Mixture Models and Representational Power of RBM's, DBN's and DBM's

Guido Montufar[*]
Max Planck Institute for Mathematics in the Sciences,
Inselstraße 22, D-04103 Leipzig, Germany.
montufar@mis.mpg.de

Abstract

Here we give a contribution intended to help working out the minimal size of Restricted Boltzmann Machines (RBM's), Deep Belief Networks (DBN's) or Deep Boltzmann Machines (DBM's) which are universal approximators of visible distributions. The representational power of these objects arises from the marginalization of hidden units, an operation which naturally produces mixtures of conditional distributions. We present results on the representational power of mixture models with factorizing mixture components, in particular a sharp bound on the required and sufficient number of mixture components to represent any arbitrary visible distribution. The methods disclose a class of visible distributions which require the maximal number of mixture components while all mixture components must be atoms. We derive a test of universal approximating properties and find that an RBM with more than $2^n - 1$ parameters is not always a universal approximator of distributions on $\{0, 1\}^n$.

1 Introduction

Lately many efforts have been put into optimizing the size of stochastic networks which are able to approximate arbitrary visible distributions as marginals through appropiate choice of their parameters, [1, 2, 3, 4]. These works are constructive, which means that a device is constructed which is shown to be a universal approximator. This yields sufficiency conditions on the number of hidden units, layers, and parameters.

We first show that 'naive' parameter counting yields a lower bound for the number of hidden units or of parameters of Restricted Boltzmann Machines (RBM's), Deep Belief Networks (DBN's) or Deep Boltzmann Machines (DBM's) must have in order to be a universal approximators. In this note we give theoretical results intended to help disclose the minimal size of an RBM's, a DBN's, or a DBM's which are universal approximators of visible distribution. We present necessary and sufficient conditions for a mixture of factorizing distributions to represent arbitrary distributions. Such conditions are relevant for understanding the representational power of DBN's, DBM's and RBM's, since mixtures are naturally produced when marginalizing the hidden units. We think also that the ideas involved can be expanded to work out classes of distributions which are best represented by a DBN, a DBM or an RBM. Based on conditions that we derive for mixture models, we derive conditions for RBM's which are universal approximators and propose a test for checking whether an RBM is a universal approximator.

[*] http://personal-homepages.mis.mpg.de/montufar/

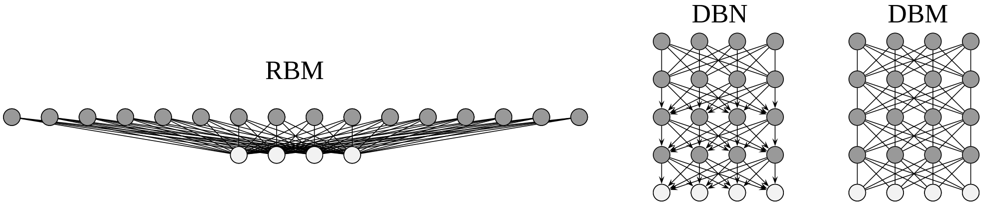

Figure 1: (Taken from [4]). Left: Graph of interactions in an RBM. Middle: Interaction graph for a DBN with $n = 4$ visible units (drawn brighter), Right: The interaction graph of a DBM with 4 visible units. An arbitrary weight can be assigned to every edge. Beside this connection weights, every node contains an individual *bias* weight. Every node represents a unit which takes value 0 or 1 with a probability that depends on the weights. A RBM and a DBN with the architectures depicted above can approximate any distributions on $\{0,1\}^4$ arbitrarily well through appropriate adjustment of parameters ([2] and [3] respectively). In [4] was even 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.

1.1 Deep Belief Networks, Deep Boltzmann Machines and Restricted Boltzmann Machines

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 $p(v, h) \propto \exp(hWv + Bv + Ch)$, where v is a binary vector of length equal to the number of visible units, and h is 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 that the interaction structure is rather deep than shallow, i.e. each hidden layer is not very large compared to the visible layer, as shown in Fig. 1. Denoting by h^k the state vector of the layer k, the joint distribution on the states of all units of a DBN is of the following form:

$$\begin{aligned} P(h^0, h^1, \dots, h^l) &= P(h^{l-1}, h^l) \prod_{k=0}^{l-2} P(h^k | h^{k+1}), \\ P(h^k | h^{k+1}) &= \prod_{j=1}^{n_k} P(h^k_j | h^{k+1}), \\ P(h^k_j | h^{k+1}) &\propto \exp\left(h^k_j b^k_j + h^k_j \sum_{i=1}^{n_{k+1}} W^{k+1}_{j,i} h^{k+1}_i \right). \end{aligned}$$

A Deep Boltzmann Machine (DBM) has the same interaction structure as a DBN, but with undirected connection weights. The distribution on the states of all units is a Boltzmann-Gibbs distribution with interaction structure of the form sketched in Fig. 1.

An RBM, a DBN or a DBM is a universal approximator of distributions on the states of n visible binary units if any distribution p_V on $\{0,1\}^n$ can be arbitrarily well approximated by a marginal distribution

$$p_V(v) = \sum_h p(v,h), \tag{1}$$

where p is the joint distribution on the states of all units of that RBM, DBN or DBM with an appropiate choice of the bias and connecting weights.

We will refer to a DBN that is capable of approximating any distribution on the visible states arbitrarily well (through appropriate choice of parameters) as a universal DBN approximator. Similarly we will use the denomination universal RBM approximator and universal DBM approximator.

1.2 Mixture Models

A probability mixture model consists of a set of distributions which can be written as convex combination of distributions belonging to some further set of distributions, see for instance [5, 6, 7].

In discrete mixture models a family of distributions $\mathcal{E} \subseteq \overline{\mathcal{P}(\mathcal{X})}$ is given, where $\overline{\mathcal{P}(\mathcal{X})}$ is the set of all joint distributions of n random variables $(X_1,\ldots,X_n) =: X$ with sample space $\mathcal{X} = \{0,1\}^n$ (we consider here binary variables). A natural way to understand mixture models, see [7], is to assume that there is a hidden random variable Y with state space $\{0,1\}^m$, and that for each $y \in \{0,1\}^m$, a mixture component is given by the conditional distribution of X given $Y = y$, $p_y \in \mathcal{E}$. If the random variable Y has distribution α, then the joint distribution of Y and X is given by

$$\Pr(Y = y, X = x) = \alpha(y)\, p_y(x).$$

Since the variable Y is assumed to be hidden, only the marginal distribution of X is visible, i.e.,

$$\Pr(X = x) = \sum_{y=1}^{m} \alpha(y)\, p_y(x).$$

Suppose for example that for any y the mixture component p_y can be chosen exclusively but arbitrarily from $\{\delta_x\}_x$. Then, the convex combinations of the form

$$\sum_y \alpha(y)\, \delta_{x_y}(x)$$

clearly cover all visible distributions (if there are as many y as x, and α is arbitrary). This is simply a direct parametrization of the visible distribution in terms of its values on the different x. On the other hand, this model has $2^n - 1 = |\mathcal{X}| - 1$ parameters (defining an arbitary α) and it is clear that a smaller number of mixture components would not suffice to represent some distributions. Not even to arbitrarily well approximate some distributions. More generally, a problem arises when α cannot be chosen arbitrarily, as is the case when it comes to approximating probability distributions as marginals as in RBM's. We will comment on this later.

We will focus on the situation where the mixture weights α (distribution on the states of the hidden units) can be chosen arbitrarily, and ask what happens when one allows more general mixture components than $\{\delta_x\}$, e.g. factorizing distributions, as are the conditional distributions of DBN's and RBM's. How many mixture components of this kind are required and sufficient if we want to represent any distribution? How many are required if we want to represent only distributions from some class?

The analysis of the representational power of general mixture models bears many difficulties. However, in the case of mixtures of factorizing distributions, appealing results can be achieved, as will be outlined below.

We denote the set of all factorizing distributions (on n binary variables) by $\overline{\mathcal{E}^1}$, and the subset of strictly positive distributions by $\mathcal{E}^1$. We consider here mixtures with components from $\overline{\mathcal{E}^1}$. This is the following set:

$$\mathrm{Mixt}^m(\overline{\mathcal{E}^1}) := \left\{ \sum_{j=1}^{m} \alpha_j f_j : \alpha_i \geq 1, \sum \alpha_i = 1 \text{ and } f_j(x_1,\ldots,x_n) = f_j^1(x_1)\cdots f_j^n(x_n) \right\},$$

where $(x_1, \dots, x_n) \in \{0,1\}^n =: \mathcal{X}$. The set of factorizing distributions contains all atoms $\{\delta_x\}_{x \in \mathcal{X}}$, and since these are the extremal points of the set of distributions on $\mathcal{X}$, any distribution can be represented as a mixture of $|\mathcal{X}|$ elements, (when the mixture weights can be chosen arbitrarily).

Does a smaller number of mixture components suffices? What is the minimal necessary number, and how does it depend on the number of random variables n? As we will see it is possible to derive conditions which also apply in the case of constrained α. Furthermore, through this analysis we find a class of distributions which needs the largest number of factorizing mixture components to be represented and are hence a good choice of distributions to test whether a DBN, RBM or a DBM is a universal approximator. We will discuss this below.

1.3 Idea for understanding the representational power of mixture models

Consider the mixtures of two factorizing distributions on two binary variables. The set of mixtures of two fixed elements can be represented as a line connecting the two elements. If the two elements are not fixed, the set of lines connecting points covers all of the probability simplex, see Fig. 2. This reproduces the content of Theorem 2 in [8].

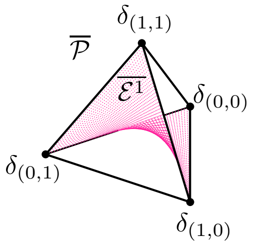

Figure 2: Simplex of distributions on two binary variables $\overline{\mathcal{P}}$, and the set of factorizing distributions $\overline{\mathcal{E}^1}$. Arbitrary mixtures of two elements in $\overline{\mathcal{E}^1}$ are given by a straight line connecting the two points. In the case of constrained mixture weights, only some points on that straight line are covered by the mixture model, and not all points on $\overline{\mathcal{E}^1}$ are necessarily covered.

The situation becomes more complicated for n larger than 2, since the dimension of the set of factorizing distributions, n, increases only as the logarithm of the dimension of the probability simplex, $2^n - 1$. However, a closer inspection reveals that mixtures of two elements lying in the intervals $[\delta_{(0,1)}, \delta_{(1,1)}]$ and $[\delta_{(1,0)}, \delta_{(0,0)}]$, which in fact belong to $\overline{\mathcal{E}^1}$, already suffice to cover all the probability simplex. The sets of distributions described by these intervals have the special property that they comprend all possible distributions with support sets $\{(0,1),(1,1)\}$ and $\{(1,0),(0,0)\}$ and that the union of those two sets is $\mathcal{X}$. That observation is elaborated in [9] for arbitrary n, where it is shown that all distributions with support restricted to some special sets are contained in the independence model and that $2^n/2$ such sets cover all the state space $\{0,1\}^n$ for arbitrary n. This can be directly used to decompose arbitrary distributions as mixtures of independent distributions. We discuss those results in Section 3.

2 Lower bound on the number of parameters

An intuitively simple, but important observation is the following:

Lemma 1. *For an RBM, a DBM and a DBN to approximate any visible distribution on $\{0,1\}^n$ arbitrarily well, necessarily the number of parameters has to be at least equal to $2^n - 1$, the dimension of the set of all distributions on $\{0,1\}^n$.*

This is a straight forward generalization of the result for DBN's presented in [4]. It is only needed that in an RBM and in a DBM the set of joint distributions on all units is a manifold (an exponential family), the closure of which contains all distributions arising for infinite parameters. In fact the proof given in [4] can be seen to work for more general networks wich produces visible distributions via marginalization, e.g. Boltzmann Machines with higher order interactions and the like.

2.1 Lower bound on the number of layers and units

The number of free parameters in a DBN and in a DBM with layers of constant size is *square of the width of each layer* $\times$ *number of hidden layers* $+$ *number of units*, which for k hidden layers of width n is $k(n^2+n)+n$. On the other hand, the number of parameters needed to describe all distributions on $\{0,1\}^n$ is 2^n-1. Therefore, a lower bound on the number of hidden layers of a universal DBN approximator and a universal DBM approximator is given by $\frac{2^n-1-n}{n(n+1)}$, (which yields 2^n-1 free parameters). This makes a number of $\frac{2^n-n-1}{n+1}$ hidden units. Otherwise the number of parameters would not be sufficient. Asymptotically, this bound is of order $\frac{2^n}{n^2}$ hidden layers and $\frac{2^n}{n}$ hidden units.

An RBM with n visible units and m hidden units contains $n+m+n\cdot m$ parameters. We therefore have the condition $n+m+n\cdot m\geq 2^n-1$ in order to have a universal approximator. This yields $m\geq\frac{2^n-n-1}{n+1}$ hidden units, which is of order $\frac{2^n}{n}$.

Corollary 2. *A DBN and a DBM with layers of width $n+c$ must have at least a number of layers of order $\frac{2^n}{n^2}$ (and $\frac{2^n}{n}$ hidden units) in order to be a universal approximator of distributions on n visible units. An RBM must have a number of order $\frac{2^n}{n}$ hidden units in order to be a universal approximator of distributions on n visible units.*

This in particular solves a problem raised by Sutskever and Hinton in [1]: *Can it be shown that a deep and narrow (with width $n+c$) network of $\ll 2^n/n^2$ layers cannot approximate every distribution?*

Interestingly we see that in both cases, deep and shallow architectures, the bound on the number of hidden units is exactly the same. Minimizing the number of hidden units $\sum_l n_l$ (n_l is the number of units in layer l) of a network with pairwise interactions between neighboring layers, while keeping the number of parameters $\sum_l n_{l-1}n_l+\sum_l n_l$ constant to the minimal necessary value 2^n-1 yields something of the form: two hidden layers of size $\sqrt{2^n}$, (and $2\cdot\sqrt{2^n}$ hidden units).

However, RBM's and DBN's, make important and distinctive restrictions on the way the parameters are used, (only pairwise interactions, interactions only among units in neighboring layers). Therefore, the bound derived above is not necessarily achievable.

For a better recognition of the virtues and drawbacks of the two architectures (deep and narrow, broad and shallow) it is necessary to understand if this theoretical bounds are actually achievable for them.

In the remainder of this note we review mixture models from the perspective of hidden causes (hidden units), and derive facts and obstructions in the representability of arbitrary distributions as mixtures of factorizing distributions. As we shall see, obstructions which apply in the case of arbitrary mixture weights already can be used to derive obstructions in the case of constrained mixture weights, and improve bounds for the number of hidden units and hidden layers required to have a RBM and a DBN or DBM be a universal approximator.

3 Constructions and Mixture Models

Here we want to discuss the to date best known bounds on the minimal number of hidden units and layers required to have a universal RBM and DBN approximator and relations to more deep results on mixture models.

In [2] was shown that any distribution on $\{0,1\}^n$ can be arbitrarily well approximated by the marginal distribution of an RBM. And the smallest to date known sufficient number of hidden units is the following (given in [4] based on a refinement of [2]):

Theorem 1 in [4] (*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 exactly one entry, and such that the support set of p is contained in the union of these pairs.*

This result allowed a refinement of the construction presented in [3], and the derivation of the to date smallest universal DBN approximator:

Theorem 3 in [4] (*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 $\{0,1\}^n$.*

A central tool used in the proofs of these statements is that the set of factorizing distributions on $\{0,1\}^n$ contains all distributions with support given by an arbitrary pair of vectors in $\{0,1\}^n$ which differ in exactly one entry.

The pairs of binary vectors differing in exactly one entry are in fact the only sets for which any distribution with support therein is contained in the set of factorizing distributions. This is a consequence of results in [9] which require a mathematical framework that we omit at this point.

Proposition 3. (Support sets of factorizing distributions, [9]) *$\mathcal{Y} \subseteq \{0,1\}^n$ is the support set of a factorizing distribution on $\{0,1\}^n$ if and only if $\mathcal{Y}$ constitutes the vertices of a face of the n-dimensional unit cube. The set of factorizing distributions contains every distribution with support $\mathcal{Y}$ if and only if $\mathcal{Y}$ has cardinality one, or consists of two binary vectors differing in exactly one entry.*

Now, a mixture of $2^n/2$ arbitrary distributions with support on disjonit pairs of vectors differing in exactly one entry yields any arbitrary distribution on $\{0,1\}^n$ if the mixture weights can be chosen arbitrarily, which yields the following result. See Fig. 3.

Theorem 4. (Sufficient number of mixture components, [9]) *Any disribution on $\{0,1\}^n$ can be written as mixture of 2^{n-1} factorizing distributions, given that the mixture weights are not constrained.*

This result tells us that in the case of arbitrary mixture weights, (the distribution on the hidden states can be made arbitrary), a number of 2^{n-1} hidden states suffices.

We now show that Theorem 4 in fact is optimal, and hence, that the representation of arbitrarily distributions indeed requires a huge amonut of mixture components from the set of factorizing distributions:

Theorem 5. (Minimal sufficient number of mixture components, [9]) *For the mixture model with mixture components from the set of factorizing distributions to approximate any distribution on $\{0,1\}^n$ arbitrarily well, it is necessary that the number of mixture components is at least 2^{n-1}. Distributions with support given by Z, defined as the set of vertices of the n-cube taking the same color in a 2-coloring, see Fig. 3, require that the mixture contains only mixture components which are atoms.*

Proof of Theorem 5. Define Z as the set of vertices of the n-cube which are assigned the same color in a 2-coloring of the graph of the n-cube. For example all vectors with an even number of ones, $Z := \{x \in \mathcal{X} : \sum_i x_i = 2k\}$ defines a 2-coloring of the n-cube, since for any edge of the n-cube with vertices $\{x_1, x_2\}$ we have that x_1 and x_2 differ in exactly one entry, and thus no edge is contained in Z. Clearly, $|Z| = 2^n/2$. Furthermore, no subset of Z of cardinality larger than one corresponds to the vertices of a face of the n-cube, and hence is not the support of a factorizing distribution, Proposition 3. To see this regard that any such face would contain an edge, whose vertices would be in Z, in contradiction to its definition.

Consider any distribution p with support Z. If p is written as a mixture of factorizing distributions, $p = \sum_i \alpha_i f_i$, then every f_i (for which $\alpha_i > 0$) must have support contained in Z and it must correspond to a face of the n-cube, Proposition 3. Hence, $|\text{supp} f_i| = 1, \forall f_i$ for which $\alpha_i > 0$. Clearly also, at least $|Z| = 2^n/2$ components are needed.

To finish notice that $\text{Mixt}^m(\mathcal{E}^1) \supseteq \mathcal{P}$ implies $\overline{\text{Mixt}^m(\mathcal{E}^1)} = \text{Mixt}^m(\overline{\mathcal{E}^1}) = \overline{\mathcal{P}}$. Hence, the representability of all strictly positive distributions (leaving aside the not strictly positive distributions) also requires $2^n/2$ mixture components. □

There exist distributions on $\{0,1\}^n$ which cannot be written as a mixture of less than 2^{n-1} elements from the independence model $\overline{\mathcal{E}^1}$, even if the mixture weights are arbitrary.

Furthermore the components in the mixture are unique and given by atoms, i.e. distributions which put mass one on one visible state, and mass 0 on all other visible states.

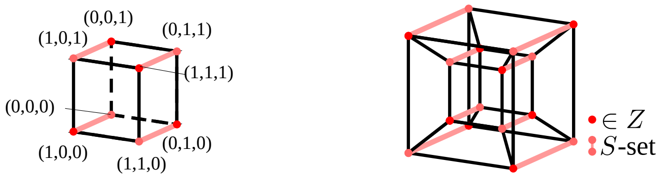

Figure 3: The graph of the 3 and 4 dimensional cubes. The faces of these objects are the support sets of factorizing distributions on $\{0,1\}^3$ and $\{0,1\}^4$. Every distribution with support in the vertices of any edge is a factorizing distribution. These support sets are denoted S-sets. Minimal coverings of $\mathcal{X}$ using disjoint edges-sets are shown. Also a set Z of elements with the same color in a 2-coloring. Distributions with support in such a set need the maximal number of factorizing mixture components, Theorem 5.

If the visible distribution is a mixture of independent distributions, ($Pr(\cdot|Y)$ is an independent distribution on the visible states for any value of Y), then Y must be a variable which takes at least 2^{n-1} different states. In case that Y takes not more than the specified 2^{n-1} different states it is furthermore required that the distribution of Y can be choosen arbitrarily.

Hence we see that for a DBN or an RBM to represent or approximate distributions with support given by a Z defined as above, all conditional visible distributions for a hidden state which occurs with positive probability h, must be atoms.

The above quoted result (from [4]) about univesal RBM approximators tells us that $2^{n-1}-1$ hidden units suffice. Regard that $2^{n-1}-1$ hidden units correspond to $2^{2^{n-1}-1}$ different hidden states. There are two reasons why a universal RBM approximator requires a large amount of hidden units:

1. The distribution on the hidden states cannot be chosen arbitrarily.
2. The factorizing mixture components of the RBM share the parameters B, and differ only through Wh, for the different states h of the hidden units.

It seems that in the construction of the universal RBM approximator a hidden unit is used for the generation of each mixture component. We think that understanding this issue can help improving the constructions of RBM's and DBN's, or showing that the present constructions are already optimal.

4 Test of universal approximating properties

The distributions with support set Z as defined above (vertices of the graph of the n-dimensional unit cube which are assigned the same color in a 2-coloring) are particularly difficult to represent as mixtures of independent distributions. We have seen that any representation of these distributions as mixtures of factorizing distributions must consist of mixture components which are atoms, and hence that $2^n/2$ components are required.

It is therefore appealing that testing the representability of distributions with support Z is a good way to show that a stochastic network is not a universal approximator.

Here we test the representability of distributions with support Z for different sizes of an RBM. To do so first observe that the marginal visible distribution of a RBM with parameters W, B, C is given by

$$p_V = \sum_h \exp(hWv + Bv + Ch) / \sum_{v,h} \exp(hWv + Bv + Ch). \tag{2}$$

From the results of the last section we have that if p_V is requested to be a distribution with support Z, then for any fixed h the function on v $\exp(hWv + Bv + Ch)$ must be proportional to an atom, i.e.:

$$hWv + Bv + Ch = \begin{cases} \lambda_{v'}, v = v' \\ -\infty, v \neq v'. \end{cases} \tag{3}$$

This yields that in order to have the right hand side representing distributions with support on Z a series of equations on W, B, C must be solvable:

$$(\underline{h}|\mathbb{1}) \begin{pmatrix} W & \underline{C} \\ \underline{B} & 0 \end{pmatrix} \begin{pmatrix} \underline{v} \\ \mathbb{1} \end{pmatrix} \overset{!}{=} \begin{pmatrix} -\infty & -\infty & \cdots & \\ \lambda^{1,1} & -\infty & \cdots & \\ -\infty & \lambda^{2,1} & -\infty & \\ \lambda^{1,k_1} & -\infty & \cdots & \end{pmatrix} =: \underline{\lambda}, \tag{4}$$

where $\underline{h} = \begin{pmatrix} h_1 \\ \vdots \\ h_{2^m} \end{pmatrix}$ is a list of all states of the hidden units, and $\underline{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_{2^n} \end{pmatrix}^T$ is a list of all visible states. The matrix in the RHS has 2^n columns, (each one corresponding to a visible state vector), and 2^m rows, (each one corresponding to a hidden state vector). In each row only one value may be different from $-\infty$, (since all mixture components must be atoms, corresponding to eq. 3), and the sum of the values differing from $-\infty$ in the column i sums up to λ_i, where $\exp(\sum_l \lambda_{i,l}) \propto p_v(v_i)$.

This equation can be reformulated as a usual linear equation on the variables contained in W, B, C using properties of matrix equations:

$$\left(\begin{pmatrix} \underline{v} \\ \mathbb{1} \end{pmatrix}^T \otimes (\underline{h}|\mathbb{1}) \right) \operatorname{vec} \begin{pmatrix} W & \underline{C} \\ \underline{B} & 0 \end{pmatrix} = \operatorname{vec} \underline{\lambda}, \tag{5}$$

where $\operatorname{vec} M$ is the vector which arises putting all columns of M in a single column, and $\otimes$ denotes the usual Kronecker product of matrices.

Preliminary results

For simplicity we tested whether the left hand side of eq. 5 can produce a RHS of the form of $\underline{\lambda}$ for the special case where all $\lambda^{i,k}$ are set to 0, and all other values are only required to be different from 0 and have common sign. This is a problem of linear programming which we treated with Matlab and found the following preliminary results on small systems:

n	2	2	3	4	3	3	4	4
m	2	3	2	2	3	4	3	5
satisfies parameter bound	yes	yes	yes	no	yes	yes	yes	yes
equation is solvable	yes	yes	no	no	yes	yes	yes	yes

In the case of 3 visible and 2 hidden units we have $3 + 2 + 6 = 11$ parameters. The lower bound on the number of parameters derived in the first section is $2^3 - 1 = 7$. Hence we see that although that bound is satisfied, the RBM fails to be a universal approximator.

5 Conclusion

Parameter counting can be used to compute lower bounds on the number of hidden units and layers of universal DBN approximators, universal RBM approximators, and universal DBM approximators. We provided insights on mixture models which allow to analyze the representational power DBN's and RBM's. Using this we showed that the bounds derived by parameter counting are not allways achievable.

We showed that a property of the set of factorizing distributions (that it contains all distributions with support given by any pair of vectors differing in exactly one entry) used in the derivation of the smallest to date known universal RBM approximators and universal DBN approximators [4] cannot be further enlarged.

We found a similarity between the sufficient number of mixture components in the mixture model of factorizing distributions, and the sufficient number of hidden for a RBM and a DBN to be a univesal approximator, which is worth to be further investigated. We also presented necessary numbers for the mixutre model, which could imply necessary numbers for RBM's and DBN's.

A natural next step is to derive results for more general mixture models, e.g. for the case when the mixture weights belong to a certain model.

References

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

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

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

[4] Guido Montufar and Nihat Ay. Refinements of universal approximation results for deep belief networks and restricted boltzmann machines. *To appear in Neural Computation,* 2010.

[5] B. G. Lindsay. *Mixture Models: theory, geometry, and applications*. NSF-CBMS Regional Conference Series in Probability and Statistics, 1995.

[6] D.M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical analysis of finite mixture distributions*. John Wiley and Sons, 1985.

[7] Mathias Drton, Bernd Sturmfels, and Seth Sullivant. *Lectures on Algebraic Statistics*. Oberwolfach Seminars vol. 39, Birkhäuser, 2009.

[8] Shun-ichi Amari. Conditional mixture model for correlated neural spikes. *Neural Computation*, 22:1718–1736, 2010.

[9] Guido Montufar. Mixture decompositions using a decomposition of the sample space. *Submitted to Kybernetika*.