Max-Planck-Institut
für Mathematik
in den Naturwissenschaften
Leipzig

Refinements of Universal Approximation Results for Deep Belief Networks and Restricted Boltzmann Machines

by

Guido Montúfar, and Nihat Ay

Preprint no.: 23 2010

Refinements of Universal Approximation Results for Deep Belief Networks and Restricted Boltzmann Machines

Guido Montufar,[1,*] **Nihat Ay**[1,2]

[1]Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany.

[2]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA.

Abstract

We improve recently published results about resources of Restricted Boltzmann Machines (RBM) and Deep Belief Networks (DBN) required to make them Universal Approximators. We show that any distribution p on the set $\{0,1\}^n$ of binary vectors of length n can be arbitrarily well approximated by an RBM with $k-1$ hidden units, where k is the minimal number of pairs of binary vectors differing in only one entry such that their union contains the support set of p. In important cases this number is half of the cardinality of the support set of p (given in Le Roux & Bengio (2008)). We construct a DBN with $\frac{2^n}{2(n-b)}$, $b \sim \log n$, hidden layers of width n that is capable of approximating any distribution on $\{0,1\}^n$ arbitrarily well. This confirms a conjecture presented in Le Roux & Bengio (2010).

1 Introduction

This work rests upon ideas presented in (Le Roux & Bengio , 2008) and (Le Roux & Bengio , 2010). We positively resolve a conjecture that was posed in Le Roux &

* montufar@mis.mpg.de

Bengio (2010). Before going into the details of this conjecture we first recall some basic notions.

The definition of RBM's and DBN's that we use is the one given in the papers mentioned above and references therein. For details the reader is referred to those works. Here we give a short description: A Boltzmann Machine consists of a collection of binary stochastic units, where any pair of units may interact. The unit set is divided into *visible* and *hidden* units. Correspondingly the state is characterized by a pair (v, h) where v denotes the state of the visible and h denotes the state of the hidden units. One is usually interested in distributions on the visible states v and would like to generate these as marginals of distributions on the states (v, h). In a general Boltzmann Machine the interaction graph is allowed to be complete. A Restricted Boltzmann Machine (RBM) is a special type of Boltzmann Machine, where the graph describing the interactions is bipartite: Only connections between visible and hidden units appear. It is not allowed that two visible units or two hidden units interact with each other (see Fig. 1). The distribution over the states of all RBM units has the form of the Boltzmann distribution $p(v, h) \propto \exp(h^T W \cdot v + B \cdot v + C \cdot h)$, where v is a binary vector of length equal to the number of visible units, and h a binary vector with length equal to the number of hidden units. The parameters of the RBM are given by the matrix W and the two vectors B and C. A Deep Belief Network consists of a chain of layers of units. Only units from neighboring layers are allowed to be connected, there are no connections within each layer. The last two layers have undirected connections between them, while the other layers have connections directed towards the first layer, the visible layer. The general idea of a DBN is to assume that all layers are of similar size, as shown in Fig. 1.

A major difficulty in the use of Boltzmann Machines always has been the slowness of learning. In order to overcome this problem, DBN's have been proposed as an alternative to classical Boltzmann Machines. An efficient learning algorithm for DBN's was given in the paper Hinton et. al. (2006).

The fundamental questions along the above-mentioned previous work are the following: Does a DBN exist that is capable of approximating any distribution on the visible states through appropriate choice of parameters? We will refer to such a DBN as a universal DBN approximator (similarly we will use the denomination universal RBM approximator). If universal DBN approximators exist, what is their minimal size?

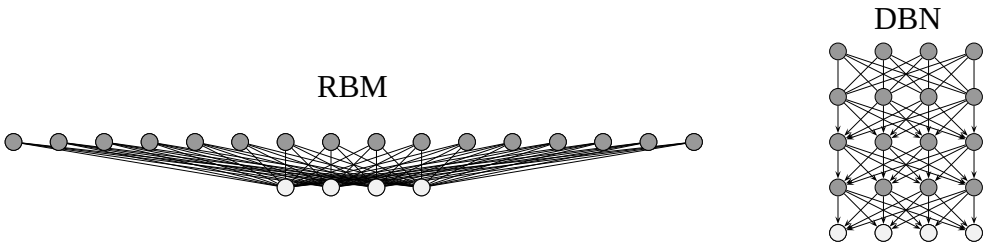

Figure 1: In the left side we sketched the graph of interactions in an RBM, in the right side the corresponding graph for a DBN with $n = 4$ visible units (drawn brighter). An arbitrary weight can be assigned to every edge. Beside this connection weights, every node contains an individual *offset* weight. Every node takes value 0 or 1 with a probability that depends on the weights. The RBM and DBN of size depicted above are examples of universal approximators of distributions on $\{0,1\}^4$ (Le Roux & Bengio (2008) and Le Roux & Bengio (2010) respectively). In the present paper is shown that the number of hidden units in the RBM can be halved, and the number of hidden layers in the DBN can be roughly halved.

Since DBN's are more difficult to study than RBM's, as a preliminary step, corresponding questions related to the representational power of RBM's have been addressed. Theorem 2 in Le Roux & Bengio (2008) shows that any distribution on $\{0,1\}^n$ with support of cardinality s is arbitrarily well approximated (with respect to the Kullback Leibler divergence) by the marginal distribution of an RBM containing $s+1$ hidden units:

Theorem 2 in Le Roux & Bengio (2008). *Any distribution on $\{0,1\}^n$ can be approximated arbitrarily well with an RBM with $s+1$ hidden units, where s is the number of input vectors whose probability does not vanish.*

This theorem proved the existence of a universal RBM approximator. The existence proof of a universal DBN approximator is due to Sutskever & Hinton (2008). More precisely, Sutskever & Hinton (2008) explicitely constructed a DBN with $\sim 3 \cdot 2^n$ hidden layers of width $n+1$ that approximates any distribution on $\{0,1\}^n$. Given that the existence problem of universal DBN approximators was positively resolved through this result, the efforts have been put into optimizing the size, i.e. reducing the number of parameters. This can be done by reducing the number of hidden layers involved in a DBN, or by making the hidden layers narrower. In terms of simple counting arguments,

we give a lower bound on the minimal number of hidden layers required for the universality of a DBN with layers of size n. The number of free parameters in such a DBN is *square of the width of each layer* $\times$ *number of hidden layers* $+$ *number of units*, which for k hidden layers is $k(n^2+n)+n$. On the other hand, the number of parameters needed to describe all distributions over 2^n elements, e.g. over binary vectors of length n, is 2^n-1. Therefore, a lower bound on the number of hidden layers of a universal DBN approximator is given by $\frac{2^n-1-n}{n(n+1)}$ (which yields 2^n-1 free parameters). Otherwise the number of parameters would not be sufficient. Asymptotically, this bound is of order $\frac{2^n}{n^2}$. Certainly, since the architecture of DBN's makes important restrictions on the way the parameters are used, such a lower bound is not expected to be achievable. In particular the approximation of a distribution through a DBN or RBM is not unambiguous, i.e. for several choices of the parameters the same distribution is produced as marginal distribution. However, in Le Roux & Bengio (2010) it has been shown that a number of hidden layers of order $\frac{2^n}{n}$ is sufficient:

Theorem 4 in Le Roux & Bengio (2010). *If $n=2^t$, a DBN composed of $\frac{2^n}{n}+1$ layers of size n is a universal approximator of distributions on $\{0,1\}^n$.*

In the paper Le Roux & Bengio (2010) the optimality of the bound given in this theorem remains an open problem. However, their proof method suggests the sufficiency of less hidden layers, which was conjectured in their paper. The proof of Theorem 4 crucially depends on the authors' previous Theorem 2 in Le Roux & Bengio (2008). Our main contribution is to sharpen Theorem 2 (see Theorem 1 in Section 2) which allows us to even better exploit their method and thereby confirm their conjecture (see Theorem 3 in Section 2). We consider our refinement as particularly interesting because there are reasons to believe that this already provides the optimal bound for the minimal number of hidden layers in a universal DBN approximator.

2 Results

2.1 Restricted Boltzmann Machines

The following Theorem 1 sharpens Theorem 2 in Le Roux & Bengio (2010). We will use it (its Corollary 2) in the proof of our main result, Theorem 3.

Theorem 1 (Reduced RBM's which are universal approximators)**.** *Any distribution p on binary vectors of length n can be approximated arbitrarily well by an RBM with $k-1$ hidden units, where k is the minimal number of pairs of binary vectors, such that the two vectors in each pair differ in only one entry, and such that the support set of p is contained in the union of these pairs.*

The set $\{0,1\}^n$ corresponds to the vertex set of the n-dimensional cube. The edges of the n-dimensional cube correspond to pairs of binary vectors of length n which differ in exactly one entry. For the graph of the n-dimensional cube there exist perfect matchings, i.e., collections of disjoint edges which cover all vertices. Therefore we have the following:

Corollary 2. *Any distribution on $\{0,1\}^n$ can be approximated arbitrarily well by an RBM with $\frac{2^n}{2}-1$ hidden units.*

The proof of Theorem 1 given below is very much in the spirit of the proof of Theorem 2 in Le Roux & Bengio (2008). The idea there consists on showing that given an RBM with some marginal visible distribution, the inclusion of an additional hidden unit allows to increment the probability mass of one visible state vector, while uniformly reducing the probability mass of all other visible vectors.

We show that the inclusion of an additional hidden unit in fact allows to increase the probabiliy mass of a pair of visible vectors, in independent ratio, given that this pair differs in one entry. At the same time, the probability of all other visible states is reduced uniformly. We also use the offset weights in the visible units to further improve the result.

Proof of Theorem 1. We stay close to the notation used in Le Roux & Bengio (2008). 1. Let p be the distribution on the states of visible and hidden units of an RBM. Its marginal probability distribution on v can be written as

$$p(v) = \frac{\sum_h z(v,h)}{\sum_{v^0,h^0} z(v^0,h^0)}.$$

Denote by $p_{w,c}$ the distribution arising through the adding of a hidden unit to the RBM connected with weigths $w = (w_1,\ldots,w_n)$ to the visible units, and with offset weight c. Its marginal distribution can be written as

$$p_{w,c}(v) = \frac{(1+\exp(w\cdot v+c))\sum_h z(v,h)}{\sum_{v^0,h^0}(1+\exp(w\cdot v^0+c))z(v^0,h^0)}.$$

2. Given any vector $v \in \{0,1\}^n$ we write $v_{\hat{j}}$ for the vector defined through $(v_{\hat{j}})_i = v_i, \forall i \neq j$, and $(v_{\hat{j}})_j = 0$. We also write $\mathbb{1} := (1,\ldots,1)$, and $e_j := \mathbb{1} - \mathbb{1}_{\hat{j}}$.

3. For any $j \in \{1,\ldots,n\}$ let $\tilde{v}$ be an arbitrary vector with $\tilde{v}_j = 1$. Define

$$\begin{aligned}
\hat{w} &:= a(\tilde{v}_{\hat{j}} - \frac{1}{2}\mathbb{1}_{\hat{j}}), \\
\bar{w} &:= a(\tilde{v}_{\hat{j}} - \frac{1}{2}\mathbb{1}_{\hat{j}}) + (\lambda_2 - \lambda_1)e_j, \\
\bar{c} &:= -\hat{w}\cdot\tilde{v} + \lambda_1 = -\hat{w}\cdot\tilde{v}_{\hat{j}} + \lambda_1.
\end{aligned}$$

For the weights $\bar{w}$ and $\bar{c}$ we have:

$$\begin{aligned}
\bar{w}\cdot v &= \frac{1}{2}a((n-1) - |\{i : (\tilde{v}_{\hat{j}})_i \neq (v_{\hat{j}})_i\}|) + (\lambda_2 - \lambda_1)v_j, \\
\bar{c} &= -\frac{1}{2}a(n-1) + \lambda_1,
\end{aligned}$$

and in the limit $a \to \infty$ we get:

$$\begin{aligned}
\lim_{a\to\infty} 1 + \exp(\bar{w}\cdot v + \bar{c}) &= 1, \quad \forall v \neq \tilde{v}, \tilde{v}_{\hat{j}}, \\
\lim_{a\to\infty} 1 + \exp(\bar{w}\cdot \tilde{v}_{\hat{j}} + \bar{c}) &= 1 + e^{\lambda_1}, \\
\lim_{a\to\infty} 1 + \exp(\bar{w}\cdot \tilde{v} + \bar{c}) &= 1 + e^{\lambda_2}.
\end{aligned}$$

Just as in the Proof of Theorem 2 in Le Roux & Bengio (2008) this yields for the marginal distribution on the visible states of the enlarged RBM the following:

$$\begin{aligned}
\lim_{a\to\infty} p_{\bar{w},\bar{c}}(v) &= \frac{p(v)}{1 + e^{\lambda_1}p(\tilde{v}_{\hat{j}}) + e^{\lambda_2}p(\tilde{v})}, \quad \forall v \neq \tilde{v}, \tilde{v}_{\hat{j}}, \\
\lim_{a\to\infty} p_{\bar{w},\bar{c}}(\tilde{v}_{\hat{j}}) &= \frac{(1 + e^{\lambda_1})p(\tilde{v}_{\hat{j}})}{1 + e^{\lambda_1}p(\tilde{v}_{\hat{j}}) + e^{\lambda_2}p(\tilde{v})}, \\
\lim_{a\to\infty} p_{\bar{w},\bar{c}}(\tilde{v}) &= \frac{(1 + e^{\lambda_2})p(\tilde{v})}{1 + e^{\lambda_1}p(\tilde{v}_{\hat{j}}) + e^{\lambda_2}p(\tilde{v})}.
\end{aligned}$$

This means that the probability of $\tilde{v}$ and of $\tilde{v}_{\hat{j}}$ can be increased independently by a multiplicative factor, while all other probabilities are reduced uniformly.

4. Now we explain how to start an induction from which the claim follows. Consider an RBM with no hidden units, RBM^0. Through a choice of the offset weigths in every visible unit, RBM^0 produces as visible distribution any arbitrary factorizable distribution $p^0(v) \propto \exp(B\cdot v) \propto \exp(B\cdot v + K)$, where B is the vector of offset weights and K is a constant that we introduce for illustrative reasons, and is not

a parameter of the RBM^0 since it cancels out with the normalization of p^0. In particular, RBM^0 can approximate arbitrarily well any distribution with support given by a pair of vectors that differ in only one entry. To see this consider any pair of vectors $\tilde{v}$ and $\tilde{v}_{\hat{j}}$ that differ in the entry j. Then, the choice $B = a(\tilde{v}_{\hat{j}} - \frac{1}{2}\mathbb{1}_{\hat{j}}) + (\lambda_2 - \lambda_1)e_j$ and $K = -a(\tilde{v}_{\hat{j}} - \frac{1}{2}\mathbb{1}_{\hat{j}})\tilde{v} + \lambda_1$ yields in the limit $\lim_{a\to\infty}$ (similarly to the equations in item 3. above) that $\lim_{a\to\infty} p^0(v) = 0$ whenever $v \neq \tilde{v}$ and $v \neq \tilde{v}_{\hat{j}}$, while $\lim_{a\to\infty} p^0(\tilde{v})/p^0(\tilde{v}_{\hat{j}}) = \exp(\lambda_2 - \lambda_1)$ can be chosen arbitrarily by modifying λ_1 and λ_2. Hence, p^0 can be made arbitrarily similar to any distribution with support $\{\tilde{v}, \tilde{v}_{\hat{j}}\}$. Notice that p^0 remains positive for all v and $a < \infty$.

By the arguments described above, every additional hidden unit allows to increase the probability of any pair of vectors which differ in one entry. Obviously, it is possible to do the same for a single vector instead of a pair. Thence, with every additional hidden unit the support set of the probabilities which can be approximated arbitrarily well is enlarged by an arbitrary pair of vectors which differ in one entry. This is, RBM$^{(i-1)}$ is an approximator of distributions with support contained in any union of i pairs of vectors which differ in exactly one entry. □

We close this passage with some remarks:

The possiblity of independent change of the probability mass of two visible vectors is due to the usability of the following two parameters: a) The offset input weigth in the added hidden unit, and b) the weight of the connection between the added hidden unit and the visible unit where the pair of visible vectors differ. See item 3. in the Proof.

The attempt to use a similar idea to increment the probability mass of three different vectors in independent ratios inducts a coupled change in the probability of a fourth vector. Three vectors differ in at least 2 entries, as do four vectors. Since only 3 parameters are available (the offset of the new hidden unit and two connection weigths), the dependency arises.

It is worth noting, that using exclusively a similar idea will not allow an exension of Theorem 2 in Le Roux & Bengio (2010) to permit the flip of a certain bit with a certain probability (only) given one of three input vectors.

2.2 Deep Belief Networks

In this section we implement our Theorem 1 to modify the construction given in the proof of Theorem 4 in Le Roux & Bengio (2010) and prove our main result, Theorem 3:

Theorem 3 (Reduced DBN's which are universal approximators)**.** *Let* $n = \frac{2^b}{2} + b$, $b \in \mathbf{N}$, $b \geq 1$. *A DBN containing* $\frac{2^n}{2(n-b)}$ *hidden layers of width* n *is a universal approximator of distributions on* $\{0,1\}^n$.

Before proving Theorem 3 we first develop some components of the proof.

An important idea of Sutskever & Hinton (2008) is that of *sharing*, by means of which in a part of a DBN the probability of a vector is increased while the probability of another vector is decreased and the probability of all other vectors remains nearly constant. This idea is refined in Theorem 2 of Le Roux & Bengio (2010):

Theorem 2 in Le Roux & Bengio (2010) (slightly different formulation)**.** *Consider two layers of units indexed by* $i \in \{1,\ldots,n\}$ *and* $k \in \{1,\ldots,n\}$*, and denote by* v *and* h *state vectors in each layer. Denote by* $\{w_{ik}\}_{i,k=1,\ldots,n}$ *the connection weights and by* $\{c_k\}_{k=1,\ldots,n}$ *the offset weights in the second layer. Given any* l *and* j*,* $l \neq j$*, let* a *be an arbitrary vector in* $\{0,1\}^n$ *and* b *another vector with* $b_i = a_i \, \forall i \neq j$*, and* $a_j \neq b_j$*. Then, it is possible to choose weights* $w_{k,l}$*,* $k \in \{1,\ldots,n\}$*, and* c_l *such that the following equations are satisfied with arbitrary accuracy:* $P(v_l = h_l|h) = 1 \forall h \notin \{a,b\}$*, while* $P(v_l = 1|h = a) = p_a$ *and* $P(v_l = 1|h = b) = p_b$ *with arbitrary* p_a, p_b*.*

By this Theorem, a sharing step can be accomplished in only one layer, whereas probability mass is transferred from a chosen vector to another vector differing in one entry. Futhermore, it demands adaptation only of the connection weights and offset weight of one single unit. Thereby, the overlay of a number of sharing steps in each layer is possible.

The main idea in Le Roux & Bengio (2010) was to exploit these circumstances using a clever sequence of transactions of probabilities. The requirements for the realizability of sharing sequences using Theorem 2 in Le Roux & Bengio (2010) can be summarized in properties of sequences of vectors. These properties are described in Theorem 3 of Le Roux & Bengio (2010), or in the items 2-3 of our appropriately modified version of that Theorem, Lemma 4 below.

How the Theorem 2 in Le Roux & Bengio (2010) and Lemma 4 brace the construction of a universal DBN approximator will become clearer in the afterwards following Lemma 5.

Lemma 4. *Let $n = \frac{2^b}{2} + b, b \in \mathbf{N}, b \geq 1$. There exist $a := 2^b = 2(n-b)$ sequences of binary vectors S_i, $0 \leq i < a-1$ composed of vectors $S_{i,k}, 1 \leq k \leq \frac{2^n}{a}$ satisfying the following:*

1. *$\{S_0, \ldots, S_{a-1}\}$ is a partition of $\{0,1\}^n$.*
2. *$\forall i \in \{0, \ldots, a-1\}$, $\forall k \in \{1, \ldots, \frac{2^n}{a} - 1\}$ we have $H(S_{i,k}, S_{i,k+1}) = 1$, where $H(\cdot, \cdot)$ denotes the Hamming distance.*
3. *$\forall i, j \in \{0, \ldots, a-1\}$ such that $i \neq j$ and $\forall k \in \{1, \ldots, \frac{2^n}{a} - 1\}$ the bit switched between $S_{i,k}$ and $S_{j,k+1}$ and the bit switched between $S_{j,k}$ and $S_{j,k+1}$ are different, unless $H(S_{i,k}, S_{j,k}) = 1$.*

Proof of Lemma 4. Let G^0_{n-b} be any Gray code for $(n-b)$ bits. Such a Gray code is a matrix of size $2^{n-b} \times (n-b)$, where every two consecutive rows have Hamming distance one to each other, and the collection of all rows is $\{0,1\}^{n-b}$. Obviously any permutation of columns of this Gray code has the same properties. Let G^i_{n-b} be the cyclic permutation of columns i positions to the left.

Now define $S_i := \begin{pmatrix} \text{bin}(i) & \\ \vdots & G^{i \bmod (n-b)}_{n-b} \\ \text{bin}(i) & \end{pmatrix}$, i.e. the first b bits of the vector $S_{i,k}$ contain the b-bit binary representation of i. The rest of the bits contain the k-th row in the Gray code G^0_{n-b} for arrays of length $n-b$ cyclically shifted i positions to the left. The cyclic permutation makes that every two sequences of vectors S_i and S_j, $i \neq j$ change the same bit in the same row (in this case they also do in every row) only if the value of the first part bin(i) and bin(j) of the two sequences differs in only one entry (in the first entry). □

Every two consecutive vectors in a sequence given in Lemma 4 differ in only one entry and this entry can be located in almost any position $\{1, \ldots, n\}$. In contrast, for the sequences given in Theorem 3 of Le Roux & Bengio (2010) that entry can be located only in a subset of $\{1, \ldots, n\}$ of cardinality $n/2$.

In the Lemma above, for any row, every one of $n-b$ entries is flipped by exactly two sequences. Regard that the attempt to produce $2n$ instead of $2(n-b)$ sequences with the properties 1-2 of the Lemma (and flips in all entries) would correspond to the following: Set $\begin{pmatrix} S_1 \\ \vdots \\ S_{2n} \end{pmatrix} = G_n$, i.e., the sequences to be overlayed are portions of the same Gray code. In this case it is difficult to achive that condition 3. is satistfied, i.e., that if S_i and S_j flip the same bit in the same row, then $H(S_{i,k}, S_{j,k}) = 1$. The condition 3. however is essential for the use of Theorem 2 of Le Roux & Bengio (2010). Most common Gray codes flip some entries more often than other entries and can be discarded. Oher sequences referred to as *totally balanced Gray codes* flip all entries equally often and exist whenever n is a power of 2, but still a strong cyclicity condition would be required. On account of this we say that the sequences given in Lemma 4 allow optimal use of Theorem 2 of Le Roux & Bengio (2010).

The following Lemma 5 is a transcription of Lemma 1 in Le Roux & Bengio (2010) with replacements of indices according to our construction. The proof is an obvious transcription which we omit here. Denote by h^i a state vector of the units in the hidden layer i, and denote by h^0 a visible state.

Lemma 5. *Let p^* be an arbitrary distribution on $\{0,1\}^n$. Consider a DBN with $\frac{2^n}{a}+1$ layers and the following properties:*

1. *$\forall i \in \{0,\ldots,a-1\}$ the top RBM between $h^{\frac{2^n}{a}}$ and $h^{\frac{2^n}{a}-1}$ assigns probability $\sum_k p^*(S_{i,k})$ to $S_{i,1}$,*

2. *$\forall i \in \{0,\ldots,a-1\}$, $\forall k \in \{1,\ldots,\frac{2^n}{a}-1\}$*

$$\begin{aligned} P(h^{\frac{2^n}{a}-(k+1)} = S_{i,k+1} | h^{\frac{2^n}{a}-k} = S_{i,k}) &= \frac{\sum_{t=k+1}^{\frac{2^n}{a}} p^*(S_{i,t})}{\sum_{t=k}^{\frac{2^n}{a}} p^*(S_{i,t})}, \\ P(h^{\frac{2^n}{a}-(k+1)} = S_{i,k} | h^{\frac{2^n}{a}-k} = S_{i,k}) &= \frac{p^*(S_{i,k})}{\sum_{t=k}^{\frac{2^n}{a}} p^*(S_{i,t})}, \end{aligned}$$

3. *$\forall k \in \{1,\ldots,\frac{2^n}{a}-1\}$ the DBN provides*

$$P(h^{\frac{2^n}{a}-(k+1)} = u | h^{\frac{2^n}{a}-k} = u) = 1, \quad \forall u \notin \cup_i \{S_{i,k}\}.$$

Such a DBN has p^* as its marginal visible distribution.

We conclude this section with the proof of Theorem 3 and some remarks:

Proof of Th. 3. The proof is analogous to the Proof of Theorem 4 in Le Roux & Bengio (2010). We just need to show the existence of a DBN with the properties of the DBN described in Lemma 5. In view of Theorem 1 it is possible to achive that the top RBM assigns arbitrary probability to the collection of vectors $S_{i,1}, i \in \{0, \ldots, a-1\}$, whenever it can be arranged in pairs of neighbouring vectors (or from Corollary 2, if all vectors are equal in a set of entries). This requirement is met for $S_{i,1}, i \in \{0, \ldots, a-1\}$ of Lemma 4, (e.g. choosing a Gray code whose first element is $(0, \ldots, 0)$ or $(1, \ldots, 1)$). The subsequent layers are just like in the Proof of Theorem 4 in Le Roux & Bengio (2010). They are possible in consideration of the mantained validity of Theorem 2 in Le Roux & Bengio (2010) using the sequences provided in Lemma 4 of the present paper. The only difference is that by our definition of S_i, $i \in \{0, \ldots, a-1\}$, at each layer $n - b$ bit flips (with correct probabilities) occur, instead of $\frac{n}{2}$. □

In the paper Le Roux & Bengio (2010) the authors overlayed n sequences of sharing steps (Theorem 3 in that paper) for the construction of a universal DBN approximator. In principle an overlay of more such sequences is possible. This is what we exploit in our proof, (the sequences given in Lemma 4). Apparently, the overlay of more sequences was not realized in that paper because for the initialization of these sequences, (property 1. in Lemma 1 in that paper), the authors use Theorem 2 of Le Roux & Bengio (2008), which only allows to assign arbitrary probability to n vectors. Our result Theorem 1 overcomes this difficulty and allows to initialize up to $2(n+1)$ sequences, which we use to obtain property 1. in Lemma 5.

Conclusion

We have shown that a Deep Belief Network (DBN) with $\frac{2^n}{2(n-b)}$, $b \sim \log n$, hidden layers of size n is capable of approximating any distribution on $\{0,1\}^n$ arbitrarily well as its marginal visible distribution. (This confirms a conjecture presented in Le Roux & Bengio (2010)). The number of layers $\frac{2^n}{2(n-b)}$ is of order $\frac{2^n}{2n}$. This DBN has $\frac{2^n}{2(n-b)}n^2 + \frac{2^n}{2(n-b)}n + n$ parameters, which is of order $\frac{n2^n}{2}$.

Furthermore, we have shown that a Restricted Boltzmann Machine (RBM) with $\frac{2^n}{2} - 1$ hidden units is capable of approximating any distribution on $\{0,1\}^n$ arbitrarily well as its marginal visible distribution. This RBM has $\frac{2^n}{2}n + \frac{2^n}{2}$ parameters, which is of order $\frac{n2^n}{2}$.

Our results improve all to date known bounds on the minimal size of universal DBN and RBN approximators. Our construction already exploits Theorem 2 in Le Roux & Bengio (2010) exhaustively, and therefore a construction using similar ideas will not allow improvements. We still do not know if our results represent the minimal sufficient size for universal DBN and RBN approximators, but we have reasons to belive that they do. This is subject of our ongoing research, Montufar & Ay (2010).

References

Hinton, G. E., Osindero, S. and Teh, Y. (2006). A fast learning algorithm for deep belief nets. *Neural Computation, 18,* 1527-1554.

Roux, Nicolas Le & Bengio, Yoshua (2010). Deep Belief Networks Are Compact Universal Approximators. *Neural Computation*.

Roux, Nicolas Le & Bengio, Yoshua (2008). Representational Power of Restricted Boltzmann Machines and Deep Belief Networks. *Neural Computation, 6,* 1631-1649.

Montufar, Guido & Ay, Nihat (2010). Faces of the Probability Simplex contained in the closure of an Exponential Family and Minimal Mixture Decompositions. *in preparation*.

Sutskever, I. & Hinton, G. E. (2008). Deep Narrow Sigmoid Belief Networks are Universal Approximators. *Neural Computation, 20,* 2629-2636.