alpha_2 = \frac{m_2}{n}$. $\gamma = 0.05$ (obtained by simulating the asymtotics of $f(x) = 1 - 0.5 \operatorname{erf}(-\sqrt{\frac{x}{\pi}})$).

Regime	ISC	Implications
$\alpha_1 > \frac{1}{\gamma}, \alpha_2 < \gamma$	$(1 + \alpha_2) \log_2(1.5)$	**ISC** determined only by α_2. For a single layer RBM ($\alpha_2 = 0$), further increase in hidden units not effective, multi-layering recommended.
$\alpha_1(1 + \alpha_2) = c$, where $c \geq 1$	$\min[1, \sqrt{c} \log_2(1.29)]$ $(\alpha_1^* = \sqrt{c})$.	Given a budget of cn^2 parameters, this is the maximum **ISC** achievable with optimal choice of α_1.
$\alpha_1(1 + \alpha_2) = c$, where $c < 1$	$c \log_2 \left[1 - \frac{1}{2} \operatorname{erf} \left(-\sqrt{\frac{1}{\pi c}} \right) \right]$ $(\alpha_1^* = c)$.	If total number of parameters $< n^2$, then multi-layering does not help.

Proof. See Supplementary material. □

Corollary 3 can be used to determine the optimal allocation of hidden units to the two layers if there is a budget on the number of parameters to be used due to computational power or time constraints. It says that if $c \geq 1$, then for optimality $\alpha_1 = 1 + \alpha_2$ and if $c < 1$, then $\alpha_2 = 0$ which means that all hidden units should be added to layer 1. The following corollary highlights the existence of a two layer architecture **RBM**$_{n,m_1,m_2}$ that has **ISC** equal to 0.585, the saturation limit for single layer RBMs.

Corollary 4. *There exists a two layer architecture* RBM_{n,m_1,m_2} *with* $\Theta(n^2)$ *parameters such that*

$$\tilde{U}(n, \alpha_1, \alpha_2) = \log_2 1.5$$

where $m_1 = \alpha_1 n, m_2 = \alpha_2 n, \alpha_1 = 1.6$ *and* $\alpha_2 = 0.6$

Proof. In Corollary 3 if we put $\tilde{U}(n, \alpha_1^*, \alpha_2^*) = \log_2(1.5)$, we get $\alpha_1^* = \sqrt{c} = \frac{\log_2(1.5)}{\log_2(1.29)} = 1.6, \alpha_2^* = \alpha_1^* - 1 = 0.6$. Number of parameters for such an RBM is $\alpha_1(1+\alpha_2)n^2 = \Theta(n^2)$. □

The number of parameters for any single layer RBM is nm where m is number of hidden units. The above corollary gives an important insight: one can construct a two layer RBM with $\Theta(n^2)$ parameters that has the same **ISC** as a single layer RBM with infinitely many hidden units. Ofcourse this is true only if the upper-bound $\tilde{U}$ is close to $\mathcal{C}$. This suggests that *lean* 2 layer networks with *order of magnitude less number of parameters can achieve the same* **ISC** *as that of a single layer RBM.*

Table 1 summarises the **ISC** values for different regimes and their respective implications for the two-hidden layered DBN. For example if $\alpha_1 > \frac{1}{\gamma}$ then the capacity is dictated only by the number of hidden units in the second layer and increasing α_1 has no effect. Multi-layering should be considered to handle distributions with multiple modes. Also, considering a practical scenario where there is a computational and memory constraint that translates into a budget

Table 2. Actual **ISC** for **RBM**$_{n,m}$ for $m = 500$, obtained by averaging brute-force enumeration from 2000 independent instantiations of weight matrix, i.e., $\mathcal{C}(n, m) = \frac{1}{n} \log_2 \left(\frac{1}{2000} \sum_{i=1}^{2000} k_i \right)$ where k_i is the number of modes enumerated in ith instantiation.

	$n = 10$	$n = 11$	$n = 12$
$\mathcal{C}(n, m)$	0.585	0.585	0.588

on the number of parameters, i.e. $\alpha_1(1 + \alpha_2) = c$, we get the optimal distribution of hidden units in the two layers that maximizes the capacity. In particular if $c < 1$ then it is recommended to allocate all hidden units to layer 1 itself instead of adding more layers.

7. Experimental Results

Our main goals are to experimentally verify Theorems 1, 2 and Corollaries 3 4. All experiments were run on CPU with 2 Xeon Quad-Core processors (2.60GHz 12MB L2 Cache) and 16GB memory running Ubuntu 16.02 [7].

7.1. Validating estimate of Number of modes

To verify our theoretical claims of Theorems 1 and 2 a number of simulation experiments for varied number of visible and hidden units were conducted. **To enable execution of exhaustive tests in reasonable time, the values of n had to be kept small**. The entries $\{w_{ij}\}$ of the weight matrix were drawn from an i.i.d. mean zero normal distribution. Each of the $2^n - 1$ vectors (leaving out the trivial all zero vector) was then tested for being *perfectly reconstructible*. A comparison of the theoretical predictions and experimental results is shown in Figures 3 and 5 for single layer and two layer RBMs respectively. It can be seen that the theoretical predictions follow similar trend as the experimental results.

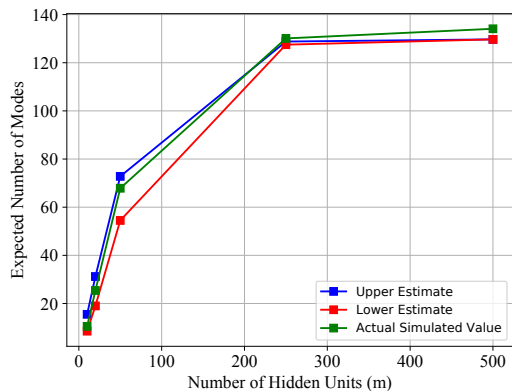
[7]The source code and instructions to run is available at http://mllab.csa.iisc.ernet.in/publications.

Figure 3. Comparison chart of the upper and lower estimates with the actual simulation value of expected number of modes $(2^{n\mathcal{C}(n,m)})$ for $n = 12$.

Discussion. Figure 3 shows that the predicted bounds on the modes are close to the actual modes enumerated. Table 2 validates the claim that for an **RBM**$_{n,m}$ as $m \to \infty$, **ISC** $\to 0.585$ (Corollary 1). To enable bruteforce enumeration in reasonable time the values for n had to be kept small. Figure 5 in the supplementary section shows the theoretical upper bound and actual simulated **ISC** values for a DBM with 2 hidden layers if we fix the total number of hidden units ($m_1 + m_2 = 10$) and vary the ratio $\beta = \frac{m_2}{m_1}$. It can be seen that both theoretical prediction of **ISC** and actual simulation results are closely aligned.

7.2. DBM design under budget on parameters

To validate the claim made in Corollary 3 we considered training a DBM with two hidden layers on the MNIST dataset. For this dataset, the standard architecture for a two hidden layer DBM uses $m_1 = 500, m_2 = 1000$ hidden units (784x500x1000) (Salakhutdinov & Hinton, 2009; Salakhutdinov & Larochelle, 2010; Hinton & Salakhutdinov, 2012). In this case $\alpha_1 = 0.64, \alpha_2 = 1.27$ and the number of parameters $= 784 \times 500 + 500 \times 1000 + (784 + 500 + 1000) = 894284$. Under a budget of fixed number of parameters Corollary 3 suggests a better split of the number of hidden units. Accordingly we trained a DBM, with architecture of 784x945x161(*Recommended*), with 894915 parameters. We note that the number of parameters are similar to the standard architecture of 784x500x1000 (*Classical*), with 894284 parameters.

We used the standard metric average log-likelihood of test data (Salakhutdinov & Hinton, 2009; Salakhutdinov & Larochelle, 2010) as the measure to compare. To estimate the models partition function we used 20,000 β_k spaced uniformly from 0 to 1.0.

Discussion. The classical *tuned* architecture for training a DBM with 2 hidden layers for the original MNIST dataset gives a log-likelihood of *-84.62*. Using our recommended architecture, we were able to get a matched log-likelihood of *-84.29* without significant tuning.

7.3. Wide single layer RBM vs *lean* two-layered DBM

To verify our claim in Corollary 4 we chose single layer RBMs with $n = 20$ and $n = 30$ and varying $\alpha = \frac{m}{n} \in \{3, 7, 10, 15\}$. We initialized weights and biases of each RBM architecture randomly and then performed gibbs sampling for 5000 steps to generate a synthetic dataset of 60,000 points. The same dataset was then used for training and evaluating corresponding multilayer DBM architecture suggested by our formula. The resulting test-set log likelihood are depicted in Figure 4.
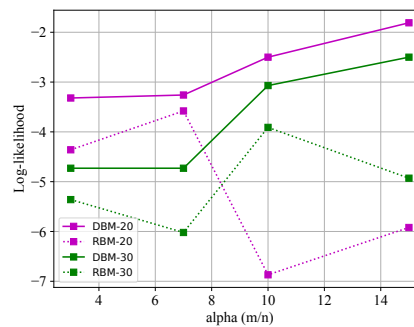

Figure 4. Comparison of test set log-likelihood attained for single layer RBM and two-layer DBM for $n = 20$ and $n = 30$. It can be seen that the DBM with much less parameters gives atleast as good log-likelihood as RBM.

Discussion. We can see that optimal DBM architecture gives same or improved log-likelihood despite the fact that it has less number of parameters than the respective single layer RBM, thus justifying our claim.

8. Conclusion

We studied the IS formalism, first introduced to study Spin glasses, to understand the energy landscape of one and two layer DBMs and proposed **ISC**, a measure of representation power of RBMs. **ISC** makes practical suggestions such as whenever number of hidden units $m > 20n$, the **ISC** saturates and multilayering should be considered. Also, **ISC** suggests alternative two layer architectures to single layer RBMs which have equal or more representational power with far fewer number of parameters.

Acknowledgment

References

Auffinger, Antonio, Arous, Gérard Ben, and Černỳ, Jiří. Random matrices and complexity of spin glasses. *Communications on Pure and Applied Mathematics*, 66(2): 165–201, 2013.

Barr, Donald R and Sherrill, E Todd. Mean and variance of truncated normal distributions. *The American Statistician*, 53(4):357–361, 1999.

Biroli, Giulio and Monasson, Rémi. From inherent structures to pure states: Some simple remarks and examples. *EPL (Europhysics Letters)*, 50(2):155, 2000.

Bray, AJ and Moore, MA. Metastable states in spin glasses. *Journal of Physics C: Solid State Physics*, 13(19):L469, 1980.

Cueto, María Angélica, Morton, Jason, and Sturmfels, Bernd. Geometry of the restricted boltzmann machine. *Algebraic Methods in Statistics and Probability,(eds. M. Viana and H. Wynn), AMS, Contemporary Mathematics*, 516:135–153, 2010.

De Dominicis, C, Gabay, M, Garel, T, and Orland, Henri. White and weighted averages over solutions of thouless anderson palmer equations for the sherrington kirkpatrick spin glass. *Journal de Physique*, 41(9):923–930, 1980.

Freund, Yoav and Haussler, David. Unsupervised learning of distributions on binary vectors using two layer networks. In *Advances in neural information processing systems*, pp. 912–919, 1992.

Hinton, Geoffrey E and Salakhutdinov, Ruslan R. A better way to pretrain deep boltzmann machines. In *Advances in Neural Information Processing Systems*, pp. 2447–2455, 2012.

Le Roux, Nicolas and Bengio, Yoshua. Representational power of restricted boltzmann machines and deep belief networks. *Neural computation*, 20(6):1631–1649, 2008.

Martens, James, Chattopadhya, Arkadev, Pitassi, Toni, and Zemel, Richard. On the representational efficiency of restricted boltzmann machines. In *Advances in Neural Information Processing Systems*, pp. 2877–2885, 2013.

Montúfar, Guido. Deep narrow boltzmann machines are universal approximators. *arXiv preprint arXiv:1411.3784*, 2014.

Montufar, Guido and Ay, Nihat. Refinements of universal approximation results for deep belief networks and restricted boltzmann machines. *Neural Computation*, 23 (5):1306–1319, 2011.

Montúfar, Guido and Rauh, Johannes. Hierarchical models as marginals of hierarchical models. *International Journal of Approximate Reasoning*, 88:531–546, 2017.

Montúfar, Guido F and Morton, Jason. When does a mixture of products contain a product of mixtures? *SIAM Journal on Discrete Mathematics*, 29(1):321–347, 2015.

Montúfar, Guido F, Rauh, Johannes, and Ay, Nihat. Expressive power and approximation errors of restricted boltzmann machines. In *Advances in neural information processing systems*, pp. 415–423, 2011.

Parisi, Giorgio and Potters, Marc. Mean-field equations for spin models with orthogonal interaction matrices. *Journal of Physics A: Mathematical and General*, 28(18): 5267, 1995.

Roberts, SA. Metastable states and innocent replica theory in an ising spin glass. *Journal of Physics C: Solid State Physics*, 14(21):3015, 1981.

Salakhutdinov, Ruslan and Hinton, Geoffrey. Deep boltzmann machines. In *Artificial Intelligence and Statistics*, pp. 448–455, 2009.

Salakhutdinov, Ruslan and Larochelle, Hugo. Efficient learning of deep boltzmann machines. In *AISTATs*, volume 9, pp. 693–700, 2010.

Stein, Daniel L and Newman, Charles M. *Spin glasses and complexity*. Princeton University Press, 2013.

Stillinger, Frank H and Weber, Thomas A. Hidden structure in liquids. *Physical Review A*, 25(2):978, 1982.

Tanaka, F and Edwards, SF. Analytic theory of the ground state properties of a spin glass. i. ising spin glass. *Journal of Physics F: Metal Physics*, 10(12):2769, 1980.

van der Maaten, Laurens. Discriminative restricted boltzmann machines are universal approximators for discrete data. Technical report, Technical Report EWI-PRB TR 2011001, Delft University of Technology, 2011.