Expressive Power and Approximation Errors of Restricted Boltzmann Machines

Guido F. Montúfar[1], **Johannes Rauh**[1], **and Nihat Ay**[1,2]

[1]Max Planck Institute for Mathematics in the Sciences, Inselstraße 22 04103 Leipzig, Germany
[2]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA
{montufar,jrauh,nay}@mis.mpg.de

Abstract

We present explicit classes of probability distributions that can be learned by Restricted Boltzmann Machines (RBMs) depending on the number of units that they contain, and which are representative for the expressive power of the model. We use this to show that the maximal Kullback-Leibler divergence to the RBM model with n visible and m hidden units is bounded from above by $(n-1) - \log(m+1)$. In this way we can specify the number of hidden units that guarantees a sufficiently rich model containing different classes of distributions and respecting a given error tolerance.

1 Introduction

A Restricted Boltzmann Machine (RBM) [24, 10] is a learning system consisting of two layers of binary stochastic units, a hidden layer and a visible layer, with a complete bipartite interaction graph. RBMs are used as generative models to simulate input distributions of binary data. They can be trained in an unsupervised way and more efficiently than general Boltzmann Machines, which are not restricted to have a bipartite interaction graph [11, 6]. Furthermore, they can be used as building blocks to progressively train and study deep learning systems [13, 4, 16, 21]. Hence, RBMs have received increasing attention in the past years.

An RBM with n visible and m hidden units generates a stationary distribution on the states of the visible units which has the following form:

$$p_{W,C,B}(v) = \frac{1}{Z_{W,C,B}} \sum_{h \in \{0,1\}^m} \exp\left(h^\top W v + C^\top h + B^\top v\right) \quad \forall v \in \{0,1\}^n \,,$$

where $h \in \{0,1\}^m$ denotes a state vector of the hidden units, $W \in \mathbf{R}^{m \times n}, C \in \mathbf{R}^m$ and $B \in \mathbf{R}^n$ constitute the model parameters, and $Z_{W,C,B}$ is a corresponding normalization constant. In the sequel we denote by $\mathrm{RBM}_{n,m}$ the set of all probability distributions on $\{0,1\}^n$ which can be approximated arbitrarily well by a visible distribution generated by the RBM with m hidden and n visible units for an appropriate choice of the parameter values.

As shown in [21] (generalizing results from [15]) $\mathrm{RBM}_{n,m}$ contains any probability distribution if $m \geq 2^{n-1} - 1$. On the other hand, if $\mathrm{RBM}_{n,m}$ equals the set $\mathcal{P}$ of all probability distributions on $\{0,1\}^n$, then it must have at least $\dim(\mathcal{P}) = 2^n - 1$ parameters, and thus at least $\lceil 2^n/(n+1) \rceil - 1$ hidden units [21]. In fact, in [8] it was shown that for most combinations of m and n the dimension of $\mathrm{RBM}_{n,m}$ (as a manifold, possibly with singularities) equals either the number of parameters or $2^n - 1$, whatever is smaller. However, the geometry of $\mathrm{RBM}_{n,m}$ is intricate, and even an RBM of dimension $2^n - 1$ is not guaranteed to contain all visible distributions, see [20] for counterexamples.

In summary, an RBM that can approximate any distribution arbitrarily well must have a very large number of parameters and hidden units. In practice, training such a large system is not desirable or even possible. However, there are at least two reasons why in many cases this is not necessary:

1

- An appropriate approximation of distributions is sufficient for most purposes.
- The interesting distributions the system shall simulate belong to a small class of distributions. Therefore, the model does not need to approximate all distributions.

For example, the set of optimal policies in reinforcement learning [25], the set of dynamics kernels that maximize predictive information in robotics [26] or the information flow in neural networks [3] are contained in very low dimensional manifolds; see [2]. On the other hand, usually it is very hard to mathematically describe a set containing the optimal solutions to general problems, or a set of interesting probability distributions (for example the class of distributions generating natural images). Furthermore, although RBMs are parametric models and for any choice of the parameters we have a resulting probability distribution, in general it is difficult to explicitly specify this resulting probability distribution (or even to estimate it [18]). Due to these difficulties the number of hidden units m is often chosen on the basis of experience [12], or m is considered as a hyperparameter which is optimized by extensive search, depending on the distributions to be simulated by the RBM.

In this paper we give an explicit description of classes of distributions that are contained in $\mathrm{RBM}_{n,m}$, and which are representative for the expressive power of this model. Using this description, we estimate the maximal Kullback-Leibler divergence between an arbitrary probability distribution and the best approximation within $\mathrm{RBM}_{n,m}$.

This paper is organized as follows: Section 2 discusses the different kinds of errors that appear when an RBM learns. Section 3 introduces the statistical models studied in this paper. Section 4 studies submodels of $\mathrm{RBM}_{n,m}$. An upper bound of the approximation error for RBMs is found in Section 5.

2   Approximation Error

When training an RBM to represent a distribution p, there are mainly three contributions to the discrepancy between p and the state of the RBM after training:

1. Usually the underlying distribution p is unknown and only a set of samples generated by p is observed. These samples can be represented as an empirical distribution p^{Data}, which usually is not identical with p.

2. The set $\mathrm{RBM}_{n,m}$ does not contain every probability distribution, unless the number of hidden units is very large, as we outlined in the introduction. Therefore, we have an approximation error given by the distance of p^{Data} to the best approximation $p^{\mathrm{Data}}_{\mathrm{RBM}}$ contained in the RBM model.

3. The learning process may yield a solution $\tilde{p}^{\mathrm{Data}}_{\mathrm{RBM}}$ in RBM which is not the optimum $p^{\mathrm{Data}}_{\mathrm{RBM}}$. This occurs, for example, if the learning algorithm gets trapped in a local optimum, or if it optimizes an objective different from Maximum Likelihood, e.g. contrastive divergence (CD), see [6].

In this paper we study the expressive power of the RBM model and the Kullback-Leibler divergence from an arbitrary distribution to its best representation within the RBM model. Estimating the approximation error is difficult, because the geometry of the RBM model is not sufficiently understood. Our strategy is to find subsets $\mathcal{M} \subseteq \mathrm{RBM}_{n,m}$ that are easy to describe. Then the maximal error when approximating probability distributions with an RBM is upper bounded by the maximal error when approximating with $\mathcal{M}$.

Consider a finite set $\mathcal{X}$. A real valued function on $\mathcal{X}$ can be seen as a real vector with $|\mathcal{X}|$ entries. The set $\mathcal{P} = \mathcal{P}(\mathcal{X})$ of all probability distributions on $\mathcal{X}$ is a $(|\mathcal{X}| - 1)$-dimensional simplex in $\mathbf{R}^{|\mathcal{X}|}$. There are several notions of distance between probability distributions, and in turn for the error in the representation (approximation) of a probability distribution. One possibility is to use the induced distance of the Euclidian space $\mathbf{R}^{|\mathcal{X}|}$. From the point of view of information theory, a more meaningful distance notion for probability distributions is the Kullback-Leibler divergence:

$$D(p\|q) := \sum_x p(x) \log \frac{p(x)}{q(x)} \ .$$

In this paper we use the basis 2 logarithm. The Kullback-Leibler (KL) divergence is non-negative and vanishes if and only if $p = q$. If the support of q does not contain the support of p it is defined

2

$q = p$ $q = \frac{1}{|\mathcal{X}|}$

relative error $\quad 0$ $\qquad\qquad\qquad\qquad\qquad \frac{128}{255}$ $\qquad\qquad\qquad 1$
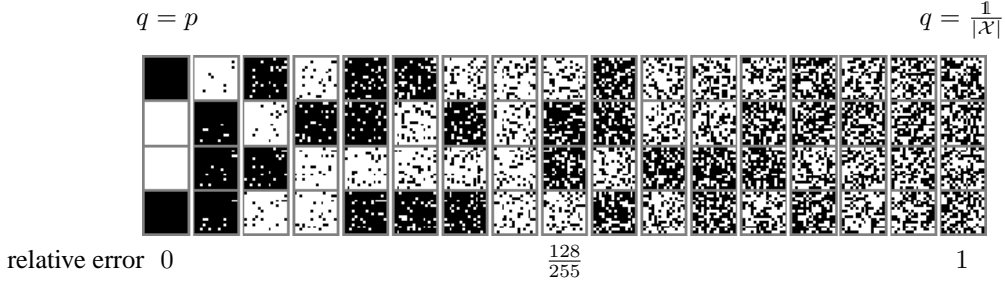
Figure 1: This figure gives an intuition on what the size of an error means for probability distributions on images with 16×16 pixels. Every column shows four samples drawn from the best approximation q of the distribution $p = \frac{1}{2}(\delta_{(1\ldots1)} + \delta_{(0\ldots0)})$ within a partition model with 2 randomly chosen cubical blocks, containing $(0\ldots0)$ and $(1\ldots1)$, of cardinality from 1 (first column) to $\frac{|\mathcal{X}|}{2}$ (last column). As a measure of error ranging from 0 to 1 we take $D(p\|q)/D\big(p\|\frac{1}{|\mathcal{X}|}\big)$. The last column shows samples from the uniform distribution, which is, in particular, the best approximation of p within $\mathrm{RBM}_{n,0}$. Note that an RBM with 1 hidden unit can approximate p with arbitrary accuracy, see Theorem 4.1.

as ∞. The summands with $p(x) = 0$ are set to 0. The KL-divergence is not symmetric, but it has nice information theoretic properties [14, 7].

If $\mathcal{E} \subseteq \mathcal{P}$ is a statistical model and if $p \in \mathcal{P}$, then any probability distribution $p_{\mathcal{E}} \in \overline{\mathcal{E}}$ satisfying

$$D(p\|p_{\mathcal{E}}) = D(p\|\mathcal{E}) := \min\{D(p\|q) : q \in \overline{\mathcal{E}}\}$$

is called a *(generalized) reversed information projection*, or rI-projection. Here, $\overline{\mathcal{E}}$ denotes the closure of $\mathcal{E}$. If p is an empirical distribution, then one can show that any rI-projection is a maximum likelihood estimate.

In order to assess an RBM or some other model $\mathcal{M}$ we use the maximal approximation error with respect to the KL-divergence when approximating arbitrary probability distributions using $\mathcal{M}$:

$$D_{\mathcal{M}} := \max\left\{D(p\|\mathcal{M}) : p \in \mathcal{P}\right\} .$$

For example, the maximal KL-divergence to the uniform distribution $\frac{1}{|\mathcal{X}|}$ is attained by any Dirac delta distributions δ_x, $x \in \mathcal{X}$, and amounts to:

$$D_{\left\{\frac{1}{|\mathcal{X}|}\right\}} = D(\delta_x\|\frac{1}{|\mathcal{X}|}) = \log|\mathcal{X}| . \tag{1}$$

3 Model Classes

3.1 Exponential families and product measures

In this work we only need a restricted class of exponential families, namely exponential families on a finite set with uniform reference measure. See [5] for more on exponential families. The boundary of discrete exponential families is discussed in [23], which uses a similar notation.

Let $A \in \mathbf{R}^{d \times |\mathcal{X}|}$ be a matrix. The columns A_x of A will be indexed by $x \in \mathcal{X}$. The rows of A can be interpreted as functions on $\mathbf{R}$. The *exponential family* $\mathcal{E}_A$ with *sufficient statistics* A consists of all probability distributions of the form p_λ, $\lambda \in \mathbf{R}^d$, where

$$p_\lambda(x) = \frac{\exp(\lambda^\top A_x)}{\sum_x \exp(\lambda^\top A_x)}, \qquad \text{for all } x \in \mathcal{X}.$$

Note that any probability distribution in $\mathcal{E}_A$ has full support. Furthermore, $\mathcal{E}_A$ is in general not a closed set. The closure $\overline{\mathcal{E}_A}$ (with respect to the usual topology on $\mathbf{R}^{\mathcal{X}}$) will be important in the following. Exponential families behave nicely with respect to rI-projection: Any $p \in \mathcal{P}$ has a unique rI-projection $p_{\mathcal{E}}$ to $\mathcal{E}_A$.

3

The most important exponential families in this work are the independence models. The *independence model* of n binary random variables consists of all probability distributions on $\{0,1\}^n$ that factorize:

$$\overline{\mathcal{E}_n} = \left\{ p \in \mathcal{P}(\mathcal{X}) : p(x_1, \dots, x_n) = \prod_{i=1}^{n} p_i(x_i) \text{ for some } p_i \in \mathcal{P}(\{0,1\}) \right\} .$$

It is the closure of an n-dimensional exponential family $\mathcal{E}_n$. This model corresponds to the RBM model with no hidden units. An element of the independence model is called a *product distribution*.

Lemma 3.1 (Corollary 4.1 of [1]) *Let $\overline{\mathcal{E}_n}$ be the independence model on $\{0,1\}^n$. If $n > 0$, then $D_{\mathcal{E}_n} = (n-1)$. The global maximizers are the distributions of the form $\frac{1}{2}(\delta_x + \delta_y)$, where $x, y \in \{0,1\}^n$ satisfy $x_i + y_i = 1$ for all i.*

This result should be compared with (1). Although the independence model is much larger than the set $\{\frac{1}{|\mathcal{X}|}\}$, the maximal divergence decreases only by 1. As shown in [22], if $\mathcal{E}$ is any exponential family of dimension k, then $D_{\mathcal{E}} \geq \log(|\mathcal{X}|/(k+1))$. Thus, this notion of distance is rather strong. The exponential families satisfying $D_{\mathcal{E}} = \log(|\mathcal{X}|/(k+1))$ are partition models; they will be defined in the following section.

3.2 Partition models and mixtures of products with disjoint supports

The *mixture* of m models $\mathcal{M}_1, \dots, \mathcal{M}_m \subseteq \mathcal{P}$ is the set of all convex combinations

$$p = \sum_i \alpha_i p_i \ , \quad \text{where } p_i \in \mathcal{M}_i, \alpha_i \geq 0, \sum_i \alpha_i = 1 \ . \tag{2}$$

In general, mixture models are complicated objects. Even if all models $\mathcal{M}_1 = \dots = \mathcal{M}_m$ are equal, it is difficult to describe the mixture [17, 19]. The situation simplifies considerably if the models have disjoint supports. Note that given any partition $\xi = \{\mathcal{X}_1, \dots, \mathcal{X}_m\}$ of $\mathcal{X}$, any $p \in \mathcal{P}$ can be written as $p(x) = p^{\mathcal{X}_i}(x)p(\mathcal{X}_i)$ for all $x \in \mathcal{X}_i$ and $i \in \{1, \dots, m\}$, where $p^{\mathcal{X}_i}$ is a probability measure in $\mathcal{P}(\mathcal{X}_i)$ for all i.

Lemma 3.2 *Let $\xi = \{\mathcal{X}_1, \dots, \mathcal{X}_m\}$ be a partition of $\mathcal{X}$ and let $\mathcal{M}_1, \dots, \mathcal{M}_m$ be statistical models such that $\mathcal{M}_i \subseteq \mathcal{P}(\mathcal{X}_i)$. Consider any $p \in \mathcal{P}$ and corresponding $p^{\mathcal{X}_i}$ such that $p(x) = p^{\mathcal{X}_i}(x)p(\mathcal{X}_i)$ for $x \in \mathcal{X}_i$. Let p_i be an rI-projection of $p^{\mathcal{X}_i}$ to $\mathcal{M}_i$. Then the rI-projection $p_{\mathcal{M}}$ of P to the mixture $\mathcal{M}$ of $\mathcal{M}_1, \dots, \mathcal{M}_m$ satisfies*

$$p_{\mathcal{M}}(x) = p(\mathcal{X}_i)p_i(x), \qquad \text{whenever } x \in \mathcal{X}_i \ .$$

Therefore, $D(p\|\mathcal{M}) = \sum_i p(\mathcal{X}_i)D(p^{\mathcal{X}_i}\|\mathcal{M}_i)$, and so $D_{\mathcal{M}} = \max_{i=1,\dots,m} D_{\mathcal{M}_i}$.

Proof Let $p \in \mathcal{M}$ be as in (2). Then $D(q\|p) = \sum_{i=1}^{m} q(\mathcal{X}_i)D(q^{\mathcal{X}_i}\|p_i)$ for all $q \in \mathcal{P}$. For fixed q this sum is minimal if and only if each term is minimal. $\square$

If each $\mathcal{M}_i$ is an exponential family, then the mixture is also an exponential family (this is not true if the supports of the models $\mathcal{M}_i$ are not disjoint). In the rest of this section we discuss two examples.

If each $\mathcal{M}_i$ equals the set containing just the uniform distribution on $\mathcal{X}_i$, then $\mathcal{M}$ is called the *partition model* of ξ, denoted with $\mathcal{P}_\xi$. The partition model $\mathcal{P}_\xi$ is given by all distributions with constant value on each block $\mathcal{X}_i$, i.e. those that satisfy $p(x) = p(y)$ for all $x, y \in \mathcal{X}_i$. This is the closure of the exponential family with sufficient statistics

$$A_x = (\chi_1(x), \chi_2(x), \dots, \chi_d(x))^\top \ ,$$

where $\chi_i := \chi_{\mathcal{X}_i}$ is 1 on $x \in \mathcal{X}_i$, and 0 everywhere else. See [22] for interesting properties of partition models.

The partition models include the set of finite exchangeable distributions (see e.g. [9]), where the blocks of the partition are the sets of binary vectors which have the same number of entries equal to one. The probability of a vector v depends only on the number of ones, but not on their position.

Corollary 3.3 *Let $\xi = \{\mathcal{X}_1, \dots, \mathcal{X}_m\}$ be a partition of $\mathcal{X}$. Then $D_{\mathcal{P}_\xi} = \max_{i=1,\dots,m} \log |\mathcal{X}_i|$.*

4
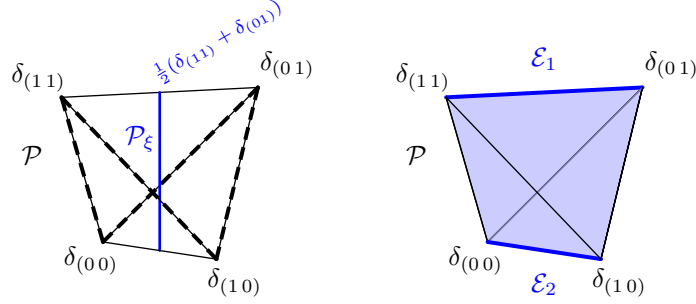
Figure 2: Models in $\mathcal{P}(\{0,1\}^2)$. Left: The blue line represents the partition model $\mathcal{P}_\xi$ with partition $\xi = \{(11),(01)\}\cup\{(00),(10)\}$. The dashed lines represent the set of KL-divergence maximizers for $\mathcal{P}_\xi$. Right: The mixture of the product distributions $\mathcal{E}_1$ and $\mathcal{E}_2$ with disjoint supports on $\{(11),(01)\}$ and $\{(00),(10)\}$ corresponding to the same partition ξ equals the whole simplex $\mathcal{P}$.

Now assume that $\mathcal{X} = \{0,1\}^n$ is the set of binary vectors of length n. As a subset of $\mathbf{R}^n$ it consists of the vertices (extreme points) of the n-dimensional hypercube. The vertices of a k-dimensional face of the n-cube are given by fixing the values of x in $n-k$ positions:

$$\{x \in \{0,1\}^n : x_i = \tilde{x}_i, \forall i \in I, \text{ for some } I \subseteq \{1,\ldots,n\}, |I| = n-k\}$$

We call such a subset $\mathcal{Y} \subseteq \mathcal{X}$ *cubical* or a *face* of the n-cube. A cubical subset of cardinality 2^k can be naturally identified with $\{0,1\}^k$. This identification allows to define independence models and product measures on $\mathcal{P}(\mathcal{Y}) \subseteq \mathcal{P}(\mathcal{X})$. Note that product measures on $\mathcal{Y}$ are also product measures on $\mathcal{X}$, and the independence model on $\mathcal{Y}$ is a subset of the independence model on $\mathcal{X}$.

Corollary 3.4 *Let $\xi = \{\mathcal{X}_1,\ldots,\mathcal{X}_m\}$ be a partition of $\mathcal{X} = \{0,1\}^n$ into cubical sets. For any i let $\mathcal{E}_i$ be the independence model on $\mathcal{X}_i$, and let $\mathcal{M}$ be the mixture of $\mathcal{E}_1,\ldots,\mathcal{E}_m$. Then*

$$D_{\mathcal{M}} = \max_{i=1,\ldots,m} \log(|\mathcal{X}_i|) - 1 .$$

See Figure 1 for an intuition on the approximation error of partition models, and see Figure 2 for small examples of a partition model and of a mixture of products with disjoint support.

4   Classes of distributions that RBMs can learn

Consider a set $\xi = \{\mathcal{X}_i\}_{i=1}^m$ of m disjoint cubical sets $\mathcal{X}_i$ in $\mathcal{X}$. Such a ξ is a partition of some subset $\cup\xi = \cup_i\mathcal{X}_i$ of $\mathcal{X}$ into m disjoint cubical sets. We write G_m for the collection of all such partitions. We have the following result:

Theorem 4.1 $\text{RBM}_{n,m}$ *contains the following distributions:*

- *Any mixture of one arbitrary product distribution, $m-k$ product distributions with support on arbitrary but disjoint faces of the n-cube, and k arbitrary distributions with support on any edges of the n-cube, for any $0 \le k \le m$. In particular:*

- *Any mixture of $m+1$ product distributions with disjoint cubical supports. In consequence, $\text{RBM}_{n,m}$ contains the partition model of any partition in G_{m+1}.*

Restricting the cubical sets of the second item to edges, i.e. pairs of vectors differing in one entry, we see that the above theorem implies the following previously known result, which was shown in [21].

Corollary 4.2 $\text{RBM}_{n,m}$ *contains the following distributions:*

- *Any distribution with a support set that can be covered by $m+1$ pairs of vectors differing in one entry. In particular, this includes:*

- *Any distribution in $\mathcal{P}$ with a support of cardinality smaller than or equal to $m+1$.*

5

Corollary 4.2 implies that an RBM with $m \geq 2^{n-1} - 1$ hidden units is a universal approximator of distributions on $\{0, 1\}^n$, i.e. can approximate any distribution to an arbitrarily good accuracy.

Assume $m + 1 = 2^k$ and let ξ be a partition of $\mathcal{X}$ into $m + 1$ disjoint cubical sets of equal size. Let us denote by $\mathcal{P}_{\xi,1}$ the set of all distributions which can be written as a mixture of $m + 1$ product distributions with support on the elements of ξ. The dimension of $\mathcal{P}_{\xi,1}$ is given by

$$\dim \mathcal{P}_{\xi,1} = (m+1)\log\left(\frac{2^n}{m+1}\right) + m + 1 + n = (m+1) \cdot n + (m+1) + n - (m+1)\log(m+1) .$$

The dimension of the set of visible distribution represented by an RBM is at most equal to the number of paramters, see [21], this is $m \cdot n + m + n$. This means that the class given above has roughly the same dimension of the set of distributions that can be represented. In fact,

$$\dim \mathcal{P}_{\xi,1} - \dim \mathrm{RBM}_{m-1} = n + 1 - (m+1)\log(m+1) .$$

This means that the class of distributions $\mathcal{P}_{\xi,1}$ which by Theorem 4.1 can be represented by $\mathrm{RBM}_{n,m}$ is not contained in $\mathrm{RBM}_{n,m-1}$ when $(m+1)^{m+1} \leq 2^{n+1}$.

Proof of Theorem 4.1 The proof draws on ideas from [15] and [21]. An RBM with no hidden units can represent precisely the independence model, i.e. all product distributions, and in particular any uniform distribution on a face of the n-cube.

Consider an RBM with $m - 1$ hidden units. For any choice of the parameters $W \in \mathbf{R}^{m-1 \times n}, B \in \mathbf{R}^n, C \in \mathbf{R}^{m-1}$ we can write the resulting distribution on the visible units as:

$$p(v) = \frac{\sum_h z(v, h)}{\sum_{v',h'} z(v', h')} , \tag{3}$$

where $z(v, h) = \exp(hWv + Bv + Ch)$. Appending one additional hidden unit, with connection weights w to the visible units and bias c, produces a new distribution which can be written as follows:

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

Consider now any set $I \subseteq [n] := \{1, \ldots, n\}$ and an arbitrary visible vector $u \in \mathcal{X}$. The values of u in the positions $[n] \setminus I$ define a face $F := \{v \in \mathcal{X} : v_i = u_i , \forall i \notin I\}$ of the n-cube of dimension $|I|$. Let $\mathbb{1} := (1, \ldots, 1) \in \mathbf{R}^n$ and denote by $u^{I,0}$ the vector with entries $u_i^{I,0} = u_i, \forall i \notin I$ and $u_i^{I,0} = 0, \forall i \in I$. Let $\lambda^I \in \mathbf{R}^n$ with $\lambda_i^I = 0 , \forall i \notin I$ and let $\lambda_c, a \in \mathbf{R}$. Define the connection weights w and c as follows:

$$w = a\left(u^{I,0} - \frac{1}{2}\mathbb{1}^{I,0}\right) + \lambda^I ,$$

$$c = -a\left(u^{I,0} - \frac{1}{2}\mathbb{1}^{I,0}\right)^\top u + \lambda_c .$$

For this choice and $a \to \infty$ equation (4) yields:

$$p_{w,c}(v) = \begin{cases} \frac{p(v)}{1 + \sum_{v' \in F} \exp(\lambda^I \cdot v' + \lambda_c)p(v')}, & \forall v \notin F \\ \frac{(1 + \exp(\lambda^I \cdot v + \lambda_c))p(v)}{1 + \sum_{v' \in F} \exp(\lambda^I \cdot v' + \lambda_c)p(v')}, & \forall v \in F \end{cases} . \tag{4}$$

If the initial p from equation (3) is such that its restriction to F is a product distribution, then $p(v) = K \exp(\eta^I \cdot v) , \forall v \in F$, where K is a constant and η^I is a vector with $\eta_i^I = 0 , \forall i \notin I$. We can choose $\lambda^I = \beta^I - \eta^I$, and $\exp(\lambda_c) = \alpha \frac{1}{K \sum_{v \in F} \exp(\beta^I \cdot v)}$. For this choice, equation (4) yields:

$$p_{w,c} = (\alpha - 1)p + \alpha\hat{p} ,$$

where $\hat{p}$ is a product distribution with support in F and arbitrary natural parameters β^I, and α is an arbitrary mixture weight in $[0, 1]$. Finally, the product distributions on edges of the cube are arbitrary, see [19] or [21] for details, and hence the restriction of any p to any edge is a product distribution. □

6
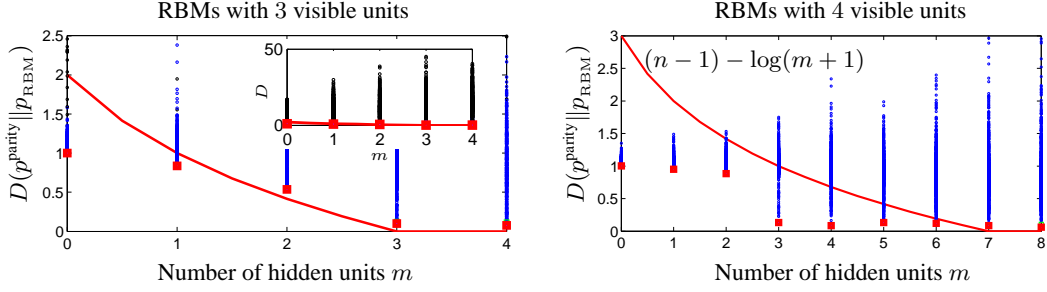
Figure 3: This figure demonstrates our results for $n = 3$ and $n = 4$ visible units. The red curves represent the bounds from Theorem 5.1. We fixed p^{parity} as target distribution, the uniform distribution on binary length n vectors with an even number of ones. The distribution p^{parity} is not the KL-maximizer from $\text{RBM}_{n,m}$, but it is in general difficult to represent. Qualitatively, samples from p^{parity} look like uniformly distributed, and representing p^{parity} requires the maximal number of product mixture components [20, 19]. For both values of n and each $m = 0, \ldots, 2^n/2$ we initialized 500 resp. 1000 RBMs at parameter values chosen uniformly at random in the range $[-10, 10]$. The inset of the left figure shows the resulting KL-divergence $D(p^{\text{parity}} \| p^{\text{rand}}_{\text{RBM}})$ (for $n = 4$ the resulting KL-divergence was larger). Randomly chosen distributions in $\text{RBM}_{n,m}$ are likely to be very far from the target distribution. We trained these randomly initialized RBMs using CD for 500 training epochs, learning rate 1 and a list of even parity vectors as training data. The result after training is given by the blue circles. After training the RBMs the result is often not better than the uniform distribution, for which $D(p^{\text{parity}} \| \frac{\mathbb{1}}{|\{0,1\}^n|}) = 1$. For each m, the best set of parameters after training was used to initialize a further CD training with a smaller learning rate (green squares, mostly covered) followed by a short maximum likelihood gradient ascent (red filled squares).

5 Maximal Approximation Errors of RBMs

Let $m < 2^{n-1} - 1$. By Theorem 4.1 all partition models for partitions of $\{0,1\}^n$ into $m + 1$ cubical sets are contained in $\text{RBM}_{n,m}$. Applying Corollary 3.3 to such a partition where the cardinality of all blocks is at most $2^{n - \lfloor \log(m+1) \rfloor}$ yields the bound $D_{\text{RBM}_{n,m}} \leq n - \lfloor \log(m+1) \rfloor$. Similarly, using mixtures of product distributions, Theorem 4.1 and Corollary 3.4 imply the smaller bound $D_{\text{RBM}_{n,m}} \leq n - 1 - \lfloor \log(m+1) \rfloor$. In this section we derive an improved bound which strictly decreases, as m increases, until 0 is reached.

Theorem 5.1 *Let $m \leq 2^{n-1} - 1$. Then the maximal Kullback-Leibler divergence from any distribution on $\{0,1\}^n$ to $\text{RBM}_{n,m}$ is upper bounded by*

$$\max_{p \in \mathcal{P}} D(p \| \text{RBM}_{n,m}) \leq (n-1) - \log(m+1) .$$

Conversely, given an error tolerance $0 \leq \epsilon \leq 1$, the choice $m \geq 2^{(n-1)(1-\epsilon)} - 1$ ensures a sufficiently rich RBM model that satisfies $D_{\text{RBM}_{n,m}} \leq \epsilon D_{\text{RBM}_{n,0}}$.

For $m = 2^{n-1} - 1$ the error vanishes, corresponding to the fact that an RBM with that many hidden units is a universal approximator. In Figure 3 we use computer experiments to illustrate Theorem 5.1. The proof makes use of the following lemma:

Lemma 5.2 *Let $n_1, \ldots, n_m \geq 0$ such that $2^{n_1} + \cdots + 2^{n_m} = 2^n$. Let $\mathcal{M}$ be the union of all mixtures of independent models corresponding to all cubical partitions of $\mathcal{X}$ into blocks of cardinalities $2^{n_1}, \ldots, 2^{n_m}$. Then $D_{\mathcal{M}} \leq \sum_{i:n_i > 1} \frac{n_i - 1}{2^{n-n_i}}$.*

Proof of Lemma 5.2 The proof is by induction on n. If $n = 1$, then $m = 1$ or $m = 2$, and in both cases it is easy to see that the inequality holds (both sides vanish). If $n > 1$, then order the n_i such that $n_1 \geq n_2 \geq \cdots \geq n_m \geq 0$. Without loss of generality assume $m > 1$.

Let $p \in \mathcal{P}(\mathcal{X})$, and let $\mathcal{Y}$ be a cubical subset of $\mathcal{X}$ of cardinality 2^{n-1} such that $p(\mathcal{Y}) \leq \frac{1}{2}$. Since the numbers $2^{n_1} + \cdots + 2^{n_i}$ for $i = 1, \ldots, m$ contain all multiples of 2^{n_1} up to 2^n and $2^n/2^{n_1}$ is even, there exists k such that $2^{n_1} + \cdots + 2^{n_k} = 2^{n-1} = 2^{n_{k+1}} + \cdots + 2^{n_m}$.

7

Let $\mathcal{M}'$ be the union of all mixtures of independence models corresponding to all cubical partitions $\xi = \{\mathcal{X}_1, \ldots, \mathcal{X}_m\}$ of $\mathcal{X}$ into m blocks of cardinalities $n_1, \ldots, n_m$ such that $\mathcal{X}_1 \cup \cdots \cup \mathcal{X}_k = \mathcal{Y}$. In the following, the symbol $\sum_i'$ shall denote summation over all indices i such that $n_i > 1$. By induction

$$D(p\|\mathcal{M}) \le D(p\|\mathcal{M}') \le p(\mathcal{Y}) \sum_{i=1}^{k}{}' \frac{n_i - 1}{2^{n-1-n_i}} + p(\mathcal{X} \setminus \mathcal{Y}) \sum_{j=k+1}^{m}{}' \frac{n_j - 1}{2^{n-1-n_j}} \ . \tag{5}$$

There exist $j_1 = k + 1 < j_2 < \cdots < j_k < j_{k+1} = m + 1$ such that $2^{n_i} = 2^{n_{j_i}} + \cdots + 2^{n_{j_{i+1}-1}}$ for all $i \le k$. Note that

$$\sum_{j=j_i}^{j_{i+1}}{}' \frac{n_j - 1}{2^{n-1-n_j}} \le \frac{n_i - 1}{2^{n-1}} (2^{n_{j_i}} + \cdots + 2^{n_{j_{i+1}-1}}) = \frac{n_i - 1}{2^{n-1-n_i}} \ ,$$

and therefore

$$(\tfrac{1}{2} - p(\mathcal{Y})) \frac{n_i - 1}{2^{n-1-n_i}} + (\tfrac{1}{2} - p(\mathcal{X} \setminus \mathcal{Y})) \sum_{j=j_i}^{j_{i+1}-1}{}' \frac{n_j - 1}{2^{n-1-n_j}} \ge 0 \ .$$

Adding these terms for $i = 1, \ldots, k$ to the right hand side of equation (5) yields

$$D(p\|\mathcal{M}) \le \frac{1}{2} \sum_{i=1}^{k}{}' \frac{n_i - 1}{2^{n-1-n_i}} + \frac{1}{2} \sum_{j=k+1}^{m}{}' \frac{n_j - 1}{2^{n-1-n_j}},$$

from which the assertions follow. $\qquad\square$

Proof of Theorem 5.1 From Theorem 4.1 we know that $\mathrm{RBM}_{n,m}$ contains the union $\mathcal{M}$ of all mixtures of independent models corresponding to all partitions with up to $m + 1$ cubical blocks. Hence, $D_{\mathrm{RBM}_{n,m}} \le D_{\mathcal{M}}$. Let $k = n - \lfloor \log(m+1) \rfloor$ and $l = 2m + 2 - 2^{n-k+1} \ge 0$; then $l 2^{k-1} + (m+1-l) 2^k = 2^n$. Lemma 5.2 with $n_1 = \cdots = n_l = k - 1$ and $n_{l+1} = \cdots = n_{m+1} = k$ implies

$$D_{\mathcal{M}} \le \frac{l(k-2)}{2^{n-k+1}} + \frac{(m+1-l)(k-1)}{2^{n-k}} = k - \frac{m+1}{2^{n-k}}.$$

The assertion follows from $\log(m+1) \le (n - k) + \frac{m+1}{2^{n-k}} - 1$, where $\log(1 + x) \le x$ for all $x > 0$ was used. $\qquad\square$

6 Conclusion

We studied the expressive power of the Restricted Boltzmann Machine model with n visible and m hidden units. We presented a hierarchy of explicit classes of probability distributions that an RBM can represent. These classes include large collections of mixtures of $m + 1$ product distributions. In particular any mixture of an arbitrary product distribution and m further product distributions with disjoint supports. The geometry of these submodels is easier to study than that of the RBM models, while these subsets still capture many of the distributions contained in the RBM models. Using these results we derived bounds for the approximation errors of RBMs. We showed that it is always possible to reduce the error to at most $(n - 1) - \log(m + 1)$. That is, given any target distribution, there is a distribution within the RBM model for which the Kullback-Leibler divergence between both is not larger than that number. Our results give a theoretical basis for selecting the size of an RBM which accounts for a desired error tolerance.

Computer experiments showed that the bound captures the order of magnitude of the true approximation error, at least for small examples. However, learning may not always find the best approximation, resulting in an error that may well exceed our bound.

Acknowledgments

Nihat Ay acknowledges support by the Santa Fe Institute.

8

References

[1] N. Ay and A. Knauf. Maximizing multi-information. *Kybernetika*, 42:517–538, 2006.

[2] N. Ay, G. Montúfar, and J. Rauh. Selection criteria for neuromanifolds of stochastic dynamics. *International Conference on Cognitive Neurodynamics*, 2011.

[3] N. Ay and T. Wennekers. Dynamical properties of strongly interacting Markov chains. *Neural Networks*, 16:1483–1497, 2003.

[4] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. *NIPS*, 2007.

[5] L. Brown. *Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory*. Inst. Math. Statist., Hayworth, CA, USA, 1986.

[6] M. A. Carreira-Perpiñan and G. E. Hinton. On contrastive divergence learning. In *Proceedings of the 10-th International Workshop on Artificial Intelligence and Statistics*, 2005.

[7] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2006.

[8] M. A. Cueto, J. Morton, and B. Sturmfels. Geometry of the Restricted Boltzmann Machine. In M. A. G. Viana and H. P. Wynn, editors, *Algebraic methods in statistics and probability II, AMS Special Session*. AMS, 2010.

[9] P. Diaconis and D. Freedman. Finite exchangeable sequences. *Ann. Probab.*, 8:745–764, 1980.

[10] Y. Freund and D. Haussler. Unsupervised learning of distributions on binary vectors using 2-layer networks. *NIPS*, pages 912–919, 1992.

[11] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Comput.*, 14:1771–1800, 2002.

[12] G. E. Hinton. A practical guide to training Restricted Boltzmann Machines, version 1. Technical report, UTML2010-003, University of Toronto, 2010.

[13] G. E. Hinton, S. Osindero, and Y. Teh. A fast learning algorithm for Deep Belief Nets. *Neural Comput.*, 18:1527–1554, 2006.

[14] S. Kullback and R. Leibler. On information and sufficiency. *Ann. Math. Stat.*, 22:79–86, 1951.

[15] N. Le Roux and Y. Bengio. Representational power of Restricted Boltzmann Machines and Deep Belief Networks. *Neural Comput.*, 20(6):1631–1649, 2008.

[16] N. Le Roux and Y. Bengio. Deep Belief Networks are compact universal approximators. *Neural Comput.*, 22:2192–2207, 2010.

[17] B. Lindsay. *Mixture models: theory, geometry, and applications*. Inst. Math. Statist., 1995.

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

[19] G. Montúfar. Mixture decompositions using a decomposition of the sample space. *ArXiv 1008.0204*, 2010.

[20] G. Montúfar. Mixture models and representational power of RBMs, DBNs and DBMs. *NIPS Deep Learning and Unsupervised Feature Learning Workshop*, 2010.

[21] G. Montúfar and N. Ay. Refinements of universal approximation results for Deep Belief Networks and Restricted Boltzmann Machines. *Neural Comput.*, 23(5):1306–1319, 2011.

[22] J. Rauh. *Finding the maximizers of the information divergence from an exponential family*. PhD thesis, Universität Leipzig, 2011.

[23] J. Rauh, T. Kahle, and N. Ay. Support sets of exponential families and oriented matroids. *Int. J. Approx. Reason.*, 52(5):613–626, 2011.

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

[25] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning)*. MIT Press, March 1998.

[26] K. G. Zahedi, N. Ay, and R. Der. Higher coordination with less control – a result of infromation maximization in the sensori-motor loop. *Adaptive Behavior*, 18(3-4):338–355, 2010.

9