**Max-Planck-Institut
für Mathematik
in den Naturwissenschaften
Leipzig**

Deep Narrow Boltzmann Machines are Universal Approximators

by

Guido Montúfar

Preprint no.: 113 2014

Deep Narrow Boltzmann Machines are Universal Approximators

Guido Montúfar

Max Planck Institute for Mathematics in the Sciences
Inselstrasse 22, 04103 Leipzig, Germany

November 13, 2014

Abstract

We show that deep narrow Boltzmann machines are universal approximators of probability distributions on the activities of their visible units, provided they have sufficiently many hidden layers, each containing the same number of units as the visible layer. Besides from this existence statement, we provide upper and lower bounds on the sufficient number of layers and parameters. These bounds show that deep narrow Boltzmann machines are at least as compact universal approximators as restricted Boltzmann machines and narrow sigmoid belief networks, with respect to the currently available bounds for those models.

Keywords: universal approximation property, Boltzmann machine, feedforward artificial neural network, deep learning

1 Introduction

It is an interesting question how the representational power of deep artificial neural networks, with several layers of hidden units, compares with that of shallow neural networks, with one single layer of hidden units. Furthermore, it is interesting how the representational power of layered networks compares in the cases of undirected and directed connections between the layers. A basic question in this respect is whether the classes of function approximators represented by the different network architectures can possibly reach any desired degree of accuracy, when endowed with sufficiently many computational units. This property, referred to as *universal approximation property*, has been established for a wide range of network architectures, including various kinds of shallow feedforward, shallow undirected, and deep feedforward networks, both in the deterministic and stochastic settings. Nevertheless, for several network architectures universal approximation has remained an open problem so far. In this paper we prove that deep narrow Boltzmann machines are universal approximators, provided they have sufficiently many layers of hidden units.

A Boltzmann machine (Ackley et al. 1985) is a network of stochastic binary units with undirected pairwise interactions. A deep Boltzmann machine (DBM) (Salakhutdinov and Hinton 2009) is a Boltzmann machine whose units build a stack of layers, where only pairs of units from subsequent layers interact, and only the units in the bottom layer are visible. The units within any given

layer are conditionally independent, given the states of the units in the adjacent layers. Figure 1 gives a schematic illustration of this architecture.

Since the first appearance of DBMs, a number of papers have addressed various practical and theoretical aspects of these networks, especially regarding training and estimation (see Montavon and Müller 2012; Goodfellow et al. 2013a; Cho et al. 2015). The undirected nature of DBMs leads to interesting and desirable properties, but it also brings with it challenges in training these networks and in their theoretical analysis. A number of anticipated properties of DBMs still are missing formal verification. In our main result we prove that narrow DBMs have the universal approximation property; they can approximate any probability distribution on the activations of their visible units arbitrarily well, provided they have sufficiently many hidden layers. We focus on DBMs with layers of constant size. We note that, in order to obtain the universal approximation property, the first hidden layer must have at least the same size as the visible layer (minus one, when this is even). As a direct corollary of our main theorem, we obtain the universal approximation of conditional probability distributions on the activations of subsets of visible units, given the activations of the remaining visible units. Furthermore, our analysis applies not only to the case of DBMs with binary units, but also to DBMs with softmax (finite-valued) units.

The general intuition is that undirected networks are more powerful than their directed equivalents, since "they allow information to flow both ways." Given that narrow deep belief networks (DBNs) (Hinton et al. 2006) have the universal approximation property (Sutskever and Hinton 2008), the natural expectation is that narrow DBMs also have the universal approximation property. DBNs can be regarded as the directed counterparts of DBMs. There are several reasons why this intuition is not straightforward to verify. While the computations carried out by feedforward networks can be studied in a sequential way, with the output of any given layer being the input of the next layer, in the undirected case, each internal layer receives inputs from both the previous layer and the next layer. This renders recurrent signals between all units and complicates a sequential analysis. We will show that it is possible to lever out these complicated recurrent signals and analyze DBMs in a sequential way. This way, we will show that, in some well defined sense, DBMs are at least as powerful as DBNs.

The proof exploits the compositional structure of DBMs. More precisely, we express the probability distributions represented by a given DBM in terms of the probability distributions represented by individual subparts of the network. The key component of the proof lies in showing that, within certain parameter regions (interaction weights and biases), the upper part of the network can "disable" the upward signals arriving from the lower part of the network. In such cases, the network can be regarded as operating effectively in a feedforward manner. With this, we can study the representational power of the DBM sequentially, increasing with each additional layer, similar to a deep belief network. This approach, based on disabling the upward signals, allows us to prove the universal approximation property of narrow DBMs, and it also reveals avenues for investigating the effects of the upward signals.

We note that Montavon, Braun, and Müller (2012) have also proposed a feedforward perspective on DBMs. Their motivation was different from ours, and they used the term "feedforward" to refer to a Gibbs sampling pass traversing the network in a feedforward manner, rather than to the structure of the joint probability distributions represented by the entire network. They showed, experimentally, that a DBM outputs a feedforward hierarchy of increasingly invariant representations.

In the remainder of this introduction we comment on (just a few) results that appear helpful to us for contextualizing the present paper. From the network architectures mentioned above (deep,

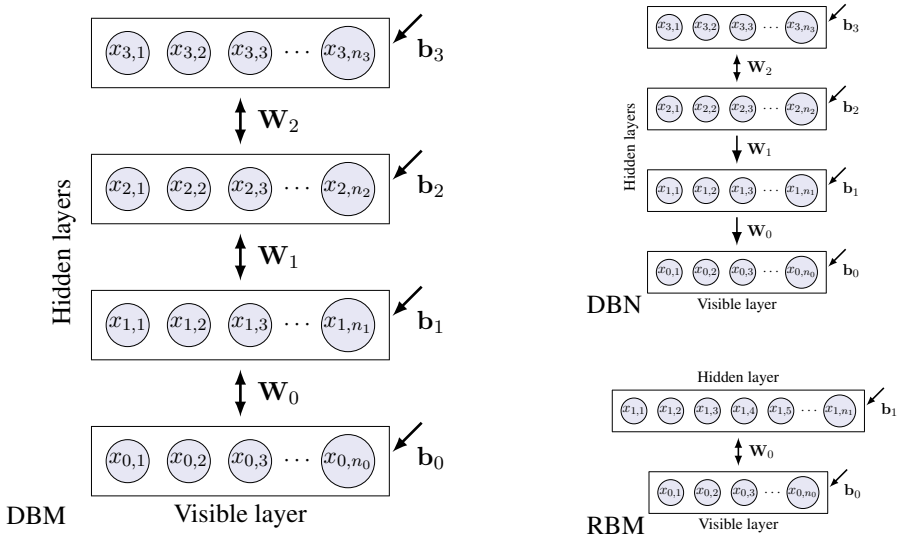

Figure 1: The left panel illustrates the architecture of a DBM. The shown DBN has a visible layer of n_0 units and three hidden layers of n_1, n_2, n_3 units. Pairs of units form consecutive layers are undirectedly connected. There are no connections between units from the same layer nor between units from non-consecutive layers. The right panel shows the architectures of a DBN and an RBM, which are the directed and the shallow versions of DBMs.

shallow, directed, undirected), presumably the most extensively studied ones are the shallow feedforward networks. A shallow feedforward network is understood as a composition of simple computational units, all having the same inputs; that is, a superposition of elementary functions defined on a common domain. For these networks it is well known that, by tuning the parameters of the individual units, they can approximate any function on the set of inputs arbitrarily well,[1] provided they have sufficiently many units (Hornik et al. 1989; Cybenko 1989). In other words, any function can be written, approximately, as a superposition (e.g., linear combination) of simple functions. This universal approximation property has been established under very general conditions both on the type of units and the type of functions being approximated (see, e.g., Leshno et al. 1993; Chen and Chen 1995). See also (Barron 1993; Burger and Neubauer 2001) for works addressing the accuracy of the approximations. An interesting recent example are shallow feedforward networks with *maxout* units (Goodfellow et al. 2013b). Besides from standard functions, i.e., deterministic output assignments given the inputs, shallow feedforward networks are also capable of approximating stochastic functions arbitrarily well, i.e., probabilistic output assignments given the inputs, when constructed with sufficiently many stochastic units. An intuitive picture is given by belief networks, where the (deterministic) state of a given unit is replaced by a probability distribution describing the likelihood of each possible state.

Deep neural networks have seen exceptional success in applications in recent years. Aiming at a better understanding and development of this success, a number of recent papers have addressed the theory of deep architectures (see Bengio and Delalleau 2011; Baldi 2012; Pascanu et al. 2014; Montúfar et al. 2014b). It is not so long ago that Sutskever and Hinton (2008) investigated deep

[1] Meant are reasonably well behaved functions and reasonable measures of approximation.

belief networks (DBNs) (Hinton et al. 2006) with narrow layers of stochastic binary units (all having about the same number of units). They showed that these architectures can approximate any binary probability distribution on the states of their visible units arbitrarily well, provided the number of hidden layers is large enough (exponentially large in the number of visible units). The minimal depth of universal approximators of this kind has been studied subsequently in more detail in (Le Roux and Bengio 2010; Montúfar and Ay 2011; Montúfar 2014). The approximation properties of DBNs with real-valued visible units and binary hidden units have been treated in recent work as well (Krause et al. 2013).

Boltzmann machines (Hinton and Sejnowski 1983; Ackley, Hinton, and Sejnowski 1985; Hinton and Sejnowski 1986) are energy based models describing the statistical behavior of pairwise interacting stochastic binary units. They have roots in statistical physics and have been studied intensively in statistics and probability theory as special types of graphical probability models and exponential families. In particular, information geometry has provided deep geometric insights about learning and approximation of probability distributions by this kind of networks (Amari et al. 1992). It is well known that Boltzmann machines are universal approximators of probability distributions over the states of their visible units, provided they have sufficiently many hidden units and there are no restrictions as for which pairs of units interact with each other (see Sussmann 1988; Younes 1996). The situation is more differentiated when a specific structure is imposed on the network, e.g., a layered structure, where only pairs of units in subsequent layers may be connected. This imposes non-trivial restrictions on the sets of representable distributions. For the shallow layered version of the Boltzmann machine, the restricted Boltzmann machine (RBM) (Smolensky 1986; Freund and Haussler 1991), the universal approximation capability has been shown in (Freund and Haussler 1991; Le Roux and Bengio 2008), provided the hidden layer is large enough (having exponentially more units than the visible layer). In fact, the proof of the universal approximation property of Boltzmann machines by Younes (1996) applies to RBMs as well. More recently, the minimal number of hidden units that is sufficient for universal approximation by RBMs and related questions have been studied in (Le Roux and Bengio 2008; Montúfar and Ay 2011; Montúfar et al. 2011; Montúfar and Morton 2013; Martens et al. 2013). Nonetheless, universal approximation results for the deep versions of RBMs, the deep Boltzmann machines (DBMs) (Salakhutdinov and Hinton 2009), have been missing so far, except when the hidden layers have exponentially many more units than the visible layer.

This paper is organized as follows. In Section 2 we provide definitions and fix notations. In Section 3 we present our main result: the universal approximation property of narrow DBMs. The proof of this result is elaborated in Sections 4 and 5. In Section 4 we address the compositional structure of DBMs. We express the probability distributions represented by a DBM in terms of the probability distributions represented by two smaller DBMs and a feedforward layer with shared parameters. In Section 5 we elaborate an approach to study DBMs from a feedforward perspective. We first present a trick to effectively disentangle the shared parameters between intermediate marginal distributions and lower conditional distributions. This is followed by a feedforward analysis proving the universal approximation property. In Section 6 we offer a discussion of the result. In the Appendix we expand on direct implications and generalizations of our main result, as well as on some possible directions for further investigations.

2 Definitions

In this section we fix notation and technical details. A layered Boltzmann machine with $L+1$ layers of $n_0, n_1, \ldots, n_L$ units is a model of joint probability distributions of the form

$$p_{\mathbf{W},\mathbf{b}}(\mathbf{x}_0, \mathbf{x}_1 \ldots, \mathbf{x}_L) = \frac{1}{Z(\mathbf{W},\mathbf{b})} \exp\Big(\sum_{l=0}^{L-1} \mathbf{x}_l^\top \mathbf{W}_l \mathbf{x}_{l+1} + \sum_{l=0}^{L} \mathbf{x}_l^\top \mathbf{b}_l\Big), \quad \text{for all } (\mathbf{x}_0, \ldots, \mathbf{x}_L) \in \{0,1\}^{n_0+\cdots+n_L}. \quad (1)$$

Here $\mathbf{x}_l = (x_{l,1}, \ldots, x_{l,n_l}) \in \{0,1\}^{n_l}$ denotes the joint state of the units in the l-th layer and $(\mathbf{x}_0, \ldots, \mathbf{x}_L) \in \{0,1\}^N$, $N = \sum_{l=0}^{L} n_l$, the joint state of all units. See Figure 1, left panel. The parameters of this model are the matrices $\mathbf{W}_l \in \mathbb{R}^{n_l \times n_{l+1}}$, $l = 0, \ldots, L-1$, of interaction weights between units from the l-th and $(l+1)$-th layers, and the vectors $\mathbf{b}_l \in \mathbb{R}^{n_l}$ of biases for the units in the l-th layer, for $l = 0, \ldots, L$. The function $Z(\mathbf{W}, \mathbf{b})$ is defined in such a way that the entries of $p_{\mathbf{W},\mathbf{b}}$ add to one, for all choices of the parameters $\mathbf{W} = (\mathbf{W}_0, \ldots, \mathbf{W}_{L-1})$ and $\mathbf{b} = (\mathbf{b}_0, \ldots, \mathbf{b}_L)$.

The set of all probability distributions of the form (1), for all possible choices of the interaction weights $\mathbf{W}$ and biases $\mathbf{b}$, is a smooth manifold, an exponential family of dimension $\sum_{l=0}^{L-1} n_l n_{l+1} + \sum_{l=0}^{L} n_l$. This manifold is embedded in the $(2^N - 1)$-dimensional set Δ_N of all possible probability distributions over $(\mathbf{x}_0, \ldots, \mathbf{x}_L) \in \{0,1\}^N$. Note that every probability distribution of the form (1) is strictly positive, meaning that it assigns strictly positive probability to every state $(\mathbf{x}_0, \ldots, \mathbf{x}_L)$. We denote this model of probability distributions by $\mathrm{DBM}_{n_0,\ldots,n_L}$, or DBM for simplicity, when $n_0, \ldots, n_L$ are clear.

The marginal probability distributions on the joint states of the units in the bottom layer are obtained by marginalizing out $\mathbf{x}_1, \ldots, \mathbf{x}_L$:

$$p_{\mathbf{W},\mathbf{b}}(\mathbf{x}_0) = \sum_{\mathbf{x}_1,\ldots,\mathbf{x}_L} p_{\mathbf{W},\mathbf{b}}(\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_L), \quad \text{for all } \mathbf{x}_0 \in \{0,1\}^{n_0}. \quad (2)$$

The set of probability distributions of this form, for all $\mathbf{W}$ and $\mathbf{b}$, is the DBM probability model with a visible layer of n_0 units and L hidden layers of $n_1, \ldots, n_L$ units. Geometrically, this set is a linear projection (marginalization) of the exponential family of distributions on the states of all layers, from the high dimensional space Δ_N to the lower dimensional space Δ_{n_0}. Note that every distribution of the form (2) is strictly positive.

In the case that the network has only one hidden layer, $L = 1$, as illustrated in the right panel of Figure 1, the model reduces to a restricted Boltzmann machine (with n_0 visible and n_1 hidden units). The corresponding set of probability distributions is denoted $\mathrm{RBM}_{n_0,n_1} \equiv \mathrm{DBM}_{n_0,n_1}$. If we replace the interactions of a DBM, except those between the top to layers, by interactions directed towards the bottom layer, we obtain a DBN. See the right panel of Figure 1 for an illustration and the Appendix for more details about RBMs and DBNs.

3 Universal Approximation

A set $\mathcal{M}$ of probability distributions on $\{0,1\}^n$ is called *universal approximator* when for any distribution q on $\{0,1\}^n$ and any $\epsilon > 0$, there is a distribution p in $\mathcal{M}$ such that $D(q\|p) \leq \epsilon$. Here

the Kullback-Leibler divergence between q and p is defined as $D(q\|p) := \sum_{\mathbf{x}} q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})}$. This is never negative and is only zero if $q = p$.

The main result of this paper is the following:

Theorem 1. *A DBM 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. More precisely, for any $n \leq n' = 2^k + k + 1$, for some $k \in \mathbb{N}$, a sufficient condition is $L \geq \frac{2^{n'}}{2(n' - \log_2(n') - 1)}$. For any n a necessary condition is $L \geq \frac{2^n - (n+1)}{n(n+1)}$.*

A direct implication of the universal approximation of probability distributions is the universal approximation of conditional probability distributions of a subset of visible units, given the states of the remaining visible units. We also note that the number of visible units (minus one) is a lower bound on the number of units in the first hidden layer of a universal approximator. See the Appendix for more details about this, and for a softmax formulation of Theorem 1.

The proof of Theorem 1 is elaborated in the next two sections. First we discuss the compositional structure of DBMs. Then we pursue a feedforward analysis leading to the universal approximation result.

4 Compositional Structure

In this section we take a look at the compositional structure of DBMs. As any other networks, DBMs are composed of simpler pieces, which are easier to analyze when taken individually. In the following we will regard a DBM as a composition of two smaller DBMs.

In order to describe these compositions, we use the renormalized entry-wise (Hadamard) product. The Hadamard product of two distributions $r, s \in \Delta_n$ is defined as

$$(r * s)(\mathbf{z}) := r(\mathbf{z}) s(\mathbf{z}) / \sum_{\mathbf{z}'} r(\mathbf{z}') s(\mathbf{z}'), \quad \text{for all } \mathbf{z} \in \{0,1\}^n. \tag{3}$$

In this definition we assume that r and s have at least one non-zero entry in common, such that $\sum_{\mathbf{z}'} r(\mathbf{z}') s(\mathbf{z}') \neq 0$. We write $r * \mathcal{M} := \{r * s \colon s \in \mathcal{M}\}$ for the set of Hadamard products of a probability distribution r and the elements of a probability model $\mathcal{M}$. The Hadamard product is a very natural operation for describing compositions of energy based models. Note that, if $r(\mathbf{z}) = \frac{1}{Z(f)} \exp(f(\mathbf{z}))$ and $s = \frac{1}{Z(g)} \exp(g(\mathbf{z}))$, then $(r * s)(\mathbf{z}) = \frac{1}{Z(f+g)} \exp(f(\mathbf{z}) + g(\mathbf{z}))$.

Now, we can write the probability distributions represented by a DBM in terms of the probability distributions represented by two smaller DBMs. More precisely, we compose $\mathrm{DBM}^{(1)}$ and $\mathrm{DBM}^{(2)}$ by identifying the bottom layer of $\mathrm{DBM}^{(1)}$ with the top layer of $\mathrm{DBM}^{(2)}$, as illustrated in Figure 2. By this composition, the distribution s that was originally represented on the states of the bottom layer of $\mathrm{DBM}^{(1)}$ becomes $r * s$, where r is the distribution that was originally represented on the states of the top layer of $\mathrm{DBM}^{(2)}$.

Proposition 2. *Consider the model $\mathrm{DBM} = \mathrm{DBM}_{n_0,\ldots,n_L}$. For any $0 < k < L$ the marginal distributions of the k-th layer's units are the distributions of the form*

$$p(\mathbf{x}_k) = (p^{(2)} * p^{(1)})(\mathbf{x}_k), \quad \textit{for all } \mathbf{x}_k \in \{0,1\}^{n_k},$$

where $p^{(1)}(\mathbf{x}_k)$ is a bottom layer marginal of $\mathrm{DBM}^{(1)} = \mathrm{DBM}_{n_k,\ldots,n_L}$ and $p^{(2)}(\mathbf{x}_k)$ is a top layer marginal of $\mathrm{DBM}^{(2)} = \mathrm{DBM}_{n_0,n_1,\ldots,n_k}$.

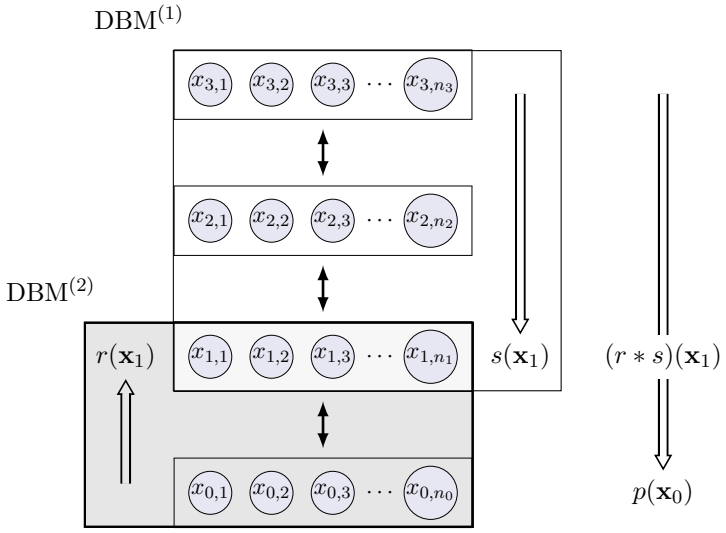

Figure 2: Composition of an upper and a lower DBM to form a larger DBM.

Proof of Proposition 2. We have

$$
\begin{aligned}
p(\mathbf{x}_k) &= \sum_{\mathbf{x}_0,\ldots,\mathbf{x}_{k-1},\mathbf{x}_{k+1},\ldots,\mathbf{x}_L} p(\mathbf{x}_0,\mathbf{x}_1,\ldots,\mathbf{x}_L) \\
&= \sum_{\mathbf{x}_0,\ldots,\mathbf{x}_{k-1},\mathbf{x}_{k+1},\ldots,\mathbf{x}_L} \frac{1}{Z(\mathbf{W},\mathbf{b})} \exp\Big(\sum_{l=0}^{L-1} \mathbf{x}_l^\top \mathbf{W}_l \mathbf{x}_{l+1} + \sum_{l=0}^{L} \mathbf{x}_l^\top \mathbf{b}_l\Big) \\
&= \sum_{\mathbf{x}_0,\ldots,\mathbf{x}_{k-1},\mathbf{x}_{k+1},\ldots,\mathbf{x}_L} \frac{1}{Z(\mathbf{W},\mathbf{b})} \exp\Big(\sum_{l=0}^{k-1} \mathbf{x}_l^\top \mathbf{W}_l \mathbf{x}_{l+1} + \sum_{l=0}^{k-1} \mathbf{x}_l^\top \mathbf{b}_l + \mathbf{x}_k^\top \mathbf{b}'_k\Big) \\
&\qquad\qquad \times \exp\Big(\sum_{l=k}^{L-1} \mathbf{x}_l^\top \mathbf{W}_l \mathbf{x}_{l+1} + \sum_{l=k}^{L} \mathbf{x}_l^\top \mathbf{b}_l - \mathbf{x}_k^\top \mathbf{b}'_k\Big) \\
&= \frac{1}{Z(\mathbf{W},\mathbf{b})} \sum_{\mathbf{x}_0,\ldots,\mathbf{x}_{k-1}} \exp\Big(\sum_{l=0}^{k-1} \mathbf{x}_l^\top \mathbf{W}_l \mathbf{x}_{l+1} + \sum_{l=0}^{k-1} \mathbf{x}_l^\top \mathbf{b}_l + \mathbf{x}_k^\top \mathbf{b}'_k\Big) \\
&\qquad \times \sum_{\mathbf{x}_{k+1},\ldots,\mathbf{x}_L} \exp\Big(\sum_{l=k}^{L-1} \mathbf{x}_l^\top \mathbf{W}_l \mathbf{x}_{l+1} + \sum_{l=k}^{L} \mathbf{x}_l^\top \mathbf{b}_l - \mathbf{x}_k^\top \mathbf{b}'_k\Big) \\
&= \frac{1}{Z(\mathbf{W},\mathbf{b})}\, Z(\mathbf{W}^{(2)},\mathbf{b}^{(2)}) p^{(2)}(\mathbf{x}_k)\; Z(\mathbf{W}^{(1)},\mathbf{b}^{(1)}) p^{(1)}(\mathbf{x}_k), \quad \text{for all } \mathbf{x}_k \in \{0,1\}^{n_k}.
\end{aligned}
$$

This shows that for any marginal $p(\mathbf{x}_k)$ representable by the compound DBM, there is a distribution $p^{(2)}(\mathbf{x}_k)$ representable as the top layer marginal of DBM$^{(2)}$ with parameters $\mathbf{W}^{(2)} = (\mathbf{W}_0,\ldots,\mathbf{W}_{k-1})$, $\mathbf{b}^{(2)} = (\mathbf{b}_0,\ldots,\mathbf{b}_{k-1},\mathbf{b}'_k)$, and a distribution $p^{(1)}(\mathbf{x}_k)$ representable as the bottom layer marginal of DBM$^{(1)}$ with parameters $\mathbf{W}^{(1)} = (\mathbf{W}_k,\ldots,\mathbf{W}_{L-1})$, $\mathbf{b}^{(1)} = (\mathbf{b}_k - \mathbf{b}'_k, \mathbf{b}_{k+1},\ldots,\mathbf{b}_L)$, such that the equation $p(\mathbf{x}_k) = (p^{(2)} * p^{(1)})(\mathbf{x}_k)$ holds, and vice versa. □

Next we define feedforward layers, as we will use them in our analysis. The feedforward layer with n_1 input and n_0 output units, denoted FF_{n_0,n_1}, is the model of conditional probability distributions of the form

$$q_{\mathbf{W}_0,\mathbf{b}_0}(\mathbf{x}_0|\mathbf{x}_1) = \frac{1}{Z(\mathbf{W}_0\mathbf{x}_1+\mathbf{b}_0)} \exp(\mathbf{x}_0^\top \mathbf{W}_0\mathbf{x}_1 + \mathbf{x}_0^\top \mathbf{b}_0), \quad \text{for all } \mathbf{x}_0 \in \{0,1\}^{n_0}, \mathbf{x}_1 \in \{0,1\}^{n_1}. \tag{4}$$

Here $\mathbf{W}_0 \in \mathbb{R}^{n_0 \times n_1}$ is a matrix of input weights and $\mathbf{b}_0 \in \mathbb{R}^{n_0}$ is a vector of biases. Clearly, these conditionals correspond exactly to the conditionals represented between first hidden layer and the visible layer of a DBM, for the same choices of parameters.

The next Proposition 3 gives an expression for the probability distributions represented by a DBM in terms of the probability distributions represented by two smaller DBMs and the conditionals represented by a feedforward layer with shared parameters.

Proposition 3. *The probability distributions representable by* $\mathrm{DBM}_{n_0,\dots,n_L}$ *are those of the from*

$$p(\mathbf{x}_0) = \sum_{\mathbf{x}_1} q(\mathbf{x}_0|\mathbf{x}_1)(r*s)(\mathbf{x}_1), \quad \textit{for all } \mathbf{x}_0 \in \{0,1\}^{n_0},$$

where $u(\mathbf{x}_0,\mathbf{x}_1) = q(\mathbf{x}_0|\mathbf{x}_1)r(\mathbf{x}_1)$ *is a joint probability distribution of the fully observable* RBM_{n_0,n_1} *and* s *is a bottom layer marginal of* $\mathrm{DBM}_{n_1,\dots,n_L}$.

Proof of Proposition 3. We have

$$p(\mathbf{x}_0) = \sum_{\mathbf{x}_1} p(\mathbf{x}_0|\mathbf{x}_1)p(\mathbf{x}_1), \quad \text{for all } \mathbf{x}_0 \in \{0,1\}^{n_0}.$$

By Proposition 2, $p(\mathbf{x}_1) = (r*s)(\mathbf{x}_1)$ for all $\mathbf{x}_1 \in \{0,1\}^{n_1}$. □

The proposition is illustrated in Figure 2. Note that $r(\mathbf{x}_1)$ is a top layer marginal of RBM_{n_0,n_1} and the conditional $q(\mathbf{x}_0|\mathbf{x}_1)$ is the top-to-bottom conditional of RBM_{n_0,n_1}, corresponding to the feedforward layer FF_{n_0,n_1}. Proposition 3 suggests that it is possible to study the representational power of DBMs in terms of the representational power of smaller DBMs composed with simple feedforward networks. The problem is that the distribution $r*s$, intended as the input of the feedforward layer, depends on the same parameters $\mathbf{W}_0, \mathbf{b}_0$ as the feedforward layer. Hence the input cannot be chosen independently from the transformation that the feedforward layer applies on it. Nonetheless, as we will show in the next section, it is possible to resolve this difficulty and analyze the representational power of the DBM in a sequential way.

5 Feedforward Analysis

Here discuss the possibility of viewing DBMs as feedforward structures. Consider a DBM composed of an upper and a lower part, as shown in Figure 2. If the upper $\mathrm{DBM}^{(1)}$ is able to "disable" or neutralize the top layer marginal r of $\mathrm{DBM}^{(2)}$, then the distribution represented at the bottom layer of the compound DBM can be regarded as the feedforward pass of the distribution s represented at the bottom layer of $\mathrm{DBM}^{(1)}$. Namely, by Proposition 3 the visible distribution of the combined network is the result of passing the marginal distribution $(r*s)(\mathbf{x}_1)$ feedforward through the conditional distribution $q(\mathbf{x}_0|\mathbf{x}_1)$.

5.1 Disabling the backward signal

In order to make the approach work, we have to deal with the problem that the marginal r and the conditional q share the same parameters. When we modify these parameters in order to obtain a specific conditional q (representing a desired feedforward transformation of the input), the marginal r changes as well, and with it also the input $r * s$. We resolve this dilemma in the following way. Instead of regarding $\mathrm{DBM}^{(1)}$ as the input model, we restrict our attention to a subset of possible input distributions $\mathcal{G} \subseteq \mathrm{DBM}^{(1)}$ with the following property:

$$r * \mathcal{G} = \mathcal{G} \quad \text{for all top layer marginals } r \text{ of } \mathrm{DBM}^{(2)}. \tag{5}$$

In this case, any desired input $s \in \mathcal{G}$, together with any desired conditional $q \in \mathrm{FF}_{n_0,n_1}$, can be obtained by the following procedure:

1. Tune the parameters of $\mathrm{DBM}^{(2)}$ to represent any desired (representable) conditional distribution q. By tuning the parameters in this way, the top layer marginal of $\mathrm{DBM}^{(2)}$ becomes a distribution r that depends on q.

2. Tune the parameters of $\mathrm{DBM}^{(1)}$ to represent a bottom layer marginal $s' \in \mathcal{G}$ with $r * s' = s$.

Now we just need to find a good choice of $\mathcal{G}$, from which we require the following.

- The set $\mathcal{G}$ has to satisfy (5).
- We have to make sure that $\mathcal{G}$ is contained in, or can be approximated arbitrarily well, by the distributions representable at the bottom layer of $\mathrm{DBM}^{(1)}$.
- Furthermore, $\mathcal{G}$ should be as large as possible, in order to account for the largest possible fraction of the representational power of $\mathrm{DBM}^{(1)}$.

It is not easy to specify the top layer marginals of $\mathrm{DBM}^{(2)}$ appearing in (5). However, at this point we can impose a stronger condition on $\mathcal{G}$ and require that $r * \mathcal{G} = \mathcal{G}$ hold for all strictly positive distributions r, in which case it automatically holds for all top layer marginals of $\mathrm{DBM}^{(2)}$. We choose $\mathcal{G}$ as the set of probability distributions on $\{0,1\}^{n_1}$ that assign positive probability only to a subset of vectors $S \subset \{0,1\}^{n_1}$, i.e., as the set

$$\Delta_{n_1}(S) := \{p \in \Delta_{n_1} : p(\mathbf{x}_1) = 0 \text{ for all } \mathbf{x}_1 \notin S\}. \tag{6}$$

In the next Proposition 4 we show that this set satisfies (5), regardless of S. In order to satisfy the second and third items of the list, we have to choose S depending on the size of $\mathrm{DBM}^{(1)}$. We will discuss the details of this further below, in Section 5.2.

Given a set of probability distributions $\mathcal{M} \subseteq \Delta_n$, let $\overline{\mathcal{M}} \subseteq \Delta_n$ denote the set of probability distributions that can be approximated arbitrarily well by elements from $\mathcal{M}$.

Proposition 4. *Let $r \in \Delta_n$ be a strictly positive probability distribution and let $\mathcal{M} \subseteq \Delta_n$ be a set of probability distributions with $\overline{\mathcal{M}} \supseteq \Delta_n(S)$. Then $\overline{r * \mathcal{M}} \supseteq \Delta_n(S)$.*

Proof of Proposition 4. The argument is simple: since $\mathcal{M}$ can approximate any distribution supported on S arbitrarily well, it can approximate any distribution of the form $s'(\mathbf{z}) = (s/r)(\mathbf{z}) :=$

$(s(\mathbf{z})/r(\mathbf{z}))\frac{1}{\sum_{\mathbf{z}'} s(\mathbf{z}')/r(\mathbf{z}')}$, $\mathbf{z} \in \{0,1\}^n$, arbitrarily well, where s is any distribution strictly supported on S. Note that any such s' is strictly supported on S, i.e., it is contained in $\Delta_n(S)$. Now, the Hadamard product of r and s' is given by

$$\begin{aligned}
(r * s')(\mathbf{z}) &= r(\mathbf{z})s'(\mathbf{z})\frac{1}{\sum_{\mathbf{z}'} r(\mathbf{z}')s'(\mathbf{z}')} \\
&= r(\mathbf{z})(s(\mathbf{z})/r(\mathbf{z}))\frac{1}{\sum_{\mathbf{z}''} s(\mathbf{z}'')/r(\mathbf{z}'')}\frac{1}{\sum_{\mathbf{z}'} r(\mathbf{z}')s'(\mathbf{z}')} \\
&= s(\mathbf{z})\frac{1}{\sum_{\mathbf{z}''} s(\mathbf{z}'')/r(\mathbf{z}'')}\frac{1}{\sum_{\mathbf{z}'} r(\mathbf{z}')(s(\mathbf{z}')/r(\mathbf{z}'))\frac{1}{\sum_{\mathbf{z}'''} s(\mathbf{z}''')/r(\mathbf{z}''')}} \\
&= s(\mathbf{z})\frac{1}{\sum_{\mathbf{z}''} s(\mathbf{z}'')/r(\mathbf{z}'')}\frac{\sum_{\mathbf{z}'''} s(\mathbf{z}''')/r(\mathbf{z}''')}{\sum_{\mathbf{z}'} s(\mathbf{z}')} \\
&= s(\mathbf{z}), \quad \text{for all } \mathbf{z} \in \{0,1\}^n.
\end{aligned}$$

Since s was an arbitrary distribution from the set $\Delta_n(S)$, this proves the claim. □

5.2 Proof of Theorem 1

In the previous subsection we have shown that, within certain parameter regimes, DBMs can be regarded as a directed models. Let us make this more explicit. Putting Propositions 3 and 4 together, we arrive at:

Proposition 5. *If* $\mathrm{DBM}_{n_1,\ldots,n_L}$ *can approximate every distribution from the set* $\Delta_{n_1}(S)$ *arbitrarily well as its bottom layer marginal, then* $\mathrm{DBM}_{n_0,n_1,\ldots,n_L}$ *can approximate every distribution from the set* $\mathrm{FF}_{n_0,n_1}(\Delta_{n_1}(S))$ *arbitrarily well as its bottom layer marginal.*

With this proposition, we can study the representational power of DBMs sequentially, from layer to layer. A feedforward layer is able to compute many interesting transformations of its input. For any choice of parameters, the conditional distribution $q_{\mathbf{W}_0,\mathbf{b}_0}$ represented by the feedforward layer FF_{n_0,n_1} defines a map $\Delta_{n_1} \to \Delta_{n_0}$ taking a probability distribution p to a probability distribution $\sum_{\mathbf{x}_1} p(\mathbf{x}_1)q_{\mathbf{W}_0,\mathbf{b}_0}(\mathbf{x}_0|\mathbf{x}_1)$. As we vary the parameters $\mathbf{W}_0, \mathbf{b}_0$, every input distribution p is mapped to a collection of output distributions. Hence the feedforward layer can augment the representational power of the input model. After a sufficient number of feedforward layers, the output distribution can be made to approximate any desired probability distribution arbitrarily well.

We focus on the DBM with layers of constant size n. First, we need to show that a DBM with n visible units and l hidden layers of n units each can approximate any distribution from $\Delta_n(S^l)$ arbitrarily well, for some $S^l \subseteq \{0,1\}^n$. Then, we need to show that by transformations with a feedforward layer, we can obtain a larger set $\Delta(S^{l+1}) \subseteq \mathrm{FF}_{n,n}(\Delta(S^l))$, which in turn can be approximated arbitrarily well by the DBM with $l+1$ hidden layers. The idea is that, by successive transformations with feedforward layers, we will obtain an increasing sequence

$$S^1 \subset S^2 \subset S^3 \subset \cdots \subset S^L = \{0,1\}^n, \tag{7}$$

meaning that the DBM with L hidden layers can approximate any distribution on $S^L = \{0,1\}^n$ arbitrarily well.

We start with $l = 1$. The representational power of RBMs (DBMs with one single hidden layer) has been studied in previous papers. We take the following Proposition 6 from (Montúfar and Ay

2011). We call a pair of states $\mathbf{x}, \mathbf{x}' \in \{0,1\}^n$ *adjacent* if the Hamming distance between the them is one, i.e., $d_H(\mathbf{x}, \mathbf{x}') := |\{i \in [n] \colon x_i \neq x'_i\}| = 1$.

Proposition 6. *The model RBM_{n_0,n_1} can approximate every distribution from $\Delta_{n_0}(S)$ arbitrarily well as its bottom layer marginal, where $S \subseteq \{0,1\}^{n_0}$ is any union of $n_1 + 1$ pairs of adjacent states.*

As mentioned in the introduction, directed networks have been studied in previous papers and we can take advantage of the tools that have been developed there. The following Proposition 7 (taken from Montúfar 2014) describes the transformations of an input set $\Delta_n(S)$ by a feedforward layer to produce an augmented set $\Delta_n(S \cup P)$ as output. The *flip* of a state vector $\mathbf{x}$ along j is the vector $\mathbf{x}_{\bar{j}}$ that results from inverting the j-th entry of $\mathbf{x}$.

Proposition 7. *The image of $\Delta_n(S)$ by $\mathrm{FF}_{n,n}$ can approximate every distribution from $\Delta_n(S \cup P)$ arbitrarily well, where $P \subseteq \{0,1\}^n$ is any set constructible by the following procedure. Take n disjoint pairs of adjacent states $p^1, \dots, p^n$ and n distinct directions $i_1, \dots, i_n$. Intersect each pair p^j with S and flip the result along the direction i_j, to obtain $\bar{p}^1 = (S \cap p^1)_{\bar{i}_1}, \dots, \bar{p}^n = (S \cap p^n)_{\bar{i}_n}$. Set $P = \{\bar{p}^1, \dots, \bar{p}^n\}$.*

Montúfar and Ay (2011) show that, for any $k \in \mathbb{N}$ and $n = 2^k + k + 1$, there is a choice of S^1 of the form described in Proposition 6 and a sequence of augmentations $S^2 = S^1 \cup P^1, \dots, S^L = S^{L-1} \cup P^{L-1}$ of the form described in Proposition 7, such that $S^L = \{0,1\}^n$ for $L = \frac{2^{n-1}}{2^k}$. This implies the existence and sufficiency statements from Theorem 1. The necessary condition results from straightforward parameter counting arguments; from comparing the dimension $\dim(\Delta_n) = 2^n - 1$ of the set being approximated and the number of parameters $Ln^2 + (L+1)n$ of the DBM. This concludes the proof of Theorem 1.

6 Conclusion

This paper proves that deep and narrow Boltzmann machines are universal approximators of probability distributions on the states of their visible units, provided they have sufficiently many layers of hidden units. Thereby, this paper settles an intuition that had been missing formal verification. This universal approximation result complements previous results addressing restricted Boltzmann machines and deep narrow sigmoid belief networks, which can be regarded the shallow and feedforward counterparts of deep narrow Boltzmann machines. Further, the presented analysis yields upper and lower bounds on the minimal number of layers and parameters of narrow DBM universal approximators. These bounds show that narrow DBMs are at least as compact universal approximators as RBMs and narrow DBNs are known to be.

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. This allowed us to trace parameter regions where DBMs can be regarded as operating in a feedforward manner, passing the probability distributions represented at the higher layers downwards from layer to layer by multiplication with conditional probability distributions. This feedforward-like behavior can be obtained when the upper part of the network is able to represent top-down distributions that neutralize the bottom-up distributions represented by the lower part of the network.

The feedforward perspective on DBMs allowed us to study their representational power sequentially, increasing from layer to layer, like DBNs, and finally prove the universal approximation

property. As a byproduct of this analysis, we obtain a picture of the classes of distributions that can be represented by both DBMs and DBNs. Our analysis also exposes a compositional structure that can be used to study the recurrent signals in DBMs, an interesting topic for future work.

There are several direct implications from our analysis, including the universal approximation of stochastic maps and the universal approximation property for DBMs with softmax units. We formulate these results explicitly in the Appendix. We provide a more detailed discussion and comparison of our results with previous results for RBMs and DBNs in the Appendix.

Appendix

Approximation of stochastic maps

A DBM can be used to define stochastic input-output relations. A stochastic map with inputs $\{0,1\}^k$ and outputs $\{0,1\}^m$ assigns a probability distribution $p(\cdot|\mathbf{i}) \in \Delta_m$ to each input vector $\mathbf{i} \in \{0,1\}^k$. DBMs define such maps by clamping the states of some of their units to the input values $\mathbf{i}$, and taking the resulting conditional probability distribution over the states of some other units as the output distributions. One way of doing this is by dividing the visible units in two groups, corresponding to inputs and outputs, as $\mathbf{x}_0 = (\mathbf{i}, \mathbf{o})$. Given that $p(\mathbf{x}_0) = p(\mathbf{i}, \mathbf{o})$ stands in one to one relation to the pair $(p(\mathbf{i}), p(\mathbf{o}|\mathbf{i}))$, Theorem 1 implies:

Corollary 8. *A DBM with a visible layer of $n = k + m$ units and L hidden layers of n units each is a universal approximator of stochastic input-output maps with $\mathbf{i} = (x_{0,1}, \dots, x_{0,k})$ and $\mathbf{o} = (x_{0,k+1}, \dots, x_{0,k+m})$, provided L is large enough.*

Note that a universal approximator of stochastic maps is also a universal approximator of deterministic maps. This is because every deterministic map $\mathbf{i} \mapsto \mathbf{o} = f(\mathbf{i})$ can be regarded as the special type of stochastic map $\mathbf{i} \mapsto \delta_{f(\mathbf{i})}(\mathbf{o})$, where $\delta_{f(\mathbf{i})}$ is the Dirac delta assigning probability one to $\mathbf{o} = f(\mathbf{i})$.

Corollary 8 complements previous results addressing universal approximation of stochastic maps by conditional RBMs (van der Maaten 2011; Montúfar et al. 2014a). As discussed in (Montúfar et al. 2014a), in contrast to joint probability distributions, stochastic maps do not need to model the input distributions, and hence universal approximators of stochastic maps need not be universal approximators of joint probability distributions. It would be interesting to investigate corresponding refinements of Corollary 8 in future work.

Softmax units

All arguments presented in the main part of this article hold for arbitrary finite valued units (not only binary units). An analysis of sequences of feedforward layers of finite-valued units is available from (Montúfar 2014). This allows us to formulate the following generalization of Theorem 1:

Theorem 9. *A DBM with a visible layer of n softmax q-valued units and L hidden layers of n softmax q-valued units each is a universal approximator of probability distributions on the states of the visible layer, provided L is large enough. More precisely, for any $n \leq n' = q^k + k + 1$, for some $k \in \mathbb{N}$, a sufficient condition is $L \geq 1 + \frac{q^{n'}-1}{q(q-1)(n'-\log_q(n'))-1)}$. For any n a necessary condition is $L \geq \frac{q^n-1}{n(q-1)(n(q-1)+2)}$.*

This result can be further refined to cases where each layer has units with different numbers of possible states. We omit further details at this point.

Minimal width of universal approximators

In a layered network, a too narrow layer represents a bottleneck. It is an interesting question how narrow a universal approximator can be. For example, if the visible layer has n_0 units, the first hidden layer of a universal approximator must have at least $n_1 \geq n_0 - 1$ units. In fact, when n_0 is odd, this has to be at least $n_1 \geq n_0$.

Proposition 10. *A DBM with n_0 visible units can be a universal approximator only if the first hidden layer contains at least $n_1 \geq n_0 - 1$ units, when n_0 is even, and at least $n_1 \geq n_0$ units, when n_0 is odd.*

Proof of Proposition 10. This follows from the fact that the visible distributions of the DBM are mixtures of the conditionals $p(\mathbf{x}_0|\mathbf{x}_1)$, for all $\mathbf{x}_1 \in \{0,1\}^{n_1}$. Each of these conditional distributions is a product distribution. There are distributions on $\{0,1\}^{n_0}$ that can only be approximated by mixtures of product distributions, if these mixtures involve mixture components that approximate all point measures assigning probability one to the binary strings with an odd number of ones (see Montúfar 2013).

Now, Montúfar and Morton (2014; Proposition 3.19) show that when n_0 is odd, there is no $(n_0 - 1)$-generated zonoset with a point in each odd (or each even) orthant of $\mathbb{R}^{n_0}$. Without going into more details, this implies that, when $n_1 = n_0 - 1$, with odd n_0, the set of conditionals $\{p(\mathbf{x}_0|\mathbf{x}_1) \colon \mathbf{x}_1 \in \{0,1\}^{n_1}\}$ cannot approximate the set of point measures that assign probability one to the binary strings with an odd (or even) number of ones. □

We note that the same width bound holds for DBNs, since the visible distributions represented by DBNs are mixtures of the same product distributions as the visible distributions of DBMs.

Comparison with narrow DBNs

DBNs have the same network topology as DBMs, but with interactions directed towards the bottom layer, except for the interactions between the deepest two layers, which are undirected. The corresponding joint probability distributions have the form

$$p_{\mathbf{W},\mathbf{b}}(\mathbf{x}_0, \mathbf{x}_1 \ldots, \mathbf{x}_L) = p_{\mathbf{W}_{L-1},\mathbf{b}_{L-1},\mathbf{b}_L}(\mathbf{x}_{L-1}, \mathbf{x}_L) \prod_{l=0}^{L-2} p_{\mathbf{W}_l,\mathbf{b}_l}(\mathbf{x}_l|\mathbf{x}_{l+1}),$$
$$\text{for all } (\mathbf{x}_0, \ldots, \mathbf{x}_L) \in \{0,1\}^{n_0+\cdots+n_L}. \quad (8)$$

Here the distributions of the states in the deepest two layers are given by

$$p_{\mathbf{W}_{L-1},\mathbf{b}_{L-1},\mathbf{b}_L}(\mathbf{x}_{L-1}, \mathbf{x}_L) = \frac{1}{Z(\mathbf{W}_{L-1}, \mathbf{b}_{L-1}, \mathbf{b}_L)} \exp(\mathbf{x}_{L-1}^\top \mathbf{W}_{L-1} \mathbf{x}_L + \mathbf{x}_{L-1}^\top \mathbf{b}_{L-1} + \mathbf{x}_L^\top \mathbf{b}_L),$$
$$\text{for all } (\mathbf{x}_{L-1}, \mathbf{x}_L) \in \{0,1\}^{n_{L-1}+n_L}. \quad (9)$$

The conditional distributions (feedforward layers), are given by

$$p_{\mathbf{W}_l,\mathbf{b}_l}(\mathbf{x}_l|\mathbf{x}_{l+1}) = \frac{1}{Z(\mathbf{W}_l\mathbf{x}_{l+1},\mathbf{b}_l)}\exp(\mathbf{x}_l^\top\mathbf{W}_l\mathbf{x}_L + \mathbf{x}_l^\top\mathbf{b}_l),$$
$$\text{for all } \mathbf{x}_l \in \{0,1\}^{n_l}, \text{ for all } \mathbf{x}_{l+1} \in \{0,1\}^{n_{l+1}}. \quad (10)$$

Although DBNs have undirected interactions between the top two layers, in the narrow case the universal approximation capability stems essentially from the feedforward part. A DBN with layers of width n is a universal approximator if the number of hidden layers satisfies $L \geq \frac{2^n}{2(n-\log_2(n)-1)}$ and only if $L \geq \frac{2^n-(n+1)}{n(n+1)}$ (Montúfar and Ay 2011). These bounds correspond exactly to the bounds we obtained in Theorem 1 for DBMs. In our proof we showed that the kinds of transformations of probability distributions exploited in (Montúfar and Ay 2011) in the context of DBNs can also be represented by DBMs. In particular, our analysis shows that many distributions that are representable by DBNs are also representable by DBMs of the same size.

Comparison with RBMs

In the case of one single hidden layer, the DBM reduces to an RBM. RBMs are universal approximators, provided the hidden layer contains sufficiently many units. The minimal number of hidden units m for which an RBM with n visible units is a universal approximator is at least $\frac{2^n-n}{n+1}$ and at most $2^{n-1}-1$ (Montúfar and Ay 2011) or 2^n-n-1 (Younes 1996), whatever is smaller. The exact value is not known, but there are examples where the lower bound is not attained. For narrow DBMs we obtained an upper bound on the minimal number of layers sufficient for universal approximation of the form $L \geq 2^n/2(n-\log_2(n)-1)$. Hence both, RBMs and narrow DBMs require at most a number of interaction weights and biases of order $O(n2^{n-1})$. We should note that in both cases, it is possible to formulate restrictions on the interaction weights and biases, such that the total number of free parameters needed for universal approximation is 2^n-1, i.e., just as large as the dimension of the set Δ_n.

Exploiting the backward activity

The product $r*s$ arising in Proposition 3 can be used to augment the input model that is passed to the feedforward layer. As long as this does not interfere with the choice of a desirable conditional q, this could be exploited to obtain a more compact construction of a universal approximator. Investigating this in detail could help us better understand the differences of DBNs and DBMs. It would be interesting to take a closer look at this in future work.

Approximation errors

The proofs presented in this paper specify sets of probability distributions that can be represented by DBMs, depending on the number of hidden layers they have. When the networks are not deep enough to reach universal approximation capacity, their approximation errors can be studied in terms of these sets. In particular, we can obtain maximal approximation error bounds for narrow DBMs, depending on the number of hidden layers. Such bounds would resemble exactly the maximal approximation error bounds obtained for DBNs in (Montúfar 2014) .

Acknowledgment

I am grateful to the Santa Fe Institute, where I was located during the work on this article.

References

H. Ackley, G. E. Hinton, and T. J. Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive Science*, pages 147–169, 1985.

S. Amari, K. Kurata, and H. Nagaoka. Information geometry of Boltzmann machines. *Neural Networks, IEEE Transactions on*, 3(2):260–271, Mar 1992. ISSN 1045-9227. doi: 10.1109/72.125867.

P. Baldi. Autoencoders, unsupervised learning, and deep architectures. In I. Guyon, G. Dror, V. Lemaire, G. W. Taylor, and D. L. Silver, editors, *ICML Unsupervised and Transfer Learning*, volume 27 of *JMLR Proceedings*, pages 37–50. JMLR.org, 2012. URL http://dblp.uni-trier.de/db/journals/jmlr/jmlrp27.html#Baldi12.

A. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *Information Theory, IEEE Transactions on*, 39(3):930–945, May 1993. ISSN 0018-9448. doi: 10.1109/18.256500.

Y. Bengio and O. Delalleau. On the expressive power of deep architectures. In J. Kivinen, C. Szepesvri, E. Ukkonen, and T. Zeugmann, editors, *Algorithmic Learning Theory*, volume 6925 of *Lecture Notes in Computer Science*, pages 18–36. Springer Berlin Heidelberg, 2011. ISBN 978-3-642-24411-7. doi: 10.1007/978-3-642-24412-4_3. URL http://dx.doi.org/10.1007/978-3-642-24412-4_3.

M. Burger and A. Neubauer. Error bounds for approximation with neural networks. *Journal of Approximation Theory*, 112(2):235 – 250, 2001. ISSN 0021-9045. doi: http://dx.doi.org/10.1006/jath.2001.3613. URL http://www.sciencedirect.com/science/article/pii/S0021904501936135.

T. Chen and H. Chen. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *Neural Networks, IEEE Transactions on*, 6(4):911–917, Jul 1995. ISSN 1045-9227. doi: 10.1109/72.392253.

K. Cho, T. Raiko, A. Ilin, and J. Karhunen. How to pretrain deep Boltzmann machines in two stages. In P. Koprinkova-Hristova, V. Mladenov, and N. K. Kasabov, editors, *Artificial Neural Networks*, volume 4 of *Springer Series in Bio-/Neuroinformatics*, pages 201–219. Springer International Publishing, 2015. ISBN 978-3-319-09902-6. URL http://dx.doi.org/10.1007/978-3-319-09903-3_10.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989. URL http://dx.doi.org/10.1007/BF02551274.

Y. Freund and D. Haussler. Unsupervised learning of distributions of binary vectors using 2-layer networks. In J. E. Moody, S. J. Hanson, and R. Lippmann, editors, *Advances in Neural Information Processing Systems 4*, NIPS '91, pages 912–919. Morgan Kaufmann, 1991.

I. J. Goodfellow, A. C. Courville, and Y. Bengio. Joint training deep Boltzmann machines for classification. *CoRR*, abs/1301.3568, 2013a. URL http://arxiv.org/abs/1301.3568.

I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. C. Courville, and Y. Bengio. Maxout networks. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Proceedings*, pages 1319–1327. JMLR.org, 2013b. URL http://jmlr.org/proceedings/papers/v28/goodfellow13.html.

G. E. Hinton and T. J. Sejnowski. Optimal Perceptual Inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1983.

G. E. Hinton and T. J. Sejnowski. Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1. chapter Learning and Relearning in Boltzmann Machines, pages 282–317. MIT Press, Cambridge, MA, USA, 1986. ISBN 0-262-68053-X. URL http://dl.acm.org/citation.cfm?id=104279.104291.

G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, July 2006. ISSN 0899-7667. URL http://dx.doi.org/10.1162/neco.2006.18.7.1527.

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359 – 366, 1989. ISSN 0893-6080. URL http://dx.doi.org/10.1016/0893-6080(89)90020-8.

O. Krause, A. Fischer, T. Glasmachers, and C. Igel. Approximation properties of DBNs with binary hidden units and real-valued visible units. In *ICML (1)*, volume 28 of *JMLR Proceedings*, pages 419–426. JMLR.org, 2013. URL http://dblp.uni-trier.de/db/conf/icml/icml2013.html#KrauseFGI13.

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

N. Le Roux and Y. Bengio. Deep belief networks are compact universal approximators. *Neural Computation*, 22:2192–2207, 2010.

M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken. Multilayer feedforward networks with a non-polynomial activation function can approximate any function. *Neural networks*, 6(6):861–867, 1993.

J. Martens, A. Chattopadhya, T. Pitassi, and R. Zemel. On the representational efficiency of restricted Boltzmann machines. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2877–2885. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5020-on-the-representational-efficiency-of-restricted-boltzmann-machines.pdf.

G. Montavon and K.-R. Müller. Deep Boltzmann machines and the centering trick. In *Neural Networks: Tricks of the Trade*, pages 621–637. Springer Berlin Heidelberg, 2012.

G. Montavon, M. L. Braun, and K.-R. Müller. Deep Boltzmann machines as feed-forward hierarchies. In N. D. Lawrence and M. A. Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS-12)*, volume 22, pages 798–804, 2012. URL http://jmlr.csail.mit.edu/proceedings/papers/v22/montavon12/montavon12.pdf.

G. Montúfar. Mixture decompositions of exponential families using a decomposition of their sample spaces. *Kybernetika*, 49(1), 2013. URL http://www.kybernetika.cz/content/2013/1/23.

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

G. Montúfar 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. Morton. Discrete restricted Boltzmann machines. In *International Conference on Learning Representations*, ICLR '13, 2013. URL http://arxiv.org/abs/1301.3529. Accepted for JMLR special topics issue Learning Representations.

G. Montúfar and J. Morton. When does a mixture of products contain a product of mixtures? *Accepted for SIAM Journal on Discrete Mathematics*, 2014. Preprint available at http://arxiv.org/abs/1206.0387.

G. Montúfar, J. Rauh, and N. Ay. Expressive power and approximation errors of restricted Boltzmann machines. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 415–423. Curran Associates, Inc., 2011. URL http://papers.nips.cc/paper/4380-expressive-power-and-approximation-errors-of-restricted-boltzmann-machines.pdf.

G. Montúfar, N. Ay, and K. Zahedi. Expressive power of conditional restricted Boltzmann machines. *arXiv preprint arXiv:1402.3346*, 2014a. URL http://arxiv.org/abs/1402.3346.

G. Montúfar, R. Pascanu, K. Cho, and Y. Bengio. On the number of linear regions of deep neural networks. In *NIPS '14*, 2014b. Preprint available at http://arxiv.org/abs/1402.1869.

R. Pascanu, G. Montúfar, and Y. Bengio. On the number of response regions of deep feed forward networks with piece-wise linear activations. In *International Conference on Learning Representations*, ICLR '14, 2014. URL http://arxiv.org/abs/1312.6098.

R. Salakhutdinov and G. E. Hinton. Deep Boltzmann machines. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 5, pages 448–455, 2009.

P. Smolensky. Information processing in dynamical systems: foundations of harmony theory. In *Symposium on Parallel and Distributed Processing*, 1986.

H. Sussmann. Learning algorithms for Boltzmann machines. In *Decision and Control, 1988., Proceedings of the 27th IEEE Conference on*, pages 786–791 vol.1, Dec 1988. doi: 10.1109/CDC.1988.194417.

I. Sutskever and G. Hinton. Deep narrow sigmoid belief networks are universal approximators. *Neural Computation*, 20:2629–2636, 2008.

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

L. Younes. Synchronous Boltzmann machines can be universal approximators. *Applied Mathematics Letters*, 9(3):109 – 113, 1996. ISSN 0893-9659. doi: http://dx.doi.org/10.1016/0893-9659(96)00041-9. URL http://www.sciencedirect.com/science/article/pii/0893965996000419.