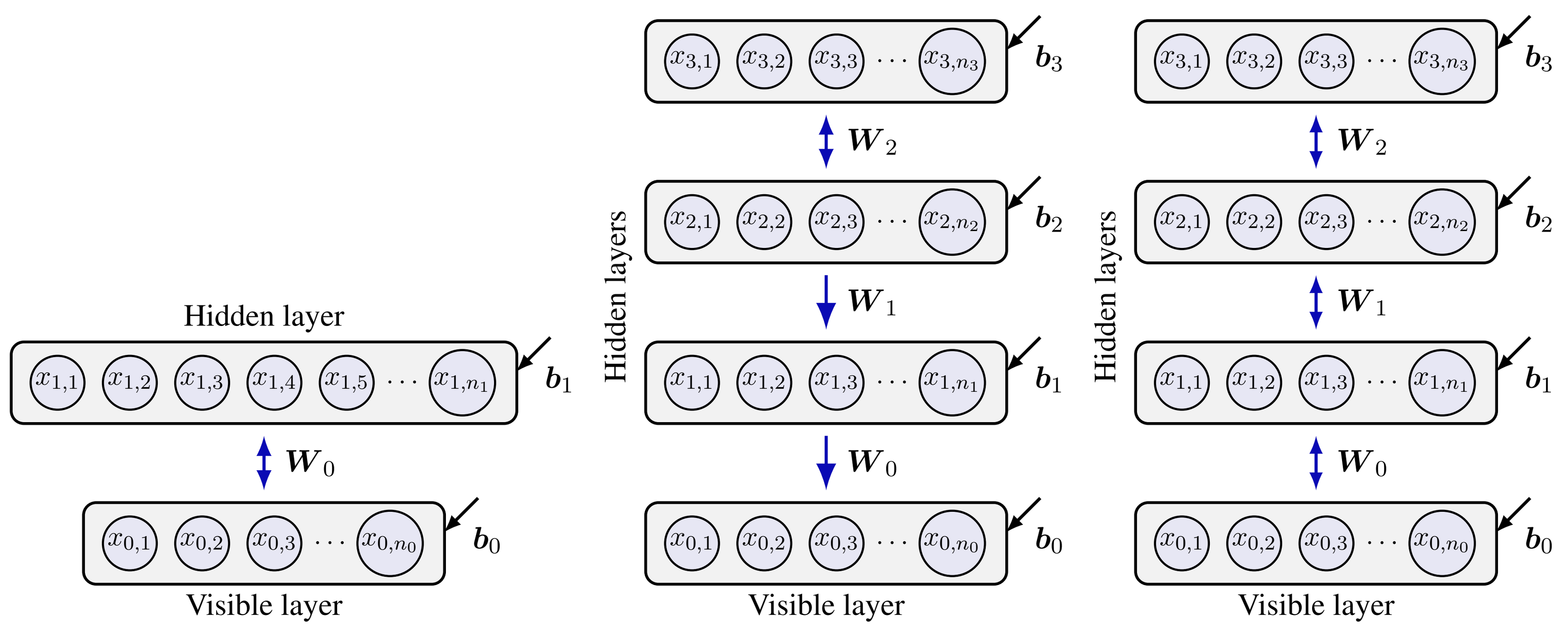
Deep Narrow Boltzmann Machines are Universal Approximators

Guido Montúfar (author)
montufar@mis.mpg.de
Max Planck Institute for Mathematics in the Sciences

KyungHyun Cho (presenter)
kyunghyun.cho@umontreal.ca
Université de Montréal

Introduction

- It is an interesting question how the representational power of **deep** artificial neural networks compares with that of **shallow** neural networks.
- Furthermore, it is interesting how the representational power of layered networks compares in the cases of **undirected** and **directed** connections.
- A **basic question** in this respect is whether a given network type can reach any degree of representation accuracy, when endowed with sufficiently many units.
- **Universal approximation** has been verified for many types of neural networks, but has remained an **open problem** for **deep narrow Boltzmann machines**.



RBM	DBN	DBM
(shallow undirected)	(deep directed)	(deep undirected)
✓ [1,5,13,8]	✓ [12,6,8,9]	✓ [this work]

Figure. Restricted Boltzmann machine (RBM), deep belief network (DBN), and deep Boltzmann machine (DBM).

Definition. A deep Boltzmann machine with n_0 visible units and L hidden layers of $n_1, \ldots, n_L$ units is a model of probability distributions of the form

$$p_{\mathbf{W},\mathbf{b}}(\mathbf{x}_0) = \sum_{\mathbf{x}_1,\ldots,\mathbf{x}_L} \frac{1}{Z(\mathbf{W},\mathbf{b})} \exp\Big(\sum_{l=0}^{L-1} \mathbf{x}_l^\top \mathbf{W}_l \mathbf{x}_{l+1} + \sum_{l=0}^{L} \mathbf{x}_l^\top \mathbf{b}_l\Big).$$

The model is **narrow**, when all layers have about the same number of units.

Overview

- At an intuitive level, undirected networks are expected to be more powerful than directed networks, since "they allow information to flow both ways."
- This intuition is not straightforward to verify. Feedforward networks can be naturally studied in a sequential way, but undirected networks are more subtle.
- We develop a method to study undirected architectures in a sequential way.

Sequential Analysis

- Express the visible probability distribution of a DBM in terms of the distributions of two smaller DBMs.

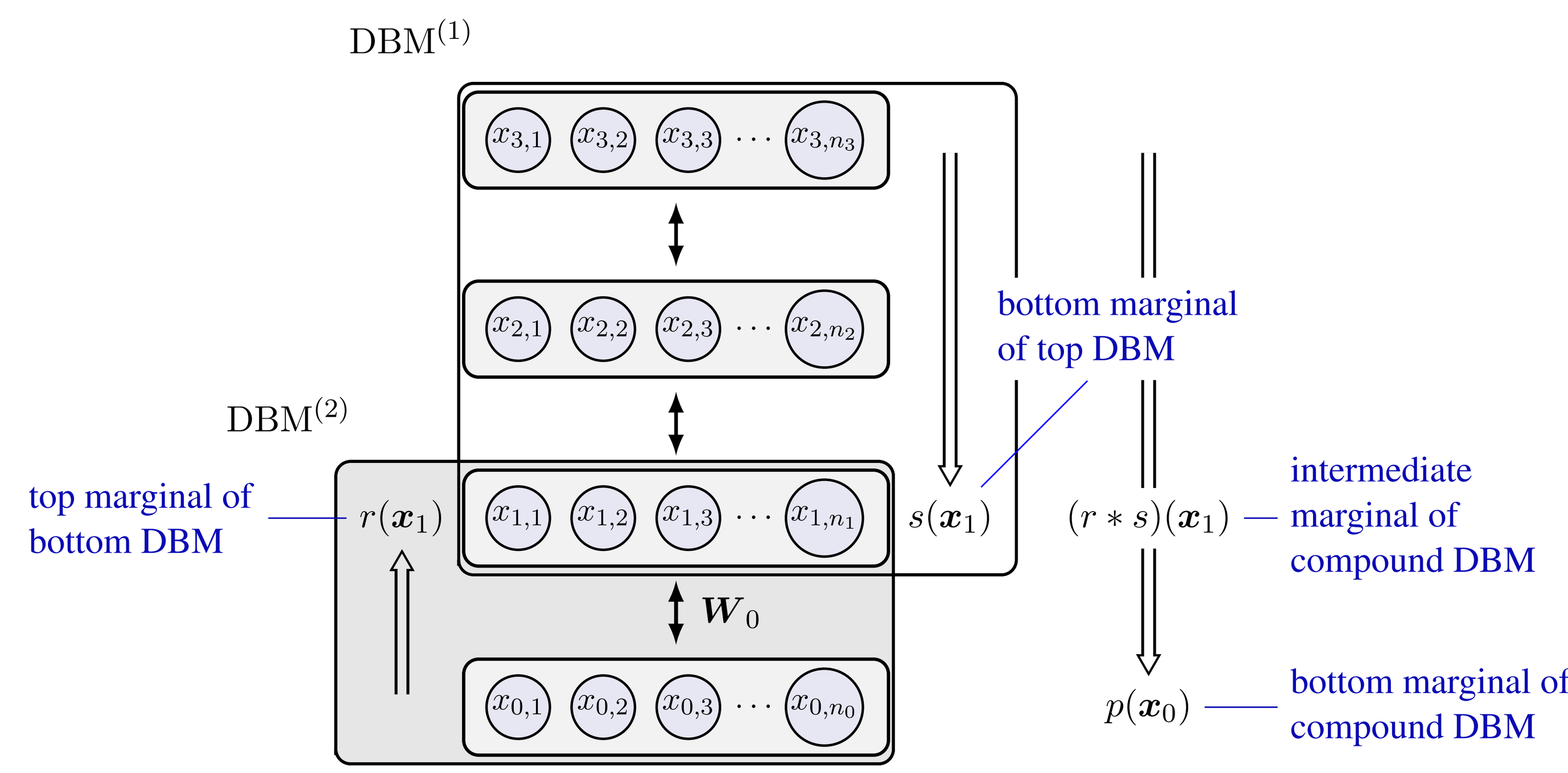


Figure. Composition of two DBMs to form a compound DBM.
Here $(r * s)$ denotes the renormalized entrywise product of r and s.

- The bottom marginal $p(\mathbf{x}_0)$ is the feedforward pass of the intermediate marginal $(r * s)(\mathbf{x}_1)$ by the feedforward map
$$q_{\mathbf{W}_0,\mathbf{b}_0}(\mathbf{x}_0|\mathbf{x}_1) = \frac{1}{Z(\mathbf{W}_0\mathbf{x}_1 + \mathbf{b}_0)} \exp(\mathbf{x}_0^\top \mathbf{W}_0 \mathbf{x}_1 + \mathbf{x}_0^\top \mathbf{b}_0).$$
- Problem: **shared parameters** of intermediate marginal and feedforward map.
- Solution: restrict attention to special marginals s from the top DBM to obtain **independent parameters** for the feedforward maps.
- In this way, with each additional layer we can transform the visible distribution by an independent feedforward map.

Universal Approximation

Theorem. A deep and narrow Boltzmann machine with a visible layer of n units and L hidden layers of n units each **is a universal approximator** of probability distributions on the states of the visible layer, provided L is large enough.

- Sufficient condition:
$$L \geq \frac{2^{n'}}{2(n' - \log_2(n') - 1)},$$
for any $n' = 2^k + k + 1 \geq n$, $k \in \mathbb{N}$.
- Necessary condition:
$$L \geq \frac{2^n - (n+1)}{n(n+1)}.$$
- For universal approximation, the first hidden layer must have at least as many units as the visible layer (minus one).
- Similar results for discriminative and multinomial models.

Conclusions

- We investigated the compositional structure of DBMs and presented a trick to separate the activities on the upper part of the network from those on the lower part of the network.
- Within certain parameter regions, deep Boltzmann machines can be studied as feedforward networks.
- We showed that deep narrow Boltzmann machines are universal approximators, and provided upper and lower bounds on the sufficient depth and width.
- In a specific sense, deep narrow Boltzmann machines are at least as powerful as narrow sigmoid belief networks and restricted Boltzmann machines.
- The methods appear valuable for studying the effects of training undirected networks sequentially, from layer to layer.

Acknowledgments

I am grateful to the Santa Fe Institute, Santa Fe, NM, USA, where I was hosted while working on this article. Many thanks to KyungHyun Cho for presenting this ICLR conference poster.

References

[1] Freund, Yoav and Haussler, David. Unsupervised learning of distributions of binary vectors using 2-layer networks. In Moody, J. E., Hanson, S. J., and Lippmann, R. (eds.), *Advances in Neural Information Processing Systems 4*, NIPS '91, pp. 912–919. Morgan Kaufmann, 1991.

[2] Bengio, Yoshua and Delalleau, Olivier. On the expressive power of deep architectures. In Kivinen, Jyrki, Szepesvári, Csaba, Ukkonen, Esko, and Zeugmann, Thomas (eds.), *Algorithmic Learning Theory*, volume 6925 of *Lecture Notes in Computer Science*, pp. 18–36. Springer Berlin Heidelberg, 2011.

[3] Cybenko, George. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4): 303–314, 1989.

[4] Hornik, Kurt, Stinchcombe, Maxwell, and White, Halbert. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359 – 366, 1989.

[5] Le Roux, Nicolas and Bengio, Yoshua. Representational power of restricted Boltzmann machines and deep belief networks. *Neural Computation*, 20(6):1631–1649, 2008.

[6] Le Roux, Nicolas and Bengio, Yoshua. Deep belief networks are compact universal approximators. *Neural Computation*, 22:2192–2207, 2010.

[7] Montavon, Grégoire, Braun, Mikio L., and Müller, Klaus-Robert. Deep Boltzmann machines as feed-forward hierarchies. In Lawrence, Neil D. and Girolami, Mark A. (eds.), *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS-12)*, volume 22, pp. 798–804, 2012.

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

[9] Montúfar, Guido. Universal approximation depth and errors of narrow belief networks with discrete units. *Neural Computation*, 26 (7):1386–1407, 2014.

[10] Salakhutdinov, Ruslan and Hinton, Geoffrey E. Deep Boltzmann machines. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 5, pp. 448–455, 2009.

[11] Salakhutdinov, Ruslan and Hinton, Geoffrey E. An efficient learning procedure for deep Boltzmann machines. *Neural Computation*, 24(8):1967–2006, August 2012.

[12] Sutskever, Ilya and Hinton, Geoffrey E. Deep narrow sigmoid belief networks are universal approximators. *Neural Computation*, 20: 2629–2636, 2008.

[13] Younes, Laurent. Synchronous Boltzmann machines can be universal approximators. *Applied Mathematics Letters*, 9(3):109 – 113, 1996.