Why do we need Depth in Deep Networks: A Statistical Mechanics perspective on DBMs

Chiranjib Bhattacharyya
Department of Computer Science and Automation
Indian Institute of Science
Bangalore-560012

December 17, 2018

The success of Deep Learning poses a fundamental question on when and why one should use multi-layering. This important issue has profound practical implications and is an extremely active area of research. However, most of these investigations have been limited to Supervised Learning architectures, such as CNNs. In this post we draw attention to unsupervised Learning architectures, more specifically DBMs, which require a very different approach from Supervised Learning frameworks.
Understanding the representation power of Deep Boltzmann Machines (DBMs[1]) is key to addressing the issue of Depth. It is well known that an DBM with one hidden layer is a universal approximator [Le Roux and Bengio, 2008]: There exists an DBM with m hidden units which can approximate any input distribution with support set size k arbitrarily well if

$$m \geq k + 1. \tag{1}$$

This is a powerful result and suggests the question, why layering is necessary at all in DBMs.

Inherent structures and Complexity of Spin Glasses Empirical analysis suggests[Bansal et al., 2018] that the estimate is extremely pessimistic when used in designing networks–finding m given k. We turn to Statistical Mechanics, in particular Spin-Glasses, to derive nore accurate alternatives. To explain Spin glasses, which motivated Boltzman Machines, the notion of *Inherent Structures*(IS) was introduced in [Stillinger and Weber, 1982]. A spin-glass is described through the probability model

$$P(\mathbf{s}|W) = \frac{e^{-\frac{1}{T}E(\mathbf{s}|W)}}{Z(W)}, Z(W) = \sum_{\mathbf{s}} e^{-\frac{1}{T}E(\mathbf{s})} \tag{2}$$

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

Definition 1. (**Complexity**)[Parisi and Potters, 1995] The *Complexity* of the model described in Eqn (2) is given by

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

where K is the number of local minima of Energy function, E defined with parameter W and $\mathbf{P}$ is a prior distribution over W.

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

Inherent Structure Capacity of RBM Borrowing the notion of *Complexity*, we define a measure which relates the *representational power* of a DBM architecture to the expected number of modes under a prior distribution on the parameters. More formally,

Definition 2 (**Inherent Structure Capacity).** For an L layered DBM with $m_1, \ldots, m_L$ hidden units and n visible units we define the *Inherent Structure Capacity* (**ISC**), denoted by $C(n, m_1, \ldots, m_L)$, 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.

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

where $\mathcal{H}$ denotes the local minima of DBM with visible units clamped to a pattern $\mathbf{v}$.

ISC of an one layer RBM Usual methods of computing *Complexity*, such as Replica method, will not apply to RBM and computation of **ISC** will require new tools. In [Bansal et al., 2018] bounds were derived to approximate **ISC** for one and two layer networks. For one layer network, the bounds throw light on **ISC** as number of hidden units approach infinity. In particular,

For a single layer RBM with m hidden units, $\lim_{m \to \infty} \mathcal{C}(n, m) = \log_2 1.5 = 0.585$

This result is interesting for two reasons.

1. The value of **ISC** can be as high as one but the limiting value is far less than one. This suggests that multilayering is needed if one wants to model distribution with larger number of modes. This result thus gives a quantitative rationale of increasing the number of layers.

2. On closer inspection of the proof it emerges that the limit is quickly attained with increasing m. A value of $m \geq 10n$ gives indeed a very close approximation. This then suggests that the increase in **ISC** is minimal, if number of hidden units are more than 10 times the visible units and at this point one can consider adding more layers. This result explains why multi-layer narrow RBMs[Montúfar, 2014] should be preferred to large wide single layer networks.

ISC of a two layer RBM A pursuant question to the issue of one layer case is: does **ISC** increase if number of layers are increased to two? The method used to compute **ISC** for an one layer DBM will not directly apply to two layer DBMs and will require more refined arguments. [Bansal et al., 2018] computes a close approximation to **ISC**, in the two layer case, and show that

For an $\boldsymbol{RBM}_{n,m_1,m_2}$*, where m_1, m_2 are number of hidden units in the first and second layer of RBM, if* $\alpha_1 = \frac{m_1}{n} > \frac{1}{20}$ *and* $\alpha_2 = \frac{m_2}{n} < 20$ *then* $\mathcal{C}(n, m_1, m_2) \leq (1 + \alpha_2) \log_2(1.5)$

The result shows that for a RBM with a wide first layer and a narrow second layer, the upper bound on **ISC** increases linearly over and above the bound achieved in the single layer case.

Future Directions: The proposal of **ISC** as a measure of representational power of a DBM is interesting as it points to specific suggestions on when to use Multi-layer DBMs. However, much needs to be done in terms of deriving algorithms for computing the value of **ISC** for the general case. Most importantly, this result suggests that there is a lot to be gained from Statistical Mechanics for understanding Model Selection in Deep Networks, ISC is a small step in this exciting journey.

References

A. Bansal, A. Anand, and C. Bhattacharyya. Using inherent structures to design lean 2-layer rbms. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 452–460, 2018. URL `http://proceedings.mlr.press/v80/bansal18a.html`.

M. A. Cueto, J. Morton, and B. Sturmfels. 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.

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

J. Martens, A. Chattopadhya, T. Pitassi, and R. Zemel. On the representational efficiency of restricted boltzmann machines. In *Advances in Neural Information Processing Systems*, pages 2877–2885, 2013.

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

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

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

G. F. Montúfar, J. Rauh, and N. Ay. Expressive power and approximation errors of restricted boltzmann machines. In *Advances in neural information processing systems*, pages 415–423, 2011.

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

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

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