Max-Planck-Institut für Mathematik in den Naturwissenschaften Leipzig

Universal Approximation Depth and Errors of Narrow Belief Networks with Discrete Units

by

Guido Montúfar

Universal Approximation Depth and Errors of Narrow Belief Networks with Discrete Units

Guido F. Montúfar[1]

[1]Department of Mathematics, Pennsylvania State University, University Park, PA 16802, USA.

Abstract

We generalize recent theoretical work on the minimal number of layers of narrow deep belief networks that can approximate any probability distribution on the states of their visible units arbitrarily well. We relax the setting of binary units (Sutskever and Hinton, 2008; Le Roux and Bengio, 2008, 2010; Montúfar and Ay, 2011) to units with arbitrary finite state spaces, and the vanishing approximation error to an arbitrary approximation error tolerance. For example, we show that a q-ary deep belief network with $L \geq 2 + \frac{q^{\lceil m-\delta \rceil}-1}{q-1}$ layers of width $n \leq m + \log_q(m) + 1$ for some $m \in \mathbb{N}$ can approximate any probability distribution on $\{0, 1, \ldots, q-1\}^n$ without exceeding a Kullback-Leibler divergence of δ. Our analysis covers discrete restricted Boltzmann machines and naïve Bayes models as special cases.

1 Introduction

A *deep belief network* (DBN) (Hinton et al., 2006) is a layered stochastic network with undirected bipartite interactions between the units in the top two layers, and directed bipartite interactions between the units in all other subsequent pairs of layers, directed towards the bottom layer. The top two layers form a *restricted Boltzmann machine* (RBM) (Smolensky, 1986). The entire network defines a model of probability distributions on the states of the units in the bottom layer, the *visible* layer. When the number of units in every layer has the same order of magnitude, the network is called *narrow*. The *depth* refers to the number of layers. Deep network architectures are believed to play a key role in information processing of intelligent agents, see (Bengio, 2009) for an overview on this exciting topic. DBNs were the first deep architectures to be envisaged together with an efficient unsupervised training algorithm (Hinton et al., 2006). Due to

their restricted connectivity, it is possible to greedily train their layers one at the time, and in this way, identify remarkably good parameter initializations for solving specific tasks (see Bengio et al., 2007). The ability to train deep architectures efficiently has pioneered a great number of applications in machine learning and in the booming field *deep learning*.

The representational power of neural networks has been studied for several decades, whereby their universal approximation properties have received special attention. For instance, a well known result (Hornik et al., 1989) shows that multilayer feedforward networks with one exponentially large layer of hidden units are universal approximators of Borel measurable functions. Although universal approximation has a limited importance for practical purposes,[1] it plays an important role as warrant for consistency and sufficiency of the complexity attainable by specific classes of learning systems. Besides the universal approximation question, it is natural to ask "how well is a given network able to approximate certain classes of probability distributions?" This note pursues an account on the ability of DBNs to approximate probability distributions.

The first universal approximation result for deep and narrow sigmoid belief networks is due to Sutskever and Hinton (2008). They showed that a narrow sigmoid belief network with $3(2^n - 1) + 1$ layers can represent probability distributions arbitrarily close to any probability distribution on the set of length n binary vectors. Their result shows that not only exponentially wide and shallow networks are universal approximators (Hornik et al., 1989), but also exponentially deep and narrow ones are. Subsequent work has studied the optimality question "how deep is deep enough?," with improved universal approximation depth bounds by Le Roux and Bengio (2010) and later by Montúfar and Ay (2011), which we will discuss below in more detail. These papers focus on the minimal depth of narrow DBN universal approximators with binary units; that is, the number of layers that these networks must have in order to be able to represent probability distributions arbitrarily close to any probability distribution on the states of their visible units. The present note complements that analysis in two ways:

First, instead of asking for the minimal size of universal approximators, we ask for the minimal size of networks that can approximate any distribution to a given error tolerance, treating the universal approximation problem as the special case of zero error tolerance. This analysis gives a theoretical basis on which to balance model accuracy and parameter count. For comparison, universal approximation is a binary property which always requires an exponential number of parameters. As it turns out, our analysis also allows us to estimate the expected value of the model approximation errors incurred when learning classes of distributions, say low-entropy distributions, with networks of given sizes.

Second, we consider networks with finite valued units, called discrete or multinomial DBNs, including binary DBNs as special cases. Non-binary units serve, obviously, to encode non-binary features directly, which may be interesting in multi-channel perception, e.g., color-temperature-distance sensory inputs. Additionally, the interactions between discrete units can carry much richer relations that those between binary units. In particular, within the non-binary discrete setting, DBNs, RBMs, and naïve Bayes

---

[1]Where a more or less good approximation of a small set of target distributions is often sufficient, or where the goal is not to model data directly but rather to obtain abstract representations of data.
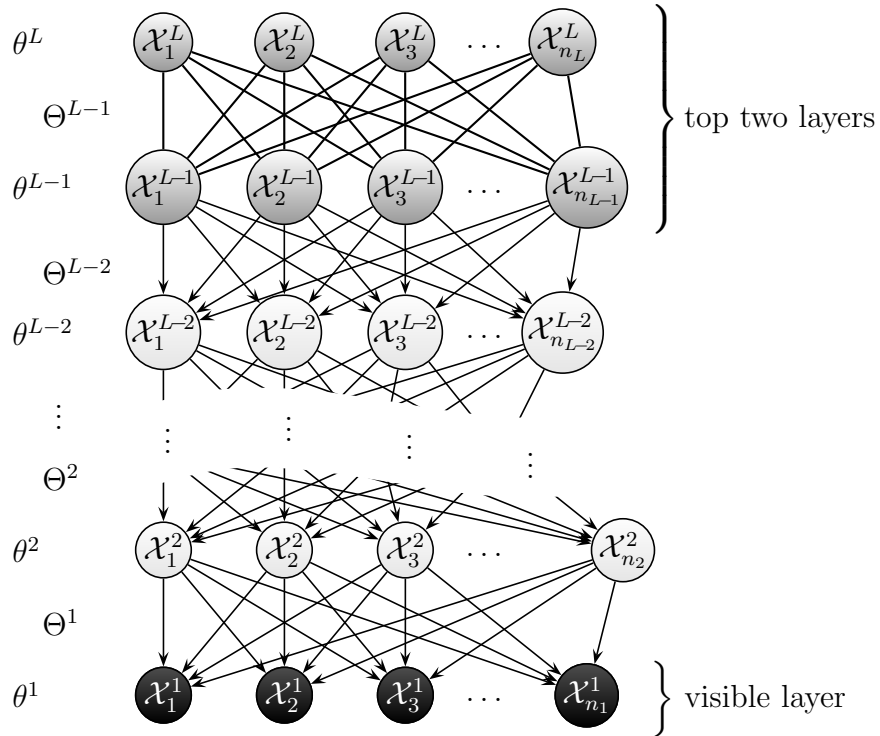
Figure 1: Graphical representation of a discrete DBN probability model. Each node represents a unit with the indicated state space. The top two layers have undirected connections; they correspond to the term $p_{L-1,L}$ described in eq. (2). All other layers receive directed connections, corresponding to the terms p_l, $l \in [L-2]$ described in eq. (3). Only the bottom layer is visible.

models can be seen as representatives of the same class of probability models.

This paper is organized as follows. Section 2 gives formal definitions, before we proceed to state our main result Theorem 2 in Section 3: a bound on the approximation errors of discrete DBNs. A universal approximation depth bound follows directly. After this, a discussion of the result is given, together with a sketch of the proof. The proof entails several steps of independent interest, developed in the next sections. Section 4 addresses the representational power and approximation errors of RBMs with discrete units. Section 5 studies the models of conditional distributions represented by feedforward discrete stochastic networks (DBN layers). Section 6 studies concatenations of layers of feedforward networks and elaborates on the patterns of probability sharing steps (transformations of probability distributions) that they can realize. Section 7 concludes the proof of the main theorem and gives a corollary about the expectation value of the approximation error of DBNs. Section A presents an empirical validation scheme and tests the approximation error bounds numerically on small networks.

2 Preliminaries

A few formal definitions are necessary before proceeding. Given a finite set $\mathcal{X}$, we denote $\Delta(\mathcal{X})$ the set of all probability distributions on $\mathcal{X}$. A *model* of probability distributions on $\mathcal{X}$ is a subset $\mathcal{M} \subseteq \Delta(\mathcal{X})$. Given a pair of distributions $p, q \in \Delta(\mathcal{X})$, the Kullback-Leibler divergence from p to q is defined as $D(p\|q) := \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$ when $\mathrm{supp}(p) \subseteq \mathrm{supp}(q)$, and $D(p\|q) := \infty$ otherwise. The divergence from a distribution p to a model $\mathcal{M} \subseteq \Delta(\mathcal{X})$ is defined as $D(p\|\mathcal{M}) := \inf_{q \in \mathcal{M}} D(p\|q)$. The divergence of any distribution on $\mathcal{X}$ to $\mathcal{M}$ is bounded by

$$D_{\mathcal{M}} := \sup_{p \in \Delta(\mathcal{X})} D(p\|\mathcal{M}).$$

We refer to $D_{\mathcal{M}}$ as the *universal* or *maximal approximation error* of $\mathcal{M}$. The model $\mathcal{M}$ is called a *universal approximator* of probability distributions on $\mathcal{X}$ iff $D_{\mathcal{M}} = 0$.

A discrete DBN probability model is specified by a number of layers (the depth of the network), the number of units in each layer (the width of each layer), and the state space of each unit in each layer. Let $L \in \mathbb{N}$, $L \geq 2$ be the number of layers. We imagine these layers arranged as a stack with layer 1 at the bottom (this will be the visible layer) and layer L at the top (this will be the *deepest* layer). See Figure 1. For each $l \in \{1, \ldots, L\} =: [L]$, let $n_l \in \mathbb{N}$ be the number of units in layer l. For each $i \in [n_l]$, let $\mathcal{X}_i^l$, $|\mathcal{X}_i^l| < \infty$ be the state space of unit i in layer l. We denote the joint state space of the units in layer l by $\mathcal{X}^l = \mathcal{X}_1^l \times \cdots \times \mathcal{X}_{n_l}^l$, and write $x^l = (x_1^l, \ldots, x_{n_l}^l)$ for a state from $\mathcal{X}^l$. We call a unit q-*valued* or q-*ary* if its state space has cardinality q, and assume that q is a finite integer larger than one.

In order to proceed with the definition of the DBN model, we consider the mixed graphical model with undirected connections between the units in the top two layers L and $L-1$, and directed connections from the units in layer $l+1$ to the units in layer l for all $l \in [L-2]$. This model consists of joint probability distributions on the states $\mathcal{X} = \mathcal{X}^1 \times \cdots \mathcal{X}^L$ of all network units, parametrized by a collection of real matrices and vectors $\Theta = \{\Theta^1, \ldots, \Theta^{L-1}, \theta^1, \ldots, \theta^L\}$. For each $l \in [L-1]$, the matrix Θ^l contains the interaction weights between units in layers l and $l+1$. It consists of row blocks $\Theta_i^l \in \mathbb{R}^{(|\mathcal{X}_i^l|-1) \times (\sum_{j \in [n_{l+1}]}(|\mathcal{X}_j^{l+1}|-1))}$ for all $i \in [n_l]$. For each $l \in [L]$, the row vector θ^l contains the bias weights of the units in layer l. It consists of blocks $\theta_i^l \in \mathbb{R}^{|\mathcal{X}_i^l|-1}$ for all $i \in [n_l]$.

Note that the bias of a unit with state space $\mathcal{X}_i^l$ is a vector with $|\mathcal{X}_i^l| - 1$ entries, and the interaction of a pair of units with state spaces $\mathcal{X}_i^l$ and $\mathcal{X}_j^{l+1}$ is described by a matrix of order $(|\mathcal{X}_i^l| - 1) \times (|\mathcal{X}_j^{l+1}| - 1)$. The number of interaction and bias parameters in the entire network adds to $\sum_{l=1}^{L-1} (\sum_{i \in [n_l]}(|\mathcal{X}_i^l| - 1))(1 + \sum_{j \in [n_{l+1}]}(|\mathcal{X}_j^{l+1}| - 1)) + \sum_{i \in [n_L]}(|\mathcal{X}_i^L| - 1)$.

For any choice Θ of these parameters, the corresponding probability distribution on the states of all units is

$$p(x^1, \ldots, x^L; \Theta) = p_{L-1,L}(x^{L-1}, x^L; \Theta^{L-1}, \theta^{L-1}, \theta^L) \prod_{l=1}^{L-2} p_l(x^l | x^{l+1}; \Theta^l, \theta^l)$$

$$\text{for all } (x^1, \ldots, x^L) \in \mathcal{X}^1 \times \cdots \times \mathcal{X}^L; \quad (1)$$

4

where

$$p_{L-1,L}(x, y; \Theta^{L-1}, \theta^{L-1}, \theta^L) = \frac{\exp(\mathbf{x}^\top \Theta^{L-1} \mathbf{y} + \theta^{L-1} \mathbf{x} + \theta^L \mathbf{y})}{Z(\Theta^{L-1}, \theta^{L-1}, \theta^L)}$$

$$\text{for all } (x, y) \in \mathcal{X}^{L-1} \times \mathcal{X}^L; \quad (2)$$

and

$$p_l(x|y; \Theta^l, \theta^l) = \prod_{i \in [n_l]} p_{l,i}(x_i|y; \Theta_i^l, \theta_i^l)$$

$$\text{for all } x \in \mathcal{X}^l \text{ and } y \in \mathcal{X}^{l+1}; \quad \text{for each } l \in [L-2]; \quad (3)$$

with factors given by

$$p_{l,i}(x_i|y; \Theta_i^l, \theta_i^l) = \frac{\exp(\mathbf{x}_i^\top \Theta_i^l \mathbf{y} + \theta_i^l \mathbf{x}_i)}{Z(\Theta_i^l \mathbf{y}, \theta_i^l)} \quad \text{for all } x_i \in \mathcal{X}_i^l \text{ and } y \in \mathcal{X}^{l+1}. \quad (4)$$

Here we use following notation. Given a state vector $x = (x_1, \ldots, x_n)$ of n units with joint state space $\mathcal{X}_1 \times \cdots \times \mathcal{X}_n = \{0, 1, \ldots, q_1-1\} \times \cdots \times \{0, 1, \ldots, q_n-1\}$, $\mathbf{x}$ denotes the x-th column of a minimal matrix of sufficient statistics for the independent distributions of these n units. To make this more concrete, we set $\mathbf{x}$ equal to a column vector with blocks $\mathbf{x}_1, \ldots, \mathbf{x}_n$, where $\mathbf{x}_i = (\delta_{y_i}(x_i))_{y_i \in \mathcal{X}_i \setminus \{0\}}$ is the one-hot representation of x_i without the first entry, for all $i \in [n]$. For example, if $x = (x_1, x_2) = (1, 0) \in \mathcal{X}_1 \times \mathcal{X}_2 = \{0, 1, 2\} \times \{0, 1, 2\}$, then $\mathbf{x} = [\begin{smallmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{smallmatrix}]$, with $\mathbf{x}_1 = [\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}]$ and $\mathbf{x}_2 = [\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}]$.

The function

$$Z(\Theta^{L-1}, \theta^{L-1}, \theta^L) = \sum_{x \in \mathcal{X}^{L-1}, y \in \mathcal{X}^L} \exp(\mathbf{x}^\top \Theta^{L-1} \mathbf{y} + \theta^{L-1} \mathbf{x} + \theta^L \mathbf{y}) \quad (5)$$

normalizes the probability distribution $p_{L-1,L}(\cdot; \Theta^{L-1}, \theta^{L-1}, \theta^L) \in \Delta(\mathcal{X}^{L-1} \times \mathcal{X}^L)$ from eq. (2). Likewise, the function

$$Z(\Theta_i^l \mathbf{y}, \theta_i^l) = \sum_{x_i \in \mathcal{X}_i^l} \exp(\mathbf{x}_i^\top \Theta_i^l \mathbf{y} + \theta_i^l \mathbf{x}) \quad (6)$$

normalizes the probability distribution $p_{l,i}(\cdot|y; \Theta_i^l, \theta_i^l) \in \Delta(\mathcal{X}_i^l)$ from eq. (4) for each $i \in [n_l]$ and $l \in [L-2]$.

The marginal of the distribution $p(\cdot; \Theta) \in \Delta(\mathcal{X}^1 \times \cdots \times \mathcal{X}^L)$ from eq. (1) on the states $\mathcal{X}^1$ of the units in the first layer is given by

$$P(x^1; \Theta) = \sum_{(x^2, \ldots, x^L) \in \mathcal{X}^2 \times \cdots \times \mathcal{X}^L} p(x^1, \ldots, x^L; \Theta) \quad \text{for all } x^1 \in \mathcal{X}^1. \quad (7)$$

The discrete DBN probability model with L layers of widths $n_1, \ldots, n_L$ and state spaces $\mathcal{X}^1, \ldots, \mathcal{X}^L$, is the set of probability distributions $P(\cdot; \Theta) \in \Delta(\mathcal{X}^1)$ expressible by eq. (7) for all possible choices of the parameter Θ. Intuitively, this set is a linear projection of a manifold parametrized by Θ, and may have self-intersections or other singularities.

The discrete DBN probability model with $L = 2$ is a discrete RBM probability model. This model consists of the marginal distributions on $\mathcal{X}^{L-1}$ of the distributions $p_{L-1,L}(\cdot; \Theta^{L-1}, \theta^{L-1}, \theta^L)$ from eq. (2) for all possible choices of $\Theta^{L-1}, \theta^{L-1}$, and θ^L.

When $L > 2$, the distributions on $\mathcal{X}^{L-1}$ defined by the top two DBN layers can be seen as the inputs of the stochastic maps defined by the conditional distributions $p_{L-2}(\cdot|\cdot; \Theta^{L-2}, \theta^{L-2})$ from eq. (3). The outputs of these maps are probability distributions on $\mathcal{X}^{L-2}$ that can be seen as the inputs of the stochastic maps defined by the next lower layer and so forth. The discrete DBN probability model can be seen as the set of images of a discrete RBM probability model by a family of sequences of stochastic maps.

The following simple class of probability models will be useful to study the approximation capabilities of DBN models. Let $\varrho = \{A_1, \ldots, A_N\}$ be a partition of a finite set $\mathcal{X}$. The *partition model* $\mathcal{P}$ with partition ϱ is the set of probability distributions on $\mathcal{X}$ which have constant value on each A_i. Geometrically, this is the simplex with vertices $\mathbb{1}_{A_i}/|A_i|$ for all $i \in [N]$, where $\mathbb{1}_{A_i}$ is the indicator function of A_i. The *coarseness* of $\mathcal{P}$ is $\max_i |A_i|$. Unlike many statistical models, partition models have a well understood Kullback-Leibler divergence. If $\mathcal{P}$ is a partition model of coarseness c, then $D_{\mathcal{P}} = \log(c)$. Furthermore, partition models are known to be optimally approximating exponential families, in the sense that they minimize the universal approximation error among all closures of exponential families of a given dimension (see Rauh, 2013).

3 Main Result

The starting point of our considerations is the following result for binary DBNs:

Theorem 1. *A deep belief network probability model with L layers of binary units of width $n = 2^{k-1} + k$ (for some $k \in \mathbb{N}$) is a universal approximator of probability distributions on $\{0,1\}^n$ whenever $L \geq 1 + 2^{2^{k-1}}$.*

Note that

$$\frac{2^n}{2(n - \log_2(n))} \leq 2^{2^{k-1}} \leq \frac{2^n}{2(n - \log_2(n) - 1)}. \tag{8}$$

This result is due to Montúfar and Ay (2011, Theorem 2). It is based on a refinement of previous work by Le Roux and Bengio (2010), who obtained the bound $L \geq 1 + \frac{2^n}{n}$ when n is a power of two.

The main result of this paper is following generalization of Theorem 1. Here we make the simplifying assumption that all layers have the same width n and the same state space. The result holds automatically for DBNs with wider hidden layers or hidden units with larger state spaces.

Theorem 2. *Let* DBN *be a deep belief network probability model with $L \in \mathbb{N}$, $L \geq 2$ layers of width $n \in \mathbb{N}$. Let the i-th unit of each layer have state space $\{0, 1, \ldots, q_i - 1\}$, $q_i \in \mathbb{N}$, $2 \leq q_i < \infty$, for each $i \in [n]$. Let m be any integer with $n \geq m \geq \prod_{j=m+2}^{n} q_j$, and let $q = q_1 \geq \cdots \geq q_m$. If $L \geq 2 + \frac{q^S - 1}{q - 1}$ for some $S \in \{0, 1, \ldots, m\}$, then the probability model* DBN *can approximate each element of a partition model*
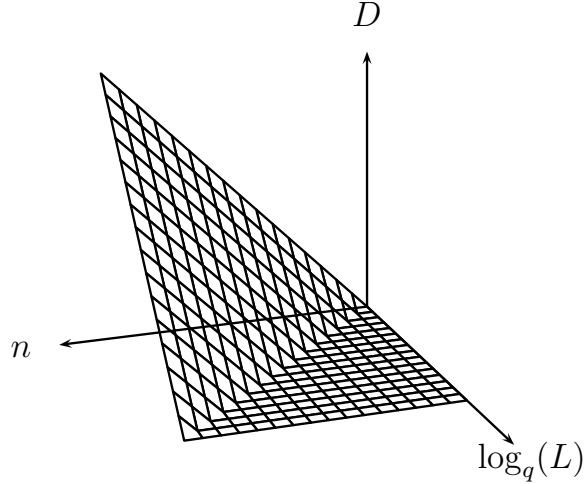
6

Figure 2: Qualitative illustration of Theorem 2. Shown is the large-scale behaviour of the DBN universal approximation error upper-bound as a function D of the layer width n and the logarithm of the number of layers $\log_q(L)$. Here it is assumed that the Kullback-Leibler divergence is computed in base q logarithm and that all units are q-ary. The number of parameters of these DBNs scales with $Ln^2(q-1)^2$.

of coarseness $\prod_{j \in [m-S]} q_j$ arbitrarily well. The Kullback-Leibler divergence from any distribution on $\{0, 1, \ldots, q_1 - 1\} \times \cdots \times \{0, 1, \ldots, q_n - 1\}$ to DBN *is bounded by*

$$D_{\text{DBN}} \leq \log(\prod_{j \in [m-S]} q_j).$$

In particular, this DBN probability model is a universal approximator whenever

$$L \geq 2 + \frac{q^m - 1}{q - 1}.$$

When all units are q-ary and the layer width is $n = q^{k-1} + k$ for some $k \in \mathbb{N}$, then the DBN probability model is a universal approximator of distributions on $\{0, 1, \ldots, q-1\}^n$ whenever $L \geq 2 + \frac{q^{q^{k-1}} - 1}{q - 1}$. Note that

$$\frac{q^n - 1}{q(q-1)(n - \log_q(n))} \leq \frac{q^{q^{k-1}} - 1}{q - 1} \leq \frac{q^n - 1}{q(q-1)(n - \log_q(n) - 1)}. \tag{9}$$

The theorem is illustrated in Figure 2.

Remarks

The number of parameters of a q-ary DBN with L layers of width n is $(L-1)(n(q-1)+1)n(q-1)+n(q-1)$. Since the set of probability distributions on $\{0, 1, \ldots, q-1\}^n$ has dimension $q^n - 1$, the DBN model is full dimensional only if $L \geq \frac{q^n - 1}{n(q-1)(n(q-1)+2)} + 1$. This is a parameter-counting lower bound for the universal approximation depth. Theorem 2 gives an upper bound for the minimal universal approximation depth. The upper

bound from the theorem surpasses the parameter-counting lower bound by roughly a factor n. We think that the upper bound is tight, up to sub-linear factors, in consideration of the following. Probability models with hidden variables can have dimension strictly smaller than their parameter count (dimension defect). Moreover, in some cases even full dimensional models represent only very restricted classes of distributions, as has been observed, for example, in binary tree models with hidden variables. It is known that for any prime power q, the smallest naïve Bayes model universal approximator of distributions on $\{0, 1, \ldots, q-1\}^n$ has $q^{n-1}(n(q-1)+1)-1$ parameters (see Montúfar, 2013, Theorem 13). Hence for these models the number of parameters needed to achieve universal approximation surpasses the corresponding parameter-counting lower bound $q^n/(n(q-1)+1)$ by a factor of order n.

Computing tight bounds for the maximum of the Kullback-Leibler divergence is a notoriously challenging problem. This is even so for simple probability models without hidden variables, like independence models with mixed discrete variables. The optimality of our DBN error bounds is not completely settled at this point, but we think that they give a close description of the large-scale approximation error behaviour of DBNs. For the limiting case of one single layer with n independent q-ary units, it is known that the maximal divergence is equal to $(n-1)\log(q)$ (see Ay and Knauf, 2006), corresponding to the line $\log_q(L) = 0$ in Figure 2. Furthermore, when our upper bounds vanish, they obviously are tight (corresponding to the points with value zero in Figure 2).

Discrete DBNs have many hyperparameters (the layer widths and the state spaces of the units), which makes their analysis combinatorially intricate. Some of these intricacies are apparent from the floor and ceiling functions in our main theorem. This theorem tries to balance accuracy, generality, and clarity. In some cases, the bounds can be improved by exhausting the representational power gain per layer described in Theorem 8. A more detailed and accurate account on the two-layer case (RBMs) is given in Section 4. In Section 7 we give results describing probability distributions contained in the DBN model (Proposition 9) and addressing the expectation value of the divergence (Corollary 11). Section A contains an empirical discussion, together with the numerical evaluation of small models.

Outline of the Proof

We will prove Theorem 2 by first studying the individual parts of the DBN: the RBM formed by the top two layers (Section 4); the individual units with directed inputs (Section 5); the *probability sharing* realized by stacks of layers (Section 6); and finally, the sets of distributions of the units in the bottom layer (Section 7). The proof steps can be summarized as follows:

- Show that the top RBM can approximate any probability distribution with support on a set of the form $\mathcal{X}_1 \times \cdots \mathcal{X}_k \times \underbrace{\{0\} \times \cdots \times \{0\}}_{n}$ arbitrarily well.
  $\underbrace{}_{k+1}$

- For a unit with state space $\mathcal{X}_1$ receiving n directed inputs, show that there is a choice of parameters for which the following holds for each state $h_n \in \mathcal{X}_n$ of the n-th input unit: If the input vector is $(h_1, h_2, \ldots, h_n)$, then the unit outputs h'_1 with probability $p^{h_n}(h'_1)$, where p^{h_n} is an arbitrary distribution on $\mathcal{X}_1$ for all $h_n \in \mathcal{X}_n$.

- Show that there is a sequence of $\frac{q^m-1}{q-1}$ stochastic maps $p(h) \mapsto p(v) = \sum_h p(v|h)p(h)$ each of which superposes nearly qn probability multi-sharing steps, which maps the probability distributions represented by the top RBM to an arbitrary probability distribution on $\mathcal{X}_1 \times \cdots \times \mathcal{X}_n$.

- Show that the DBN approximates certain classes of tractable probability distributions arbitrarily well, and estimate their maximal approximation errors.

The superposition of probability sharing steps is inspired by (Le Roux and Bengio, 2010), together with the refinements of that work devised in (Montúfar and Ay, 2011). By *probability sharing* we refer to the process of transferring an arbitrary amount of probability from a state vector x' to another state vector x''. In contrast to the binary proofs, where each layer superposes about $2n$ sharing steps, here each layer superposes about qn multi-sharing steps, whereby each multi-sharing step transfers probability from one state to $q-1$ states (when the units are q-ary). With this, a more general treatment of models of conditional distributions is required. Further, additional considerations are required in order to derive tractable submodels of probability distributions which allow to bound the DBN model approximation errors.

4 Restricted Boltzmann Machines

We denote by $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$ the restricted Boltzmann machine probability model with hidden units $Y_1, \ldots, Y_m$ taking states in $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_m$ and visible units $X_1, \ldots, X_n$ taking states in $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$. Recall the definitions made in pg. 6. In the literature RBMs are defined by default with binary units; however, RBMs with discrete units have appeared in (Welling et al., 2005), and their representational power has been studied in (Montúfar and Morton, 2013). The results from this section are closely related to the analysis given in (Montúfar and Morton, 2013).

Theorem 3. *The model* $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$ *can approximate any mixture distribution* $p = \sum_{i=0}^m \lambda_i p_i$ *arbitrarily well, where* p_0 *is any product distribution, and* p_i *is any mixture of* $(|\mathcal{Y}_i| - 1)$ *product distributions for all* $i \in [m]$ *satisfying* $\mathrm{supp}(p_i) \cap \mathrm{supp}(p_j) = \emptyset$ *for all* $1 \leq i < j \leq m$.

Here, a *product distribution* p is a probability distribution on $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ that factorizes as $p(x_1, \ldots, x_n) = \prod_{j \in [n]} p_j(x_j)$ for all $x \in \mathcal{X}$, where p_j is a distribution on $\mathcal{X}_j$ for all $j \in [n]$. A *mixture* is a weighted sum with non-negative weights adding to one. The support of a distribution p is $\mathrm{supp}(p) := \{x \in \mathcal{X} \colon p(x) > 0\}$.

Proof of Theorem 3. Let $\mathcal{E}_{\mathcal{X}}$ denote the set of strictly positive product distributions of $X_1, \ldots, X_n$. Let $\mathcal{M}_{\mathcal{X}}^k$ denote the set of all mixtures of k product distributions from $\mathcal{E}_{\mathcal{X}}$. The closure $\overline{\mathcal{M}_{\mathcal{X}}^k}$ contains all mixtures of k product distributions, including those which are not strictly positive. Let $q \circ q'$ denote the renormalized entry-wise product with $(q \circ q')(x) = q(x)q'(x) / \sum_{x' \in \mathcal{X}} q(x')q'(x')$ for all $x \in \mathcal{X}$. Let $\mathbb{1}$ denote the constant function on $\mathcal{X}$ with value 1. The model $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$ can be written, up to normalization, as the set

$$\mathcal{M}_{\mathcal{X}}^{|\mathcal{Y}_1|} \circ \cdots \circ \mathcal{M}_{\mathcal{X}}^{|\mathcal{Y}_m|} = \mathbb{R}_+ \mathcal{E}_{\mathcal{X}} \circ (\mathbb{1} + \mathbb{R}_+ \mathcal{M}_{\mathcal{X}}^{|\mathcal{Y}_1|-1}) \circ \cdots \circ (\mathbb{1} + \mathbb{R}_+ \mathcal{M}_{\mathcal{X}}^{|\mathcal{Y}_m|-1}). \quad (10)$$

9

Now consider any probability distributions $p_0 \in \mathcal{E}_\mathcal{X}$, $p_1' \in \overline{\mathcal{M}_\mathcal{X}^{|\mathcal{Y}_1|-1}}, \ldots, p_m' \in \overline{\mathcal{M}_\mathcal{X}^{|\mathcal{Y}_m|-1}}$. If $\mathrm{supp}(p_i') \cap \mathrm{supp}(p_j') = \emptyset$ for all $1 \le i < j \le m$, then the product $(\mathbb{1} + \lambda_1' p_1') \circ \cdots \circ (\mathbb{1} + \lambda_m' p_m')$ is equal to $\mathbb{1} + \sum_{i \in [m]} \lambda_i' p_i'$, up to normalization. Let $\lambda_i' = \lambda_i / \lambda_0 \sum_x p_i'(x) p_0(x)$ and $p_i'(x) = p_i(x)/p_0(x)$. Then $\lambda_0 p_0 \circ (\mathbb{1} + \sum_{i \in [m]} \lambda_i p_i') = \sum_{i=0}^m \lambda_i p_i = p$. Hence the mixture distribution p is contained in the closure of the RBM model. $\square$

RBMs can approximate certain partition models arbitrarily well:

Lemma 4. *Let $\mathcal{P}$ be the partition model with partition blocks $\{x_1\} \times \cdots \times \{x_k\} \times \mathcal{X}_{k+1} \times \cdots \times \mathcal{X}_n$ for all $(x_1, \ldots, x_k) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_k$. If $1 + \sum_{j \in [m]}(|\mathcal{Y}_j| - 1) \ge (\prod_{i \in [k]} |\mathcal{X}_i|)/\max_{j \in [k]} |\mathcal{X}_j|$, then each distribution contained in $\mathcal{P}$ can be approximated arbitrarily well by distributions from $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$.*

Proof. Any point in $\mathcal{P}$ is a mixture of the uniform distributions on the partition blocks. These mixture components have disjoint supports, since the partition blocks are disjoint. They are product distributions, since they can be written as $p_{x_1,\ldots,x_k} = \prod_{i \in [k]} \delta_{x_i} \prod_{i \in [n] \setminus [k]} u_i$, where u_i denotes the uniform distribution on $\mathcal{X}_i$. For any $j \in [k]$, any mixture of the form $\sum_{x_j \in \mathcal{X}_j} \lambda_{x_j} p_{x_1,\ldots,x_k}$ is also a product distribution which factorizes as

$$\big(\sum_{x_j \in \mathcal{X}_j} \lambda_{x_j} \delta_{x_j}\big) \prod_{i \in [k] \setminus \{j\}} \delta_{x_i} \prod_{i \in [n] \setminus [k]} u_i. \tag{11}$$

Hence any point in $\mathcal{P}$ is a mixture of $(\prod_{i \in [k]} |\mathcal{X}_i|)/\max_{j \in [k]} |\mathcal{X}_j|$ product distributions of the form given in eq. (11). The claim follows from Theorem 3. $\square$

Lemma 4, together with the divergence formula for partition models given in pg. 6, implies:

Theorem 5. *If $1 + \sum_{j \in [m]}(|\mathcal{Y}_j| - 1) \ge \big(\prod_{i \in \Lambda} |\mathcal{X}_i|\big)/\max_{i' \in \Lambda} |\mathcal{X}_{i'}|$ for some $\Lambda \subseteq [n]$, then*

$$D_{\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}} \le \log\big(\prod_{i \in [n] \setminus \Lambda} |\mathcal{X}_i|\big).$$

In particular, the model $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$ is a universal approximator whenever

$$1 + \sum_{j \in [m]}(|\mathcal{Y}_j| - 1) \ge |\mathcal{X}|/\max_{i \in [n]} |\mathcal{X}_i|.$$

When all units are q-ary, the RBM with $(q^{n-1}-1)/(q-1)$ hidden units is a universal approximator of distributions on $\{0, 1, \ldots, q-1\}^n$. Theorem 5 generalizes previous results on binary RBMs (Montúfar and Ay, 2011, Theorem 1) and (Montúfar et al., 2011, Theorem 5.1), where it is shown that a binary RBM with $2^{n-1}-1$ hidden units is a universal approximator of distributions on $\{0, 1\}^n$ and that the maximal approximation error of binary RBMs decreases at least logarithmically in the number of hidden units. A previous result by Freund and Haussler (1991, Section 2.5) shows that a binary RBM with 2^n hidden units is a universal approximator of distributions on $\{0, 1\}^n$. See also the work by Le Roux and Bengio (2008, Theorem 2).

5 The Internal Node of a Star

Consider an inwards directed star graph with leaf variables taking states in $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_m$ and internal node variable taking states in $\mathcal{V}$. Denote by $\mathcal{S}_{\mathcal{V},\mathcal{Y}}$ the set of conditional distributions on $\mathcal{V}$ given the states $y \in \mathcal{Y}$ of the leaf units, defined by this network. Each of these distributions can be written as

$$p(v|y;\Theta) = \exp(\mathbf{v}^\top \Theta \left[\begin{smallmatrix} 1 \\ \mathbf{y} \end{smallmatrix}\right])/Z(\Theta \left[\begin{smallmatrix} 1 \\ \mathbf{y} \end{smallmatrix}\right]), \quad \text{for all } v \in \mathcal{V} \text{ and } y \in \mathcal{Y}. \tag{12}$$

The distributions from eq. (4) are of this form, with Θ corresponding to $[(\theta_i^l)^\top | \Theta_i^l]$.

A conditional distribution $p(\cdot|\cdot)$ is naturally identified with the stochastic map defined by the matrix $(p(x|y))_{y,x}$. The following lemma describes some stochastic maps that are representable by the model $\mathcal{S}_{\mathcal{V},\mathcal{Y}}$, and which we will use to define a probability sharing scheme in Section 6.

Lemma 6. *Let* $\mathcal{Z} = \{y_1\} \times \cdots \times \{y_{k-1}\} \times \mathcal{Y}_k \times \{y_{k+1}\} \times \cdots \times \{y_m\} \subseteq \mathcal{Y}$, $k \neq m$. *Furthermore, let* $\mathcal{V} = \mathcal{Y}_m$, *and let* $\{q^z : z \in \mathcal{Z}\}$ *be any distributions on* $\mathcal{V}$. *Then there is a choice of the parameters* Θ *of* $\mathcal{S}_{\mathcal{V},\mathcal{Y}}$ *for which*

$$p(\cdot|y;\Theta) = \begin{cases} q^y, & \text{if } y \in \mathcal{Z} \\ \delta_{y_m}, & \text{otherwise} \end{cases}.$$

Proof. Let $\mathcal{Y}_j = \{0, 1, \ldots, r_j - 1\}$ for all $j \in [m]$, and $r = |\mathcal{V}| = r_m$. The set of strictly positive probability distributions on $\mathcal{V}$ is an exponential family $\mathcal{E}_\mathcal{V} = \{p(v;\theta) = \exp(\mathbf{v}^\top \theta)/Z(\theta)$ for all $v \in \mathcal{V} : \theta \in \mathbb{R}^d\}$ with $d = r - 1$. For some $v \in \mathcal{V}$ let $\vartheta_v \in \mathbb{R}^d$ be the parameter vector of a distribution which attains a unique maximum at v. Then for any fixed $\eta \in \mathbb{R}^d$ we have

$$\lim_{K\to\infty} p(x;\eta + K\vartheta_v) = \delta_v(x) \quad \text{for all } x \in \mathcal{V}. \tag{13}$$

To see this, note that $p(x;K\vartheta_v) \propto p(x;\vartheta_v)^K$ and hence $\lim_{K\to\infty} p(x;K\vartheta_v) = \delta_v$. Furthermore, $p(x;\eta + K\vartheta_v) \propto p(x;\eta)p(x;K\vartheta_v)$.

Without loss of generality let $\mathcal{Z} = \mathcal{Y}_1 \times \{0\} \times \cdots \times \{0\}$. For each $z = (z_1, \ldots, z_m) \in \mathcal{Z}$ let $\theta^{z_1} \in \mathbb{R}^d$ be such that $p(v;\theta^{z_1}) = q^z(v)$ for all $v \in \mathcal{V}$. The matrix Θ can be set as follows:

$$\Theta = \left[\begin{array}{c|c|c|c|c} \theta^0 & \Theta_1 & \Theta_2 & \cdots & \Theta_m \end{array} \right]; \tag{14}$$

where Θ_j contains the columns corresponding to $\mathbf{y}_j$ in eq. (12) and

$$\begin{aligned} \Theta_1 &= [\theta^1 - \theta^0 | \cdots | \theta^{r_1-1} - \theta^0] && \in \mathbb{R}^{d \times r_1}; \\ \Theta_j &= [K_0\vartheta_0 | K_0\vartheta_0 | \cdots | K_0\vartheta_0] && \in \mathbb{R}^{d \times r_j}, \text{ for } j = 2, \ldots, m-1; \tag{15} \\ \Theta_m &= [K_1\vartheta_1 | K_2\vartheta_2 | \cdots | K_{r-1}\vartheta_{r-1}] \in \mathbb{R}^{d \times r}. \end{aligned}$$

The matrix Θ maps $\{\left[\begin{smallmatrix} 1 \\ \mathbf{y} \end{smallmatrix}\right] : y \in \mathcal{Z}\}$ to the parameter vectors $\{\theta^{z_1} : z_1 \in \mathcal{Y}_1\}$ with corresponding distributions $\{q^z : z \in \mathcal{Z}\}$. When $K_0, \ldots, K_{r-1} \in \mathbb{R}$ are chosen such that $\|\theta^0\|, \ldots, \|\theta^{r_1-1}\| \ll K_0 \ll K_1, \ldots, K_{r-1}$, then for each $y \in \mathcal{Y} \setminus \mathcal{Z}$ the vector $\left[\begin{smallmatrix} 1 \\ \mathbf{y} \end{smallmatrix}\right]$ is mapped to a parameter vector $\Theta \left[\begin{smallmatrix} 1 \\ \mathbf{y} \end{smallmatrix}\right]$ with $p(\cdot|y;\Theta \left[\begin{smallmatrix} 1 \\ \mathbf{y} \end{smallmatrix}\right])$ arbitrarily close to δ_{y_m}. □
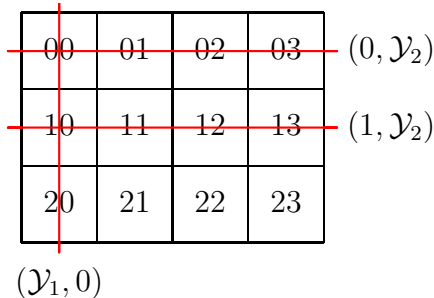
Figure 3: Three multi-sharing steps on $\{0, 1, 2\} \times \{0, 1, 2, 3\}$.

Remark 7. In order to prove Lemma 6 for any subset $\mathcal{Z} \subseteq \mathcal{Y}$ it is sufficient to show that (i) the vectors $\{\mathbf{y} \colon y \in \mathcal{Z}\}$ are affinely independent, and (ii) there is a linear map Θ mapping $\{\begin{bmatrix} 1 \\ \mathbf{y} \end{bmatrix} : y \in \mathcal{Z}\}$ into the zero vector and $\begin{bmatrix} 1 \\ \mathbf{y} \end{bmatrix}$ into the relative interior of the normal cone of $Q_{\mathcal{V}} := \mathrm{conv}\{\mathbf{v} \colon v \in \mathcal{V}\}$ at the vertex $v = y_m$ for all $y \in \mathcal{Y} \setminus \mathcal{Z}$.

6 Probability Sharing

A single directed layer

Consider an input layer of units $Y_1, \ldots, Y_m$ with bipartite connections directed towards an output layer of units $X_1, \ldots, X_n$. Denote by $\mathcal{L}_{\mathcal{X}, \mathcal{Y}}$ the model of conditional distributions defined by this network. Recall the definition from eq. (3). Each conditional distribution $p(\cdot | \cdot; \Theta) \in \mathcal{L}_{\mathcal{X}, \mathcal{Y}}$ defines a linear stochastic map $F_\Theta \colon q \mapsto \sum_{y \in \mathcal{Y}} p(x|y; \Theta) q(y)$ from the simplex $\Delta(\mathcal{Y})$ of distributions on $\mathcal{Y}$ to the simplex $\Delta(\mathcal{X})$ of distributions on $\mathcal{X}$. Here the parameter Θ corresponds to the parameters Θ^l, θ^l of the conditional distributions from eq. (3) for a given l.

For any $y \in \mathcal{Y}$ and $j \in [m]$, we denote by $y[j]$ the one-dimensional cylinder set $\{y_1\} \times \cdots \times \{y_{j-1}\} \times \mathcal{Y}_j \times \{y_{j+1}\} \times \cdots \times \{y_m\}$. Similarly, for any $\Lambda \subseteq [m]$, we denote by $y[\Lambda]$ the cylinder set consisting of all arrays in $\mathcal{Y}$ with fixed values $\{y_i\}_{i \in [m] \setminus \Lambda}$ in the entries $[m] \setminus \Lambda$.

Applying Lemma 6 to each output unit of $\mathcal{L}_{\mathcal{X}, \mathcal{Y}}$ shows:

Theorem 8. *Consider some $\{y^{(s)}\}_{s \in [k]} \subseteq \mathcal{Y}$. Let $\{j_s\}_{s \in [k]}$ be a multiset and $\{i_s\}_{s \in [k]}$ a set of indices from $[m]$. If the cylinder sets $y^{(s)}[j_s]$ are disjoint and $\mathcal{Z}$ is a subset of $\mathcal{Y}$ containing them, then the image of $\Delta(\mathcal{Z})$ by the family of stochastic maps $\mathcal{L}_{\mathcal{Y}, \mathcal{Y}}$ contains $\Delta(\mathcal{Z} \cup_{s \in [k]} y^{(s)}[\{j_s, i_s\}])$.*

This result describes the image of a set of probability distributions by the collection of stochastic maps defined by a DBN layer for all choices of its parameters. In turn, it describes part of the DBN representational power contributed by a layer of units.

A stack of directed layers

In the case of binary units, sequences of probability sharing steps can be defined conveniently using Gray codes, as done in (Le Roux and Bengio, 2010). A Gray code is

an ordered list of vectors, where each two subsequent vectors differ in only one entry. A binary Gray code can be viewed as a sequence of one-dimensional cylinder sets. In the non-binary case, this correspondence is no longer given. Instead, motivated by Theorem 8, we will use one-dimensional cylinder sets in order to define sequences of multi-sharing steps, as shown in Figure 3.

Let $q_i = |\mathcal{Y}_i|$ be the cardinality of $\mathcal{Y}_i$ for $i \in [n]$, and let $m \leq n$. The set $\mathcal{Z} = \{0\} \times \cdots \times \{0\} \times \mathcal{Y}_{m+1} \times \cdots \times \mathcal{Y}_n \subseteq \mathcal{Y}$ can be written as the disjoint union of $k = \prod_{i=m+2}^{n} |\mathcal{Y}_i|$ one-dimensional cylinder sets, as $\mathcal{Z} = \cup_{s=1}^{k} y^{(s)}[m+1]$, where $y^{(s)} = (0, \ldots, 0 \mid \underset{m+1}{0}, y_{m+2}^{(s)}, \ldots, y_n^{(s)})$ and $\{(y_{m+2}^{(s)}, \ldots, y_n^{(s)})\}_{s=1}^{k} = \mathcal{Y}_{m+2} \times \cdots \times \mathcal{Y}_n$.

In the following, each set $y^{(s)}[m+1]$ will be the starting point of a sequence of sharing steps. By Theorem 8, a directed DBN layer maps the simplex of distributions $\Delta(y^{(1)}[m+1] \cup \cdots \cup y^{(k)}[m+1])$ surjectively to the simplex of distributions $\Delta(y^{(1)}[m+1, 1] \cup \cdots \cup y^{(k)}[m+1, k])$. The latter can be mapped by a further DBN layer onto a larger simplex and so forth. Starting with $y^{(1)}[m+1]$, consider the sequence

$$
\begin{aligned}
&(0, 0, \ldots, 0 \mid 0, y_{m+2}^{(1)}, \ldots, y_n^{(1)})[m+1, 1] \\
&(0, 0, \ldots, 0 \mid 0, y_{m+2}^{(1)}, \ldots, y_n^{(1)})[m+1, 2] \quad\quad\quad (16) \\
&(1, 0, \ldots, 0 \mid 0, y_{m+2}^{(1)}, \ldots, y_n^{(1)})[m+1, 2]
\end{aligned}
$$

continued as shown in Table 1. We denote this sequence of cylinder sets by G^1, and its l-th row (a cylinder set) by $G^1(l)$. The union $\cup_{l \in [K]} G^1(l)$ of the first K rows, with $K = 1 + q_1 + q_1 q_2 + \cdots + \prod_{j=1}^{m-1} q_j$, is equal to $\mathcal{Y}_1 \times \cdots \times \mathcal{Y}_m \times \mathcal{Y}_{m+1} \times \{y_{m+2}^{(1)}\} \times \cdots \times \{y_n^{(1)}\}$.

We define k sequences $G^1, \ldots, G^k$ as follows: The first m coordinates of G^s are equal to a permutation of the first m coordinates of G^1, defined by shifting each of these m columns cyclically s positions to the right. The last $n - m$ coordinates of G^s are equal to $(\mathcal{Y}_{m+1}, y_{m+2}^{(s)}, \ldots, y_n^{(s)})$.

We use the abbreviation $\{s+t\} := (s+t-1) \mod (m) + 1$. Within the first m columns, the free coordinate of the l-th row of G^s is $s + \kappa$, where κ is the least integer with $l \leq \sum_{i=0}^{\kappa} \prod_{j=s}^{\{s+i-1\}} q_j$. Here the empty product is defined as 1. Let $q = \max_{j \in \{1, \ldots, m\}} q_j$. We can modify each sequence G^s, by repeating rows if necessary, such that the free coordinate of the l-th row of the resulting sequence $\tilde{G}^s$ is $s + \kappa$, where κ is the least integer with $l \leq \sum_{t=0}^{\kappa} q^t$. This κ does not depend on s.

The sequences $\tilde{G}^s$ for $s \in \{1, \ldots, k\}$ are all different from each other in the last $n - m$ coordinates and have a different 'sharing' free-coordinate in each row. The union of cylinder sets in all rows of these sequences is equal to $\mathcal{Y}_1 \times \cdots \times \mathcal{Y}_n$.

7 Deep Belief Networks

Proposition 9. *Consider a DBN with $L \geq 2$ layers of width n, each layer containing units with state spaces of cardinalities $q_1, \ldots, q_n$. Let m be any integer with $n \geq m \geq \prod_{j=m+2}^{n} q_j =: k$. The corresponding probability model can approximate a distribution p on $\{0, 1, \ldots, q_1 - 1\} \times \cdots \times \{0, 1, \ldots, q_n - 1\}$ arbitrarily well whenever the support of p is contained in $\cup_{s \in [k]} \cup_{l \in [L-2]} \tilde{G}^s(l)$.*

13

$\mathcal{Y}_1$	0	0	0	0
0	$\mathcal{Y}_2$	0	0	0
1	$\mathcal{Y}_2$	0	0	0
⋮	⋮			
q_1-1	$\mathcal{Y}_2$	0	0	0
0	0	$\mathcal{Y}_3$	0	0
0	1	$\mathcal{Y}_3$	0	0
⋮	⋮	⋮		
0	q_2-1	$\mathcal{Y}_3$	0	0
1	0	$\mathcal{Y}_3$	0	0
1	1	$\mathcal{Y}_3$	0	0
⋮	⋮	⋮		
1	q_2-1	$\mathcal{Y}_3$	0	0
⋮	⋮			
q_1-1	0	$\mathcal{Y}_3$	0	0
q_1-1	1	$\mathcal{Y}_3$	0	0
⋮	⋮	⋮		
q_1-1	q_2-1	$\mathcal{Y}_3$	0	0
0	0	0	$\mathcal{Y}_4$	0
0	0	1	$\mathcal{Y}_4$	0
⋮	⋮	⋮	⋮	
0	0	q_3-1	$\mathcal{Y}_4$	0

⋮		⋮			
q_1-1	q_2-1	$\cdots$	$q_{m-2}-1$	0	$\mathcal{Y}_m$
q_1-1	q_2-1	$\cdots$	$q_{m-2}-1$	1	$\mathcal{Y}_m$
⋮	⋮		⋮	⋮	⋮
q_1-1	q_2-1	$\cdots$	$q_{m-2}-1$	$q_{m-1}-1$	$\mathcal{Y}_m$

Table 1: Sequence of one-dimensional cylinder sets.

Proof. Note that $\prod_{j=m+2}^{n} q_j \leq n \leq 1 + \sum_{j \in [n]} (q_j - 1)$. By Theorem 3 the top RBM can approximate each distribution in the probability simplex on $\{0\} \times \cdots \times \{0\} \times \mathcal{X}_{m+1} \times \cdots \times \mathcal{X}_n$ arbitrarily well. By Theorem 8, this simplex can be mapped iteratively into larger simplices, according to the sequences $\tilde{G}^s$ from Section 6. $\square$

Theorem 10. *Consider a DBN with L layers of width n, each layer containing units with state spaces of cardinalities $q_1, \ldots, q_n$. Let m be any integer with $n \geq m \geq \prod_{j=m+2}^{n} q_j$ and $q = q_1 \geq \cdots \geq q_m$. If $L \geq 2 + 1 + q + \cdots + q^{S-1} = 2 + \frac{q^S - 1}{q - 1}$, then the DBN model can approximate each distribution in a partition model $\mathcal{P}$ of coarseness $\prod_{j=1}^{m-S} q_j$ arbitrarily well.*

Proof. When $L \leq 2$ the result follows from Lemma 4. Assume therefore that $L \geq 2 + 1 + q + q^2 + \cdots + q^r$, $r \geq 0$. We use the abbreviation $\{s + t\} := (s + t - 1)_{\mod (m)} + 1$. Let $k = \prod_{j=m+2}^{n} q_j$ and $\{(y_{m+2}^{(s)}, \ldots, y_n^{(s)}) : s \in [k]\} = \mathcal{Y}_{m+2} \times \cdots \times \mathcal{Y}_n$. The top RBM can approximate each distribution from a partition model $\mathcal{P}$ (on a subset of $\mathcal{Y}$) arbitrarily well, whose partition blocks are the cylinder sets with fixed coordinate values

$$y_s = 0, y_{\{s+1\}} = 0, \ldots, y_{\{s+r\}} = 0, y_{m+1}, y_{m+2}^{(s)}, \ldots, y_n^{(s)};$$

for all $y_{m+1} \in \mathcal{Y}_{m+1}$, for all $s \in [k]$. After $L - 2$ probability sharing steps starting from $\mathcal{P}$, the DBN can approximate the distributions from the partition model arbitrarily well, whose partition blocks are the cylinder sets with fixed coordinate values

$$y_s, y_{\{s+1\}}, \ldots, y_{\{s+r\}}, y_{m+1}, y_{m+2}^{(s)}, \ldots, y_n^{(s)};$$

for all possible choices of $y_s, y_{\{s+1\}}, \ldots, y_{\{s+r\}}, y_{m+1}$, for all $s \in [k]$. The maximal cardinality of such a block is $q_1 \cdots q_{m-r-1}$, and the union of all blocks equals $\mathcal{Y}$. $\square$

Proof of Theorem 2. The claim follows bounding the divergence of the partition models described in Theorem 10. $\square$

As a corollary we obtain the following bound for the expectation value of the divergence from distributions drawn from a Dirichlet prior, to the DBN model.

Corollary 11. *The expectation value of the divergence from a probability distribution p drawn from the symmetric Dirichlet distribution $\mathrm{Dir}_{(a,\ldots,a)}$ to the model DBN from Theorem 2 is bounded by*

$$\int_{\Delta} D(p \| \mathrm{DBN}) \, \mathrm{Dir}_{(a,\ldots,a)}(p) \, \mathrm{d}p \leq (\psi(a + 1) - \psi(ca + 1) + \ln(c)) \log(e),$$

where $c = \prod_{j \in [m-S]} q_j$, ψ is the digamma function, and e is Euler's constant.

Proof. This is a consequence of analytical work (Montúfar and Rauh, 2012) on the expectation value of Kullback-Leibler divergences of standard probability models, applied to the partition models described in Theorem 2. $\square$

A Small Experiments

We run some computer experiments, not with the purpose of validating the quality of our bounds in general, but with the purpose of giving a first empirical insight. It is important to emphasize that numerical experiments evaluating the divergence from probability models defined by neural networks are only feasible for small networks, since otherwise the model distributions are too hard to compute (see, e.g., Long and Servedio, 2010). For large models one still could try to sample the distributions and replace the divergence by a proxy, like the discrepancy of low-level statistics, but here we will focus on small networks.

We generate artificial data in the following natural way: For a given visible state space $\mathcal{X}$ and the corresponding probability simplex $\Delta(\mathcal{X})$, we draw a set of distributions $\{p^i \in \Delta(\mathcal{X}) : i = 1, \ldots, T\}$ from the Dirichlet distribution $\mathrm{Dir}_{(a,\ldots,a)}$ on $\Delta(\mathcal{X})$. For the purpose of our experiments, we choose the concentration parameter a in such a way that the Dirichlet density is higher for low-entropy distributions (most distributions in practice have relatively few preferred states and hence a small entropy). Next, for each $i = 1, \ldots, T$, we generate N i.i.d. samples from p^i, which results in a data vector $X^i = (X^i_1, \ldots, X^i_N) \in \mathcal{X}^N$ with empirical distribution $P^i = \frac{1}{N} \sum_{j=1}^N \delta_{X^i_j}$.

A network $\mathcal{N}$ (with visible states $\mathcal{X}$) is then tested on all data sets X^i, $i = 1, \ldots, T$. For each data set we train $\mathcal{N}$ using contrastive divergence (CD) (Hinton, 2002; Hinton et al., 2006) and maximum likelihood (ML) gradient. This gives us a maximum likelihood estimate p_{θ_i} of P^i within $\mathcal{N}$. Finally, we compute the Kullback-Leibler divergence $D(P^i \| p_{\theta_i})$, the maximum value over all data sets $\mathsf{max}_{\mathsf{CD+ML}} = \max_{i=1,\ldots,T} D(P^i \| p_{\theta_i})$, and the mean value over all data sets $\mathsf{mean}_{\mathsf{CD+ML}} = \frac{1}{T} \sum_{i=1}^T D(P^i \| p_{\theta_i})$. We do not need cross validation, or $D(p^i \| p_{\theta_i})$, because we are interested in the representational power of $\mathcal{N}$, rather than on its generalization properties.

We note that the number of distributions which have the largest divergence from $\mathcal{N}$ is relatively small, and hence the random variable $\mathsf{max}_{\mathsf{CD+ML}}$ has a large variance (unless the number of data sets tends to infinity, $T \to \infty$). Moreover, we note that it is hard to find the best approximations of a given target P^i. Since the likelihood function $L_{X^i}(\theta) = \prod_{j=1}^N p_\theta(X^i_j)$ has many local maxima, the distribution p_{θ_i} is often not a global maximizer of L_{X^i}, even if training is arranged with many parameter initializations. Many times the estimated value p_{θ_i} is a good local minimizer of the divergence, but sometimes it is relatively poor (especially for the larger networks). This contributes again to the variance of $\mathsf{max}_{\mathsf{CD+ML}}$. The mean values $\mathsf{mean}_{\mathsf{CD+ML}}$, on the other hand, are more stable.

Figure 4 shows the results for small binary RBMs with 3 and 4 visible units, and Figure 5 shows the results for small constant-width binary DBNs with 4 visible units. In both figures the maximum and mean divergence is captured relatively well by our theoretical bounds. The empirical maximum values have a well recognizable discrepancy from the theoretical bound. This is explained by the large variance of $\mathsf{max}_{\mathsf{CD+ML}}$, given the limited number of target distributions used in these experiments. Finding a maximizer of the divergence (a data vector $X \in \mathcal{X}^N$ that is hardest to represent) is hard. Most target distributions can be approximated much better than the hardest distributions. A second observation is that with increasing network complexity (more hidden
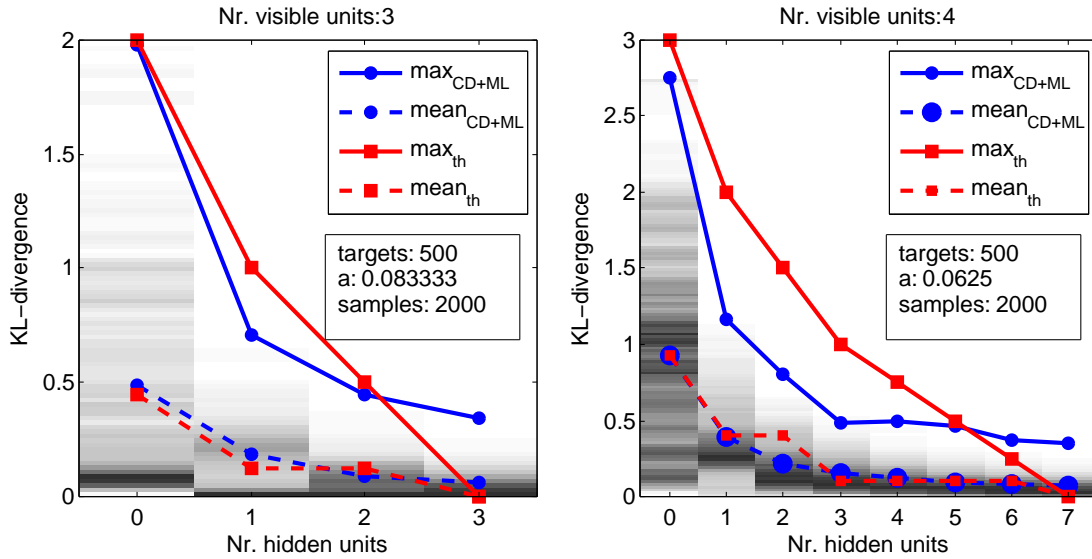
16

Figure 4: Empirical evaluation of the representational power of small binary RBMs. The gray shading indicates the frequency at which the target distribution P^i had a given divergence from the trained RBM distribution p_{θ_i} (the darker a value, the more frequent). The lines with round markers show the mean divergence (dashed) and maximal divergence (solid) over all target distributions. The lines with square markers show the theoretical upper bounds of the mean divergence (dashed) and maximal divergence (solid) over the continuum of all possible target distributions drawn from the symmetric Dirichlet distribution $\mathrm{Dir}_{(a,\dots,a)}$.
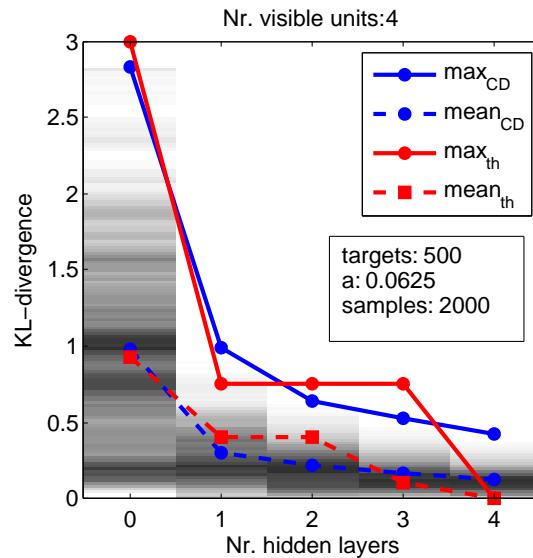

Figure 5: Empirical evaluation of the representational power of small binary DBNs. The details are as in Figure 4, whereby the theoretical upper bounds shown here for the maximal and mean divergence, $\max_{\mathrm{th}}$ and $\mathrm{mean}_{\mathrm{th}}$, are a combination of our results for RBMs and DBNs.

units), finding the best approximations of the target distributions becomes harder (even increasing the training efforts). This causes the empirical maximum divergence to actually surpass the theoretical bounds. In other words, although the models are in principle able to approximate the targets accurately, according to our theoretical bounds, in practice they may not, because of the difficult training, and their capacity remains wasted. The empirical mean values, on the other hand, have a much lower variance and are captured quite accurately by our theoretical bounds.

Acknowledgment

References

Ay, N. and Knauf, A. (2006). Maximizing multi-information. *Kybernetika*, 42:517–538.

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127. Also published as a book. Now Publishers, 2009.

Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. (2007). Greedy layer-wise training of deep networks. In *Advances in Neural Information Processing Systems 19*, pages 153–160.

Freund, Y. and Haussler, D. (1991). Unsupervised learning of distributions of binary vectors using 2-layer networks. In *Advances in Neural Information Processing Systems 4*, pages 912–919.

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

Hinton, G. E., Osindero, S., and Teh, Y. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554.

Hornik, K., Stinchcombe, M. B., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366.

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

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

Long, P. M. and Servedio, R. A. (2010). Restricted Boltzmann machines are hard to approximately evaluate or simulate. In *Proceedings of the 27-th ICML*, pages 703–710.

Montúfar, G. (2013). Mixture decompositions of exponential families using a decomposition of their sample spaces. *Kybernetika*, 49(1):23–39.

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

Montúfar, G. and Morton, J. (2013). Discrete restricted Boltzmann machines. In *Online Proceedings of the 1-st International Conference on Learning Representations (ICLR2013)*.

Montúfar, G. and Rauh, J. (2012). Scaling of model approximation errors and expected entropy distances. In *Proceedings of the WUPES'12*, pages 137–148.

Montúfar, G., Rauh, J., and Ay, N. (2011). Expressive power and approximation errors of restricted Boltzmann machines. In *Advances in Neural Information Processing Systems 24*, pages 415–423.

Rauh, J. (2013). Optimally approximating exponential families. *Kybernetika*, 49(2):199–215.

Smolensky, P. (1986). Information processing in dynamical systems: foundations of harmony theory. In *Parallel distributed processing: explorations in the microstructure of cognition, vol. 1*, pages 194–281. MIT Press.

Sutskever, I. and Hinton, G. E. (2008). Deep narrow sigmoid belief networks are universal approximators. *Neural Computation*, 20(11):2629–2636.

Welling, M., Rosen-Zvi, M., and Hinton, G. E. (2005). Exponential family harmoniums with an application to information retrieval. In *Advances in Neural Information Processing Systems 17*, pages 1481–1488.