Approximation properties of DBNs with binary hidden units and real-valued visible units

Oswin Krause OSWIN.KRAUSE@DIKU.DK
Department of Computer Science, University of Copenhagen, 2100 Copenhagen, Denmark

Asja Fischer ASJA.FISCHER@INI.RUB.DE
Institut für Neuroinformatik, Ruhr-Universität Bochum, 44780 Bochum, Germany, and
Department of Computer Science, University of Copenhagen, 2100 Copenhagen, Denmark

Tobias Glasmachers TOBIAS.GLASMACHERS@INI.RUB.DE
Institut für Neuroinformatik, Ruhr-Universität Bochum, 44780 Bochum, Germany

Christian Igel IGEL@DIKU.DK
Department of Computer Science, University of Copenhagen, 2100 Copenhagen, Denmark

Abstract

Deep belief networks (DBNs) can approximate any distribution over fixed-length binary vectors. However, DBNs are frequently applied to model real-valued data, and so far little is known about their representational power in this case. We analyze the approximation properties of DBNs with two layers of binary hidden units and visible units with conditional distributions from the exponential family. It is shown that these DBNs can, under mild assumptions, model any additive mixture of distributions from the exponential family with independent variables. An arbitrarily good approximation in terms of Kullback-Leibler divergence of an m-dimensional mixture distribution with n components can be achieved by a DBN with m visible variables and n and $n+1$ hidden variables in the first and second hidden layer, respectively. Furthermore, relevant infinite mixtures can be approximated arbitrarily well by a DBN with a finite number of neurons. This includes the important special case of an infinite mixture of Gaussian distributions with fixed variance restricted to a compact domain, which in turn can approximate any strictly positive density over this domain.

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

1. Introduction

Restricted Boltzmann machines (RBMs, Smolensky, 1986; Hinton, 2002) and deep belief networks (DBNs, Hinton et al., 2006; Hinton & Salakhutdinov, 2006) are probabilistic models with latent and observable variables, which can be interpreted as stochastic neural networks. Binary RBMs, in which each variable conditioned on the others is Bernoulli distributed, are able to approximate arbitrarily well any distribution over the observable variables (Le Roux & Bengio, 2008; Montufar & Ay, 2011). Binary deep belief networks are built by layering binary RBMs, and the representational power does not decrease by adding layers (Le Roux & Bengio, 2008; Montufar & Ay, 2011). In fact, it can be shown that a binary DBN never needs more variables than a binary RBM to model a distribution with a certain accuracy (Le Roux & Bengio, 2010).

However, arguably the most prominent applications in recent times involving RBMs consider models in which the visible variables are real-valued (e.g., Salakhutdinov & Hinton, 2007; Lee et al., 2009; Taylor et al., 2010; Le Roux et al., 2011). Welling et al. (2005) proposed a notion of RBMs where the conditional distributions of the observable variables given the latent variables and vice versa are (almost) arbitrarily chosen from the exponential family. This includes the important special case of the Gaussian-binary RBM (GB-RBM, also Gaussian-Bernoulli RBM), an RBM with binary hidden and Gaussian visible variables.

Despite their frequent use, little is known about the

approximation capabilities of RBMs and DBNs modeling continuous distributions. Clearly, orchestrating a set of Bernoulli distributions to model a distribution over binary vectors is easy compared to approximating distributions over $\Omega \subseteq \mathbb{R}^m$. Recently, Wang et al. (2012) have emphasized that the distribution of the visible variables represented by a GB-RBM with n hidden units is a mixture of 2^n Gaussian distributions with means lying on the vertices of a projected n-dimensional hyperparallelotope. This limited flexibility makes modeling even a mixture of a finite number of Gaussian distributions with a GB-RBM difficult.

This work is a first step towards understanding the representational power of DBNs with binary latent and real-valued visible variables. We will show for a subset of distributions relevant in practice that DBNs with two layers of binary hidden units and a fixed family of conditional distribution for the visible units can model finite mixtures of that family arbitrarily well. As this also holds for infinite mixtures of Gaussians with fixed variance restricted to a compact domain, our results imply universal approximation of strictly positive densities over compact sets.

2. Background

This section will recall basic results on approximation properties of mixture distributions and binary RBMs. Furthermore, the considered models will be defined.

2.1. Mixture distributions

A mixture distribution $p_{\text{mix}}(\boldsymbol{v})$ over Ω is a convex combination of simpler distributions which are members of some family G of distributions over Ω parameterized by $\boldsymbol{\theta} \in \Theta$. We define $\text{MIX}(n, G) = \{\sum_{i=1}^n p_{\text{mix}}(\boldsymbol{v}|i)p_{\text{mix}}(i) \mid \sum_{i=1}^n p_{\text{mix}}(i) = 1 \text{ and } \forall i \in \{1, \dots, n\} : p_{\text{mix}}(i) \geq 0 \wedge p_{\text{mix}}(\boldsymbol{v}|i) \in G\}$ as the family of mixtures of n distributions from G. Furthermore, we denote the family of infinite mixtures of distributions from G as $\text{CONV}(G) = \{\int_\Theta p(\boldsymbol{v}|\boldsymbol{\theta})p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \mid \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} = 1 \text{ and } \forall \boldsymbol{\theta} \in \Theta : p(\boldsymbol{\theta}) \geq 0 \wedge p(\boldsymbol{v}|\boldsymbol{\theta}) \in G\}$.

Li & Barron have shown that for some family of distributions G every element from $\text{CONV}(G)$ can be approximated arbitrarily well by finite mixtures with respect to the Kullback-Leibler divergence (KL-divergence):

Theorem 1 (Li & Barron, 2000)**.** *Let $f \in CONV(G)$. There exists a finite mixture $p_{mix} \in MIX(n, G)$ such that*

$$\text{KL}(f\|p_{mix}) \leq \frac{c_f^2 \gamma}{n} \ ,$$

where

$$c_f^2 = \int_\Omega \frac{\int f^2(\boldsymbol{v}|\boldsymbol{\theta}) f(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int f(\boldsymbol{v}|\boldsymbol{\theta}) f(\boldsymbol{\theta})\, d\boldsymbol{\theta}}\, d\boldsymbol{v}$$

and $\gamma = 4[\log(3\sqrt{e}) + a]$ with

$$a = \sup_{\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \boldsymbol{v}} \log \frac{f(\boldsymbol{v}|\boldsymbol{\theta_1})}{f(\boldsymbol{v}|\boldsymbol{\theta_2})} \ .$$

The bound is not necessarily finite. However, it follows from previous results by Zeevi & Meir (1997) that for every f and every $\epsilon > 0$ there exists a mixture p_{mix} with n components such that $\text{KL}(f\|p_{\text{mix}}) \leq \epsilon + \frac{c}{n}$ for some constant c if $\Omega \subset \mathbb{R}^m$ is a compact set and f is continuous and bounded from below by some $\eta > 0$ (i.e, $\forall x \in \Omega : f(x) \geq \eta > 0$).

Furthermore, it follows that for compact $\Omega \subset \mathbb{R}^m$ every continuous density f on Ω can be approximated arbitrarily well by an infinite but countable mixture of Gaussian distributions with fixed variance σ^2 and means restricted to Ω, that is, by a mixture of distributions from the family

$$G_\sigma(\Omega) = \Bigg\{ p(\boldsymbol{x}) = \frac{1}{(2\pi\sigma^2)^{m/2}} \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{\mu}\|^2}{2\sigma^2}\right) \Big| \boldsymbol{x}, \boldsymbol{\mu} \in \Omega \Bigg\} \ , \quad (1)$$

for sufficient small σ.

2.2. Restricted Boltzmann Machines

An RBM is an undirected graphical model with a bipartite structure (Smolensky, 1986; Hinton, 2002) consisting of one layer of m visible variables $\boldsymbol{V} = (V_1, \dots, V_m) \in \Omega$ and one layer of n hidden variables $\boldsymbol{H} = (H_1, \dots, H_n) \in \Lambda$. The modeled joint distribution is a Gibbs distribution $p(\boldsymbol{v}, \boldsymbol{h}) = \frac{1}{Z} e^{-\mathcal{E}(\boldsymbol{v}, \boldsymbol{h})}$ with energy $\mathcal{E}$ and normalization constant $Z = \int_\Omega \int_\Lambda e^{-\mathcal{E}(\boldsymbol{v}, \boldsymbol{h})}\, \mathrm{d}\boldsymbol{h}\, \mathrm{d}\boldsymbol{v}$, where the variables of one layer are mutually independent given the state of the other layer.

2.2.1. Binary-Binary-RBMs

In the standard binary RBMs the state spaces of the variables are $\Omega = \{0, 1\}^m$ and $\Lambda = \{0, 1\}^n$. The energy is given by $\mathcal{E}(\boldsymbol{v}, \boldsymbol{h}) = -\boldsymbol{v}^T \boldsymbol{W} \boldsymbol{h} - \boldsymbol{v}^T \boldsymbol{b} - \boldsymbol{c}^T \boldsymbol{h}$ with weight matrix $\boldsymbol{W}$ and bias vectors $\boldsymbol{b}$ and $\boldsymbol{c}$.

Le Roux & Bengio showed that binary RBMs are universal approximators for distributions over binary vectors:

Theorem 2 (Le Roux & Bengio, 2008)**.** *Any distribution over $\Omega = \{0, 1\}^m$ can be approximated arbitrarily*

well (with respect to the KL-divergence) with an RBM with $k+1$ hidden units, where k is the number of input vectors whose probability is not zero.

The number of hidden neurons required can be reduced to the minimum number of pairs of input vectors differing in only one component with the property that their union contains all observable patterns having positive probability (Montufar & Ay, 2011).

2.2.2. Exponential-Family RBMs

Welling et al. (2005) introduced a framework for constructing generalized RBMs called exponential family harmoniums. In this framework, the conditional distributions $p(h_i|\boldsymbol{v})$ and $p(v_j|\boldsymbol{h})$, $i = 1,\dots,n$, $j = 1,\dots,m$, belong to the exponential family. Almost all types of RBMs encountered in practice, including binary RBMs, can be interpreted as exponential family harmoniums.

The exponential family is the class $\mathcal{F}$ of probability distributions that can be written in the form

$$p(\boldsymbol{x}) = \frac{1}{Z}\exp\left(\sum_{r=1}^{k}\Phi^{(r)}(\boldsymbol{x})^T\boldsymbol{\mu}^{(r)}(\boldsymbol{\theta})\right) \quad , \qquad (2)$$

where $\boldsymbol{\theta}$ are the parameters of the distribution and Z is the normalization constant.[1] The functions $\Phi^{(r)}$ and $\boldsymbol{\mu}^{(r)}$, for $r = 1,\dots,k$, transform the sample space and the distribution parameters, respectively. Let $\mathcal{I}$ be the subset of $\mathcal{F}$ where the components of $\boldsymbol{x} = (x_1,\dots,x_m)$ are independent from each other, that is, $\mathcal{I} = \{p \in \mathcal{F} \mid \forall \boldsymbol{x} : p(x_1,\dots,x_m) = p(x_1)p(x_2)\cdots p(x_m)\}$. For elements of $\mathcal{I}$ the function $\Phi^{(r)}$ can be written as $\Phi^{(r)}(\boldsymbol{x}) = (\phi_1^{(r)}(x_1),\dots,\phi_m^{(r)}(x_m))$. A prominent subset of $\mathcal{I}$ is the family of Gaussian distributions with fixed variance σ^2, $G_\sigma(\Omega) \subset \mathcal{I}$, see equation (1).

Following Welling et al., the energy of an RBM with binary hidden units and visible units with $p(\boldsymbol{v}|\boldsymbol{h}) \in \mathcal{I}$ is given by

$$\mathcal{E}(\boldsymbol{v},\boldsymbol{h}) = -\sum_{r=1}^{k}\Phi^{(r)}(\boldsymbol{v})^T\boldsymbol{W}^{(r)}\boldsymbol{h} - \sum_{r=1}^{k}\Phi^{(r)}(\boldsymbol{v})^T\boldsymbol{b}^{(r)} - \boldsymbol{c}^T\boldsymbol{h} \quad , \qquad (3)$$

where $\Phi^{(r)}(\boldsymbol{v}) = (\phi_1^{(r)}(x_1),\dots,\phi_m^{(r)}(x_m))$. Note that not every possible choice of parameters necessarily leads to a finite normalization constant and thus to a proper distribution.

[1] By setting $k = 1$ and rewriting Φ and $\boldsymbol{\mu}$ accordingly, one obtains the standard formulation.

If the joint distribution is properly defined, the conditional probability of the visible units given the hidden is

$$p(\boldsymbol{v}|\boldsymbol{h}) = \frac{1}{Z_{\boldsymbol{h}}}\exp\left(\sum_{r=1}^{k}\Phi^{(r)}(\boldsymbol{v})^T\left(\boldsymbol{W}^{(r)}\boldsymbol{h} + \boldsymbol{b}^{(r)}\right)\right) \quad , \qquad (4)$$

where $Z_{\boldsymbol{h}}$ is the corresponding normalization constant. Thus, the marginal distribution of the visible units $p(\boldsymbol{v})$ can be expressed as a mixture of 2^n conditional distributions:

$$p(\boldsymbol{v}) = \sum_{\boldsymbol{h}\in\{0,1\}^n} p(\boldsymbol{v}|\boldsymbol{h})p(\boldsymbol{h}) \in \mathrm{MIX}(2^n,\mathcal{I})$$

2.3. Deep Belief Networks

A DBN is a graphical model with more than two layers built by stacking RBMs (Hinton et al., 2006; Hinton & Salakhutdinov, 2006). A DBN with two layers of hidden variables $\boldsymbol{H}$ and $\hat{\boldsymbol{H}}$ and a visible layer $\boldsymbol{V}$ is characterized by a probability distribution $p(\boldsymbol{v},\boldsymbol{h},\hat{\boldsymbol{h}})$ that fulfills

$$\begin{aligned} p(\boldsymbol{v},\boldsymbol{h},\hat{\boldsymbol{h}}) &= p(\boldsymbol{v}|\boldsymbol{h})p(\boldsymbol{h},\hat{\boldsymbol{h}}) \\ &= p(\hat{\boldsymbol{h}}|\boldsymbol{h})p(\boldsymbol{v},\boldsymbol{h}) \quad . \end{aligned}$$

In this study we are interested in the approximation properties of DBNs with two binary hidden layers and real-valued visible neurons. We will refer to such a DBN as a *B-DBN*. With *B-DBN(G)* we denote the family of all B-DBNs having conditional distributions $p(\boldsymbol{v}|\boldsymbol{h}) \in G$ for all $\boldsymbol{h} \in \boldsymbol{H}$.

3. Approximation properties

This section will present our results on the approximation properties of DBNs with binary hidden units and real-valued visible units. It consists of the following steps:

- Lemma 3 gives an upper bound on the KL-divergence between a B-DBN and a finite additive mixture model – however, under the assumption that the B-DBN "contains" the mixture components. For mixture models from a subset of $\mathcal{I}$, Lemma 4 and Theorem 5 show that such B-DBNs actually exist and that the KL-divergence can be made arbitrarily small.
- Corollary 6 specifies the previous theorem for the special case of Gaussian mixtures, showing how the bound can be applied to distributions used in practice.

- Finally, Theorem 7 generalizes the results to infinite mixture distributions, and thus to the approximation of arbitrary strictly positive densities on a compact set.

3.1. Finite mixtures

We first introduce a construction that will enable us to model mixtures of distributions by DBNs. For some family G an arbitrary mixture of distributions $p_{\text{mix}}(\boldsymbol{v}) = \sum_{i=1}^{n} p_{\text{mix}}(\boldsymbol{v}|i)p_{\text{mix}}(i) \in \text{MIX}(n, G)$ over $\boldsymbol{v} \in \Omega$ can be expressed in terms of a joint probability distribution of $\boldsymbol{v}$ and $\boldsymbol{h} \in \{0,1\}^n$ by defining the distribution

$$q_{\text{mix}}(\boldsymbol{h}) = \begin{cases} p_{\text{mix}}(i), & \text{if } \boldsymbol{h} = \boldsymbol{e}_i, \\ 0, & \text{else} \end{cases} \tag{5}$$

over $\{0,1\}^n$, where $\boldsymbol{e}_i$ is the ith unit vector. Then we can rewrite $p_{\text{mix}}(\boldsymbol{v})$ as $p_{\text{mix}}(\boldsymbol{v}) = \sum_{\boldsymbol{h}} q_{\text{mix}}(\boldsymbol{v}|\boldsymbol{h})q_{\text{mix}}(\boldsymbol{h})$, where $q_{\text{mix}}(\boldsymbol{v}|\boldsymbol{h}) \in G$ for all $\boldsymbol{h} \in \{0,1\}^n$ and $q_{\text{mix}}(\boldsymbol{v}|\boldsymbol{e}_i) = p_{\text{mix}}(\boldsymbol{v}|i)$ for all $i = 1, \ldots, n$. This can be interpreted as expressing $p_{\text{mix}}(\boldsymbol{v})$ as an element of $\text{MIX}(2^n, G)$ with $2^n - n$ mixture components having a probability (or weight) equal to zero. Now we can model $p_{\text{mix}}(\boldsymbol{v})$ by the marginal distribution of the visible variables $p(\boldsymbol{v}) = \sum_{\boldsymbol{h},\hat{\boldsymbol{h}}} p(\boldsymbol{v}|\boldsymbol{h})p(\boldsymbol{h}, \hat{\boldsymbol{h}}) = \sum_{\boldsymbol{h}} p(\boldsymbol{v}|\boldsymbol{h})p(\boldsymbol{h})$ of a B-DBN $p(\boldsymbol{v}, \boldsymbol{h}, \hat{\boldsymbol{h}}) \in \text{B-DBN}(G)$ with the following properties:

1. $p(\boldsymbol{v}|\boldsymbol{e}_i) = p_{\text{mix}}(\boldsymbol{v}|i)$ for $i = 1, \ldots, n$ and
2. $p(\boldsymbol{h}) = \sum_{\hat{\boldsymbol{h}}} p(\boldsymbol{h}, \hat{\boldsymbol{h}})$ approximates $q_{\text{mix}}(\boldsymbol{h})$.

Following this line of thoughts we can formulate our first result. It provides an upper bound on the KL-divergence of any element from $\text{MIX}(n, G)$ and the marginal distribution of the visible variables of a B-DBN with the properties stated above, where $p(\boldsymbol{h})$ models $q_{\text{mix}}(\boldsymbol{h})$ with an approximation error smaller than a given ϵ.

Lemma 3. *Let $p_{mix}(\boldsymbol{v}) = \sum_{i=1}^{n} p_{mix}(\boldsymbol{v}|i)p_{mix}(i) \in MIX(n, G)$ be a mixture with n components from a family of distributions G, and $q_{mix}(\boldsymbol{h})$ be defined as in (5). Let $p(\boldsymbol{v}, \boldsymbol{h}, \hat{\boldsymbol{h}}) \in$ B-DBN(G) with the properties $p(\boldsymbol{v}|\boldsymbol{e}_i) = p_{mix}(\boldsymbol{v}|i)$ for $i = 1, \ldots, n$ and $\forall \boldsymbol{h} \in \{0,1\}^n : |p(\boldsymbol{h}) - q_{mix}(\boldsymbol{h})| < \epsilon$ for some $\epsilon > 0$. Then the KL-divergence between p_{mix} and p is bounded by*

$$\text{KL}(p\|p_{mix}) \leq \mathcal{B}(G, p_{mix}, \epsilon) \ ,$$

where

$$\mathcal{B}(G, p_{mix}, \epsilon) = \epsilon \int_{\Omega} \alpha(\boldsymbol{v})\beta(\boldsymbol{v}) \, d\boldsymbol{v} + 2^n(1+\epsilon)\log(1+\epsilon)$$

with

$$\alpha(\boldsymbol{v}) = \sum_{\boldsymbol{h}} p(\boldsymbol{v}|\boldsymbol{h})$$

and

$$\beta(\boldsymbol{v}) = \log\left(1 + \frac{\alpha(\boldsymbol{v})}{p_{mix}(\boldsymbol{v})}\right) \ .$$

Proof. Using $|p(\boldsymbol{h}) - q_{\text{mix}}(\boldsymbol{h})| < \epsilon$ for all $\boldsymbol{h} \in \{0,1\}^n$ and $p_{\text{mix}}(\boldsymbol{v}) = \sum_{\boldsymbol{h}} p(\boldsymbol{v}|\boldsymbol{h})q_{\text{mix}}(\boldsymbol{h})$ we can write

$$\begin{aligned} p(\boldsymbol{v}) &= \sum_{\boldsymbol{h}} p(\boldsymbol{v}|\boldsymbol{h})p(\boldsymbol{h}) \\ &= \sum_{\boldsymbol{h}} p(\boldsymbol{v}|\boldsymbol{h})(q_{\text{mix}}(\boldsymbol{h}) + p(\boldsymbol{h}) - q_{\text{mix}}(\boldsymbol{h})) \\ &= p_{\text{mix}}(\boldsymbol{v}) + \sum_{\boldsymbol{h}} p(\boldsymbol{v}|\boldsymbol{h})(p(\boldsymbol{h}) - q_{\text{mix}}(\boldsymbol{h})) \\ &\leq p_{\text{mix}}(\boldsymbol{v}) + \alpha(\boldsymbol{v})\epsilon \ , \end{aligned}$$

where $\alpha(\boldsymbol{v})$ is defined as above. Thus, we get for the KL-divergence

$$\begin{aligned} \text{KL}(p\|p_{\text{mix}}) &= \int_{\Omega} p(\boldsymbol{v}) \log\left(\frac{p(\boldsymbol{v})}{p_{\text{mix}}(\boldsymbol{v})}\right) \mathrm{d}\boldsymbol{v} \\ &\leq \int_{\Omega} \underbrace{(p_{\text{mix}}(\boldsymbol{v}) + \alpha(\boldsymbol{v})\epsilon) \log\left(\frac{p_{\text{mix}}(\boldsymbol{v}) + \alpha(\boldsymbol{v})\epsilon}{p_{\text{mix}}(\boldsymbol{v})}\right)}_{=F(\epsilon,\boldsymbol{v})} \mathrm{d}\boldsymbol{v} \\ &= \int_{\Omega} \int_{0}^{\epsilon} \frac{\partial}{\partial\bar{\epsilon}} F(\bar{\epsilon}, \boldsymbol{v}) \, \mathrm{d}\bar{\epsilon} \, \mathrm{d}\boldsymbol{v} \end{aligned}$$

using $F(0, \boldsymbol{v}) = 0$. Because $1 + x\epsilon \leq (1+x)(1+\epsilon)$ for all $x, \epsilon \geq 0$, we can upper bound $\frac{\partial}{\partial\epsilon} F(\epsilon, \boldsymbol{v})$ by

$$\begin{aligned} \frac{\partial}{\partial\epsilon} F(\epsilon, \boldsymbol{v}) &= \alpha(\boldsymbol{v}) \left[1 + \log\left(1 + \frac{\alpha(\boldsymbol{v})}{p_{\text{mix}}(\boldsymbol{v})}\epsilon\right)\right] \\ &\leq \alpha(\boldsymbol{v}) \left[1 + \log\left((1 + \frac{\alpha(\boldsymbol{v})}{p_{\text{mix}}(\boldsymbol{v})})(1+\epsilon)\right)\right] \\ &= \alpha(\boldsymbol{v}) \left[1 + \beta(\boldsymbol{v}) + \log(1+\epsilon)\right] \end{aligned}$$

with $\beta(\boldsymbol{v})$ as defined above. By integration we get

$$\begin{aligned} F(\epsilon, \boldsymbol{v}) &= \int_0^{\epsilon} \frac{\partial}{\partial\bar{\epsilon}} F(\bar{\epsilon}, \boldsymbol{v}) \, \mathrm{d}\bar{\epsilon} \\ &\leq \alpha(\boldsymbol{v})\beta(\boldsymbol{v})\epsilon + \alpha(\boldsymbol{v})(1+\epsilon)\log(1+\epsilon) \ . \end{aligned}$$

Integration with respect to $\boldsymbol{v}$ completes the proof. □

The proof does not use the independence properties of $p(\boldsymbol{v}|\boldsymbol{h})$. Thus, it is possible to apply this bound also to mixture distributions which do not have conditionally independent variables. However, in this case one has to

show that a generalization of the B-DBN exists which can model the target distribution, as the formalism introduced in formula (3) does not cover distributions which are not in $\mathcal{I}$.

For a family $G \subseteq \mathcal{I}$ it is possible to construct a B-DBN with the properties required in Lemma 3 under weak technical assumptions. The assumptions hold for families of distributions used in practice, for instance Gaussian and truncated exponential distributions.

Lemma 4. *Let $G \subset \mathcal{I}$ and $p_{mix}(\boldsymbol{v}) = \sum_{i=1}^{n} p_{mix}(\boldsymbol{v}|i)p_{mix}(i) \in MIX(n, G)$ with*

$$p_{mix}(\boldsymbol{v}|i) = \frac{1}{Z_i} \exp\left(\sum_{r=1}^{k} \Phi^{(r)}(\boldsymbol{v})^T \boldsymbol{\mu}^{(r)}(\boldsymbol{\theta}^{(i)})\right) \quad (6)$$

for $i = 1, \ldots, n$ and corresponding parameters $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(n)}$. Let the distribution $q_{mix}(\boldsymbol{h})$ be defined by equation (5). Assume that there exist parameters $\boldsymbol{b}^{(r)}$ such that for all $\boldsymbol{c} \in \mathbb{R}^n$ the joint distribution $p(\boldsymbol{v}, \boldsymbol{h})$ of $\boldsymbol{v} \in \mathbb{R}^m$ and $\boldsymbol{h} \in \{0,1\}^n$ with energy

$$\mathcal{E}(\boldsymbol{v}, \boldsymbol{h}) = \sum_{r=1}^{k} \Phi^{(r)}(\boldsymbol{v})^T \left(\boldsymbol{W}^{(r)}\boldsymbol{h} + \boldsymbol{b}^{(r)}\right) + \boldsymbol{c}^T\boldsymbol{h}$$

is a proper distribution (i.e., the corresponding normalization constant is finite), where the ith column of $\boldsymbol{W}^{(r)}$ is $\boldsymbol{\mu}^{(r)}(\boldsymbol{\theta}^{(i)}) - \boldsymbol{b}^{(r)}$. Then the following holds:

For all $\epsilon > 0$ there exists a B-DBN with joint distribution $p(\boldsymbol{v}, \boldsymbol{h}, \hat{\boldsymbol{h}}) = p(\boldsymbol{v}|\boldsymbol{h})p(\boldsymbol{h}, \hat{\boldsymbol{h}}) \in$ B-DBN(G) such that

i) $p_{mix}(\boldsymbol{v}|i) = p(\boldsymbol{v}|\boldsymbol{e}_i)$ for $i = 1, \ldots, n$ and

ii) $\forall \boldsymbol{h} \in \{0,1\}^n : |p(\boldsymbol{h}) - q_{mix}(\boldsymbol{h})| < \epsilon$.

Proof. Property i) follows from equation (4) by setting $\boldsymbol{h} = \boldsymbol{e}_i$ and the ith column of $\boldsymbol{W}^{(r)}$ to $\boldsymbol{\mu}^{(r)}(\boldsymbol{\theta}^{(i)}) - \boldsymbol{b}^{(r)}$. Property ii) follows directly from applying Theorem 2 to p.

□

For some families of distributions, such as truncated exponential or Gaussian distributions with uniform variance, choosing $\boldsymbol{b}^{(r)} = \boldsymbol{0}$ for $r = 1, \ldots, k$ is sufficient to yield a proper joint distribution $p(\boldsymbol{v}, \boldsymbol{h})$ and thus a B-DBN with the desired properties. If such a B-DBM exists, one can show, under weak additional assumptions on $G \subset \mathcal{I}$, that the bound shown in Lemma 3 is finite. It follows that the bound decreases to zero as ϵ does.

Theorem 5. *Let $G \subset \mathcal{I}$ be a family of densities and $p_{mix}(\boldsymbol{v}) = \sum_{i=1}^{n} p_{mix}(\boldsymbol{v}|i)p_{mix}(i) \in MIX(n, G)$ with $p_{mix}(\boldsymbol{v}|i)$ given by equation (6). Furthermore, let $q_{mix}(\boldsymbol{h})$ be given by equation (5) and let $p(\boldsymbol{v}, \boldsymbol{h}, \hat{\boldsymbol{h}}) \in$ B-DBN(G) with*

(i) $p_{mix}(\boldsymbol{v}|i) = p(\boldsymbol{v}|\boldsymbol{e}_i)$ for $i = 1, \ldots, n$

(ii) $\forall \boldsymbol{h} \in \{0,1\}^n : |p(\boldsymbol{h}) - q_{mix}(\boldsymbol{h})| < \epsilon$

(iii) $\forall \boldsymbol{h} \in \{0,1\}^n : \int_{\Omega} p(\boldsymbol{v}|\boldsymbol{h}) \|\Phi^{(r)}(\boldsymbol{v})\|_1 \, d\boldsymbol{v} < \infty$.

Then $\mathcal{B}(G, p_{mix}, \epsilon)$ is finite and thus in $O(\epsilon)$.

Proof. We have to show that under the conditions given above $\int_{\Omega} \alpha(\boldsymbol{v})\beta(\boldsymbol{v}) \, \mathrm{d}\boldsymbol{v}$ is finite.

We will first find an upper bound for $\beta(\boldsymbol{v}) = \log\left(1 + \frac{\alpha(\boldsymbol{v})}{p_{\mathrm{mix}}(\boldsymbol{v})}\right)$ for an arbitrary but fixed $\boldsymbol{v}$. Since $p_{\mathrm{mix}}(\boldsymbol{v}) = \sum_i p_{\mathrm{mix}}(\boldsymbol{v}|i)p_{\mathrm{mix}}(i)$ is a convex combination, by defining $i^* = \arg\min_i p_{\mathrm{mix}}(\boldsymbol{v}|i)$ and $\boldsymbol{h}^* = \arg\max_{\boldsymbol{h}} p(\boldsymbol{v}|\boldsymbol{h})$ we get

$$\frac{\alpha(\boldsymbol{v})}{p_{\mathrm{mix}}(\boldsymbol{v})} = \sum_{\boldsymbol{h}} \frac{p(\boldsymbol{v}|\boldsymbol{h})}{\sum_i p_{\mathrm{mix}}(\boldsymbol{v}|i)p_{\mathrm{mix}}(i)} \leq 2^n \frac{p(\boldsymbol{v}|\boldsymbol{h}^*)}{p_{\mathrm{mix}}(\boldsymbol{v}|i^*)} \ . \quad (7)$$

The conditional distribution $p_{\mathrm{mix}}(\boldsymbol{v}|i)$ of the mixture can be written as in equation (6) and the conditional distribution $p(\boldsymbol{v}|\boldsymbol{h})$ of the RBM can be written as in formula (4). We define

$$\boldsymbol{u}^{(r)}(\boldsymbol{h}) = \boldsymbol{W}^{(r)}\boldsymbol{h} + \boldsymbol{b}^{(r)}$$

and get

$$\begin{aligned}
\frac{p(\boldsymbol{v}|\boldsymbol{h}^*)}{p_{\mathrm{mix}}(\boldsymbol{v}|i^*)} &= \frac{\exp\left(\sum_{r=1}^{k} \Phi^{(r)}(\boldsymbol{v})^T \boldsymbol{u}^{(r)}(\boldsymbol{h}^*)\right)}{\exp\left(\sum_{r=1}^{k} \Phi^{(r)}(\boldsymbol{v})^T \boldsymbol{\mu}^{(r)}(\boldsymbol{\theta}^{(i^*)})\right)} \\
&= \exp\left(\sum_{r=1}^{k} \Phi^{(r)}(\boldsymbol{v})^T \left[\boldsymbol{u}^{(r)}(\boldsymbol{h}^*) - \boldsymbol{\mu}^{(r)}(\boldsymbol{\theta}^{(i^*)})\right]\right) \\
&\leq \exp\left(\sum_{r=1}^{k}\sum_{j=1}^{m} \left|\phi_j^{(r)}(\boldsymbol{v})\right| \cdot \left|u_j^{(r)}(\boldsymbol{h}^*) - \mu_j^{(r)}(\boldsymbol{\theta}^{(i^*)})\right|\right) \ .
\end{aligned}$$

Note that the last expression is always larger or equal to one. We can further bound this term by defining

$$\xi^{(r)} = \max_{j,\boldsymbol{h},i} \left|u_j^{(r)}(\boldsymbol{h}^*) - \mu_j^{(r)}(\boldsymbol{\theta}^{(i)})\right|$$

and arrive at

$$\frac{p(\boldsymbol{v}|\boldsymbol{h}^*)}{p_{\mathrm{mix}}(\boldsymbol{v}|i^*)} \leq \exp\left(\sum_{r=1}^{k} \xi^{(r)} \|\Phi^{(r)}(\boldsymbol{v})\|_1\right) \ . \quad (8)$$

By plugging these results into the formula for $\beta(\boldsymbol{v})$ we obtain

$$\begin{aligned}\beta(\boldsymbol{v}) &\overset{(7)}{\leq} \log\left[1+2^n\frac{p(\boldsymbol{v}|\boldsymbol{h}^*)}{p_{\text{mix}}(\boldsymbol{v}|i^*)}\right]\\ &\overset{(8)}{\leq} \log\left[1+2^n\exp\left(\sum_{r=1}^k \xi^{(r)}\|\Phi^{(r)}(\boldsymbol{v})\|_1\right)\right]\\ &\leq \log\left[2^{n+1}\exp\left(\sum_{r=1}^k \xi^{(r)}\|\Phi^{(r)}(\boldsymbol{v})\|_1\right)\right]\\ &= (n+1)\log(2)+\sum_{r=1}^k \xi^{(r)}\|\Phi^{(r)}(\boldsymbol{v})\|_1 \ .\end{aligned}$$

In the third step, we used that the second term is always larger than 1. Insertion into $\int_\Omega \alpha(\boldsymbol{v})\beta(\boldsymbol{v})\,\mathrm{d}\boldsymbol{v}$ leads to

$$\begin{aligned}&\int_\Omega \alpha(\boldsymbol{v})\beta(\boldsymbol{v})\,\mathrm{d}\boldsymbol{v}\\ &\leq \int_\Omega \alpha(\boldsymbol{v})\left[(n+1)\log(2)+\sum_{r=1}^k \xi^{(r)}\|\Phi^{(r)}(\boldsymbol{v})\|_1\right]\mathrm{d}\boldsymbol{v}\\ &= 2^n(n+1)\log(2)\\ &\quad+\sum_{\boldsymbol{h}}\sum_{r=1}^k \xi^{(r)}\int_\Omega p(\boldsymbol{v}|\boldsymbol{h})\|\Phi^{(r)}(\boldsymbol{v})\|_1\,\mathrm{d}\boldsymbol{v} \ , \quad (9)\end{aligned}$$

which is finite by assumption. □

3.2. Finite Gaussian mixtures

Now we apply Lemma 4 and Theorem 5 to mixtures of Gaussian distributions with uniform variance.

The KL-divergence is continuous for strictly positive distributions. Our previous results thus imply that for every mixture p_{mix} of Gaussian distributions with uniform variance and every $\delta \geq 0$ we can find a B-DBN p such that $\mathrm{KL}(p\|p_{\text{mix}}) \leq \delta$. The following corollary gives a corresponding bound:

Corollary 6. *Let $\Omega = \mathbb{R}^m$ and $G_\sigma(\Omega)$ be the family of Gaussian distributions with variance σ^2. Let $\epsilon > 0$ and $p_{mix}(\boldsymbol{v}) = \sum_{i=1}^n p_{mix}(\boldsymbol{v}|i)p_{mix}(i) \in MIX(n, G_\sigma(\Omega))$ a mixture of n distributions with means $\boldsymbol{z}^{(i)} \in \mathbb{R}^m$, $i = 1,\ldots,n$. By*

$$D = \max_{\substack{r,s\in\{1,\ldots,n\}\\ k\in\{1,\ldots,m\}}}\left\{\left|z_k^{(r)} - z_k^{(s)}\right|\right\}$$

we denote the edge length of the smallest hypercube containing all means. Then there exists $p(\boldsymbol{v},\boldsymbol{h},\hat{\boldsymbol{h}}) \in$ B-DBN$(G_\sigma(\Omega))$, with $\forall \boldsymbol{h}\in\{0,1\}^n : |p(\boldsymbol{h}) - q_{mix}(\boldsymbol{h})| <$ ϵ and $p_{mix}(\boldsymbol{v}|i) = p(\boldsymbol{v}|\boldsymbol{e}_i)$, $i = 1,\ldots,n$, such that

$$\begin{aligned}&\mathrm{KL}(p\|p_{mix})\\ &\leq \epsilon\cdot 2^n\left((n+1)\log(2)+m\left(\frac{n^2}{(\sigma/D)^2}+\frac{\sqrt{2}n}{\sqrt{\pi}(\sigma/D)}\right)\right)\\ &\qquad+2^n(1+\epsilon)\log(1+\epsilon) \ .\end{aligned}$$

Proof. In a first step we apply an affine linear transformation to map the hypercube of edge length D to the unit hypercube $[0,1]^m$. Note that doing this while transforming the B-DBN-distribution accordingly does not change the KL-divergence, but it does change the standard deviation of the Gaussians from σ to σ/D. In other words, it suffices to show the above bound for $D = 1$ and $\boldsymbol{z}^{(i)} \in [0,1]^m$.

The energy of the Gaussian-Binary-RBM $p(\boldsymbol{v},\boldsymbol{h})$ is typically written as

$$\mathcal{E}(\boldsymbol{v},\boldsymbol{h}) = \frac{1}{2\sigma^2}\boldsymbol{v}^T\boldsymbol{v} - \frac{1}{\sigma^2}\boldsymbol{v}^T\boldsymbol{b} - \frac{1}{\sigma^2}\boldsymbol{v}^T\boldsymbol{W}\boldsymbol{h} - \boldsymbol{c}^T\boldsymbol{h} \ ,$$

with weight matrix $\boldsymbol{W}$ and bias vectors $\boldsymbol{b}$ and $\boldsymbol{c}$. This can be brought into the form of formula (3) by setting $k = 2$, $\phi_j^{(1)}(v_j) = v_j$, $\phi_j^{(2)}(v_j) = v_j^2$, $\boldsymbol{W}^{(1)} = \boldsymbol{W}/\sigma^2$, $\boldsymbol{W}^{(2)} = \boldsymbol{0}$, $b_j^{(1)} = b_j/\sigma^2$, and $b_j^{(2)} = 1/2\sigma^2$. With $\boldsymbol{b} = \boldsymbol{0}$ (and thus $\boldsymbol{b}^{(1)} = \boldsymbol{0}$), it follows from Lemma 4 that a B-DBN $p(\boldsymbol{v},\boldsymbol{h},\hat{\boldsymbol{h}}) = p(\boldsymbol{v}|\boldsymbol{h})p(\boldsymbol{h},\hat{\boldsymbol{h}})$ with properties (i) and (ii) from Theorem 5 exists.

It remains to show that property (iii) holds. Since the conditional probability factorizes, it suffices to show that (iii) holds for every visible variable individually. The conditional probability of the jth visible neuron of the constructed B-DBN is given by

$$p(v_j|\boldsymbol{h}) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(v_j - z_j(\boldsymbol{h}))^2}{2\sigma^2}\right) \ ,$$

where the mean $z_j(\boldsymbol{h})$ is the jth element of $\boldsymbol{W}\boldsymbol{h}$. Using this, it is easy to see that

$$\int_{-\infty}^{\infty} p(v_j|\boldsymbol{h})|\phi^{(2)}(v_j)|\,\mathrm{d}v_j = \int_{-\infty}^{\infty} p(v_j|\boldsymbol{h})v_j^2\,\mathrm{d}v_j < \infty \ ,$$

because it is the second moment of the normal distri-

bution. For $\int_{-\infty}^{\infty} p(v_j|\boldsymbol{h})|\phi^{(1)}(v_j)|\,\mathrm{d}v_j$ we get

$$\begin{aligned}
&\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}\exp\left(-\frac{(v_j-z_j(\boldsymbol{h}))^2}{2\sigma^2}\right)|v_j|\,\mathrm{d}v_j\\
&=-z_j(\boldsymbol{h})+\frac{2}{\sqrt{2\pi\sigma^2}}\int_{0}^{\infty}\exp\left(-\frac{(v_j-z_j(\boldsymbol{h}))^2}{2\sigma^2}\right)v_j\,\mathrm{d}v_j\\
&=-z_j(\boldsymbol{h})+\frac{2}{\sqrt{2\pi\sigma^2}}\int_{-z_j(\boldsymbol{h})}^{\infty}\exp\left(-\frac{t^2}{2\sigma^2}\right)(t+z_j(\boldsymbol{h}))\mathrm{d}t\\
&=-z_j(\boldsymbol{h})+2z_j(\boldsymbol{h})\int_{0}^{\infty}p(v_j|\boldsymbol{h})\mathrm{d}v_j\\
&\quad+\frac{2}{\sqrt{2\pi\sigma^2}}\int_{-z_j(\boldsymbol{h})}^{\infty}\exp\left(-\frac{t^2}{2\sigma^2}\right)t\,\mathrm{d}t\\
&\leq z_j(\boldsymbol{h})+\frac{\sqrt{2}\sigma}{\sqrt{\pi}}\exp\left(-\frac{z_j^2(\boldsymbol{h})}{2\sigma^2}\right)\leq n+\frac{\sqrt{2}\sigma}{\sqrt{\pi}}\ . \qquad (10)
\end{aligned}$$

In the last step we used that $z_j(\boldsymbol{e}_i)=z_j^{(i)}\in[0,1]$ by construction and thus $z_j(\boldsymbol{h})$ can be bounded from above by

$$z_j(\boldsymbol{h})=\sum_{i=0}^{n}h_i z_j(\boldsymbol{e}_i)\leq n\ . \qquad (11)$$

Thus it follows from Theorem 5 that the bound from Lemma 3 holds and is finite. To get the actual bound, we only need to find the constants $\xi^{(1)}$ and $\xi^{(2)}$ to be inserted into (9). The first constant is given by $\xi^{(1)}=\max_{j,\boldsymbol{h},i}\left|\frac{z_j(\boldsymbol{h})}{\sigma^2}-\frac{z_j^{(i)}}{\sigma^2}\right|$. It can be upper bounded by $\max_{j,\boldsymbol{h}}\frac{z_j(\boldsymbol{h})}{\sigma^2}\leq\frac{n}{\sigma^2}$, as an application of equation (11) shows. The second constant is given by $\xi^{(2)}=\max_{j,\boldsymbol{h},i}|\frac{1}{2\sigma^2}-\frac{1}{2\sigma^2}|=0$. Inserting these variables into inequality (9) leads to the bound. □

The bound $\mathcal{B}(G_\sigma(\Omega),p_{\text{mix}},\epsilon)$ is also finite when Ω is restricted to a compact subset of $\mathbb{R}^m$. This can easily be verified by adapting equation (10) accordingly.

Similar results can be obtained for other families of distributions. A prominent example are B-DBMs with truncated exponential distributions. In this case the energy function of the first layer is the same as for the binary RBM, but the values of the visible neurons are chosen from the interval $[0,1]$ instead of $\{0,1\}$. It is easy to see that for every choice of parameters the normalization constant as well as the bound are finite.

3.3. Infinite mixtures

We will now transfer our results for finite mixtures to the case of infinite mixtures following Li & Barron (2000).

Theorem 7. *Let G be a family of continuous distributions and $f\in CONV(G)$ such that the bound from Theorem 1 is finite for all $p_{mix\text{-}n}\in MIX(n,G)$, $n\in\mathbb{N}$. Furthermore, for all $p_{mix\text{-}n}\in MIX(n,G)$, $n\in\mathbb{N}$, and for all $\hat{\epsilon}>0$ let there exist a B-DBN in B-DBN(G) such that $\mathcal{B}(G,p_{mix},\hat{\epsilon})$ is finite. Then for all $\epsilon>0$ there exists $p(\boldsymbol{v},\boldsymbol{h},\hat{\boldsymbol{h}})\in$ B-DBN(G) with $KL(f\|p)\leq\epsilon$.*

Proof. From Theorem 1 and the assumption that the corresponding bound is finite it follows that for all $\epsilon>0$ there exists a mixture $p_{\text{mix-}n'}\in\text{MIX}(n',G)$ with $n'\geq 2c_f^2\gamma/\epsilon$ such that $\text{KL}(f\|p_{\text{mix-}n'})\leq\frac{\epsilon}{2}$.

By assumption there exists a B-DBN $\in$ B-DBN(G) such that $\mathcal{B}(G,p_{\text{mix-}n'},\hat{\epsilon})$ is finite. Thus, one can define a sequence of B-DBNs $(p_{\hat{\epsilon}})_{\hat{\epsilon}}\in$ B-DBN(G) with $\hat{\epsilon}$ decaying to zero (where the B-DBNs only differ in the weights between the hidden layers) for which it holds $\text{KL}(p_{\hat{\epsilon}}\|p_{\text{mix-}n'})\overset{\hat{\epsilon}\to 0}{\longrightarrow}0$. This implies that $p_{\hat{\epsilon}}\overset{\hat{\epsilon}\to 0}{\longrightarrow}p_{\text{mix-}n'}$ uniformly. It follows $\text{KL}(f\|p_{\hat{\epsilon}})\overset{\hat{\epsilon}\to 0}{\longrightarrow}\text{KL}(f\|p_{\text{mix-}n'})$. Thus, there exists ϵ' such that $|\text{KL}(f\|p_{\epsilon'})-\text{KL}(f\|p_{\text{mix-}n'})|<\epsilon/2$. A combination of these inequalities yields

$$\begin{aligned}
&\text{KL}(f\|p_{\epsilon'})\\
&\leq|\,\text{KL}(f\|p_{\epsilon'})-\text{KL}(f\|p_{\text{mix-}n'})|+\text{KL}(f\|p_{\text{mix-}n'})\leq\epsilon\ .
\end{aligned}$$

□

This result applies to infinite mixtures of Gaussians with the same fixed but arbitrary variance σ^2 in all components. In the limit $\sigma\to 0$ such mixtures can approximate strictly positive densities over compact sets arbitrarily well (Zeevi & Meir, 1997).

4. Conclusions

We presented a step towards understanding the representational power of DBNs for modeling real-valued data. When binary latent variables are considered, DBNs with two hidden layers can already achieve good approximation results. Under mild constraints, we showed that for modeling a mixture of n pairwise independent distributions, a DBN with only $2n+1$ binary hidden units is sufficient to make the KL-divergence between the mixture p_{mix} and the DBN distribution p arbitrarily small (i.e., for every $\delta>0$ we can find a DBN such that $\text{KL}(p\|p_{\text{mix}})<\delta$). This holds for deep architectures used in practice, for instance DBNs having visible neurons with Gaussian or truncated exponential conditional distributions, and corresponding mixture distributions having components of the same type as the visible units of the DBN. Furthermore, we extended these results to infinite mixtures and showed that these can be approximated arbitrarily well by

a DBN with a finite number of neurons. Therefore, Gaussian-binary DBNs inherit the universal approximation properties from additive Gaussian mixtures, which can model any strictly positive density over a compact domain with arbitrarily high accuracy.

Acknowledgments

OK and CI acknowledge support from the Danish National Advanced Technology Foundation through project "Personalized breast cancer screening", AF and CI acknowledge support from the German Federal Ministry of Education and Research within the Bernstein Fokus "Learning behavioral models: From human experiment to technical assistance".

References

Hinton, Geoffrey E. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

Hinton, Geoffrey E. and Salakhutdinov, Ruslan. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

Hinton, Geoffrey E., Osindero, Simon, and Teh, Yee-Whye. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006.

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

Le Roux, Nicolas and Bengio, Yoshua. Deep belief networks are compact universal approximators. *Neural Computation*, 22(8):2192–2207, 2010.

Le Roux, Nicolas, Heess, Nicolas, Shotton, Jamie, and Winn, John M. Learning a generative model of images by factoring appearance and shape. *Neural Computation*, 23(3):593–650, 2011.

Lee, Honglak, Grosse, Roger, Ranganath, Rajesh, and Ng, Andrew Y. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pp. 609–616. ACM, 2009.

Li, Jonathan Q. and Barron, Andrew R. Mixture density estimation. In Solla, Sara A., Leen, Todd K., and Müller, Klaus-Robert (eds.), *Advances in Neural Information Processing Systems 12 (NIPS 1999)*, pp. 279–285. MIT Press, 2000.

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

Salakhutdinov, Ruslan and Hinton, Geoffrey E. Learning a nonlinear embedding by preserving class neighbourhood structure. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 2, pp. 412 – 419, 2007.

Smolensky, Paul. Information processing in dynamical systems: foundations of harmony theory. In Rumelhart, David E. and McClelland, James L. (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1*, pp. 194–281. MIT Press, 1986.

Taylor, Graham W., Fergus, Rob, LeCun, Yann, and Bregler, Christoph. Convolutional learning of spatio-temporal features. In *Computer Vision – ECCV 2010*, volume 6316 of *LNCS*, pp. 140–153. Springer, 2010.

Wang, Nan, Melchior, Jan, and Wiskott, Laurenz. An analysis of Gaussian-binary restricted Boltzmann machines for natural images. In Verleysen, Michel (ed.), *Proceedings of the 20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2012)*, pp. 287–292. Evere, Belgium: d-side publications, 2012.

Welling, Max, Rosen-Zvi, Michal, and Hinton, Geoffrey E. Exponential family harmoniums with an application to information retrieval. In Saul, Lawrence K., Weiss, Yair, and Bottou, Léon (eds.), *Advances in Neural Information Processing Systems 17 (NIPS 2004)*, pp. 1481–1488. MIT Press, 2005.

Zeevi, Assaf J. and Meir, Ronny. Density estimation through convex combinations of densities: Approximation and estimation bounds. *Neural Networks*, 10(1):99 – 109, 1997.