Journal of Machine Learning Research 16 (2015) 2405-2436 Submitted 7/14; Revised 3/15; Published 12/15

Geometry and Expressive Power of Conditional Restricted Boltzmann Machines

Guido Montúfar MONTUFAR@MIS.MPG.DE
Max Planck Institute for Mathematics in the Sciences
04103 Leipzig, Germany

Nihat Ay NAY@MIS.MPG.DE
Max Planck Institute for Mathematics in the Sciences
04103 Leipzig, Germany

Department of Mathematics and Computer Science
Leipzig University
04009 Leipzig, Germany

Santa Fe Institute
Santa Fe, NM 87501, USA

Keyan Ghazi-Zahedi ZAHEDI@MIS.MPG.DE
Max Planck Institute for Mathematics in the Sciences
04103 Leipzig, Germany

Editor: Ruslan Salakhutdinov

Abstract

Conditional restricted Boltzmann machines are undirected stochastic neural networks with a layer of input and output units connected bipartitely to a layer of hidden units. These networks define models of conditional probability distributions on the states of the output units given the states of the input units, parameterized by interaction weights and biases. We address the representational power of these models, proving results on their ability to represent conditional Markov random fields and conditional distributions with restricted supports, the minimal size of universal approximators, the maximal model approximation errors, and on the dimension of the set of representable conditional distributions. We contribute new tools for investigating conditional probability models, which allow us to improve the results that can be derived from existing work on restricted Boltzmann machine probability models.

Keywords: conditional restricted Boltzmann machine, universal approximation, Kullback-Leibler approximation error, expected dimension

1. Introduction

Restricted Boltzmann Machines (RBMs) (Smolensky, 1986; Freund and Haussler, 1994) are generative probability models defined by undirected stochastic networks with bipartite interactions between visible and hidden units. These models are well-known in machine learning applications, where they are used to infer distributed representations of data and to train the layers of deep neural networks (Hinton et al., 2006; Bengio, 2009). The restricted connectivity of these networks allows to train them efficiently on the basis of cheap inference

and finite Gibbs sampling (Hinton, 2002, 2012), even when they are defined with many units and parameters. An RBM defines Gibbs-Boltzmann probability distributions over the observable states of the network, depending on the interaction weights and biases. An introduction is offered by Fischer and Igel (2012). The expressive power of these probability models has attracted much attention and has been studied in numerous papers, treating, in particular, their universal approximation properties (Younes, 1996; Le Roux and Bengio, 2008; Montúfar and Ay, 2011), approximation errors (Montúfar et al., 2011), efficiency of representation (Martens et al., 2013; Montúfar and Morton, 2015), and dimension (Cueto et al., 2010).

In certain applications, it is preferred to work with conditional probability distributions, instead of joint probability distributions. For example, in a classification task, the conditional distribution may be used to indicate a belief about the class of an input, without modeling the probability of observing that input; in sensorimotor control, it can describe a stochastic policy for choosing actions based on world observations; and in the context of information communication, to describe a channel. RBMs naturally define models of conditional probability distributions, called conditional restricted Boltzmann machines (CRBMs). These models inherit many of the nice properties of RBM probability models, such as the cheap inference and efficient training. Specifically, a CRBM is defined by clamping the states of an *input* subset of the visible units of an RBM. For each input state one obtains a conditioned distribution over the states of the *output* visible units. See Figure 1 for an illustration of this architecture. This kind of conditional models and slight variants thereof have seen success in many applications; for example, in classification (Larochelle and Bengio, 2008), collaborative filtering (Salakhutdinov et al., 2007), motion modeling (Taylor et al., 2007; Zeiler et al., 2009; Mnih et al., 2011; Sutskever and Hinton, 2007), and reinforcement learning (Sallans and Hinton, 2004).

So far, however, there is not much theoretical work addressing the expressive power of CRBMs. We note that it is relatively straightforward to obtain some results on the expressive power of CRBMs from the existing theoretical work on RBM probability models. Nevertheless, an accurate analysis requires to take into account the specificities of the conditional case. Formally, a CRBM is a collection of RBMs, with one RBM for each possible input value. These RBMs differ in the biases of the hidden units, as these are influenced by the input values. However, these hidden biases are not independent for all different inputs, and, moreover, the same interaction weights and biases of the visible units are shared for all different inputs. This sharing of parameters draws a substantial distinction of CRBM models from independent tuples of RBM models.

In this paper we address the representational power of CRBMs, contributing theoretical insights to the optimal number of hidden units. Our focus lies on the classes of conditional distributions that can possibly be represented by a CRBM with a fixed number of inputs and outputs, depending on the number of hidden units. Having said this, we do not discuss the problem of finding the optimal parameters that give rise to a desired conditional distribution (although our derivations include an algorithm that does this), nor problems related to incomplete knowledge of the target conditional distributions and generalization errors. A number of training methods for CRBMs have been discussed in the references listed above, depending on the concrete applications. The problems that we deal with here are the following: 1) are distinct parameters of the model mapped to distinct conditional

distributions; what is the smallest number of hidden units that suffices for obtaining a model that can 2) approximate any target conditional distribution arbitrarily well (a universal approximator); 3) approximate any target conditional distribution without exceeding a given error tolerance; 4) approximate selected classes of conditional distributions arbitrarily well? We provide non-trivial solutions to all of these problems. We focus on the case of binary units, but the main ideas extend to the case of discrete non-binary units.

This paper is organized as follows. Section 2 contains formal definitions and elementary properties of CRBMs. Section 3 investigates the geometry of CRBM models in three subsections. In Section 3.1 we study the dimension of the sets of conditional distributions represented by CRBMs and show that in most cases this is the dimension expected from counting parameters (Theorem 4). In Section 3.2 we address the universal approximation problem, deriving upper and lower bounds on the minimal number of hidden units that suffices for this purpose (Theorem 7). In Section 3.3 we analyze the maximal approximation errors of CRBMs (assuming optimal parameters) and derive an upper bound for the minimal number of hidden units that suffices to approximate every conditional distribution within a given error tolerance (Theorem 11). Section 4 investigates the expressive power of CRBMs in two subsections. In Section 4.1 we describe how CRBMs can represent natural families of conditional distributions that arise in Markov random fields (Theorem 14). In Section 4.2 we study the ability of CRBMs to approximate conditional distributions with restricted supports. This section addresses, especially, the approximation of deterministic conditional distributions (Theorem 21). In Section 5 we offer a discussion and an outlook. In order to present the main results in a concise way, we have deferred all proofs to the appendices. Nonetheless, we think that the proofs are interesting in their own right, and we have prepared them with a fair amount of detail.

2. Definitions

We will denote the set of probability distributions on $\{0,1\}^n$ by Δ_n. A probability distribution $p \in \Delta_n$ is a vector of 2^n non-negative entries $p(y)$, $y \in \{0,1\}^n$, adding to one, $\sum_{y\in\{0,1\}^n} p(y) = 1$. The set Δ_n is a (2^n-1)-dimensional simplex in $\mathbb{R}^{2^n}$.

We will denote the set of conditional distributions of a variable $y \in \{0,1\}^n$, given another variable $x \in \{0,1\}^k$, by $\Delta_{k,n}$. A conditional distribution $p(\cdot|\cdot) \in \Delta_{k,n}$ is a $2^k \times 2^n$ row-stochastic matrix with rows $p(\cdot|x) \in \Delta_n$, $x \in \{0,1\}^k$. The set $\Delta_{k,n}$ is a $2^k(2^n-1)$-dimensional polytope in $\mathbb{R}^{2^k\times 2^n}$. It can be regarded as the 2^k-fold Cartesian product $\Delta_{k,n} = \Delta_n \times \cdots \times \Delta_n$, where there is one probability simplex Δ_n for each possible input state $x \in \{0,1\}^k$. We will use the abbreviation $[N] := \{1,\ldots,N\}$, where N is a natural number.

Definition 1 The conditional restricted Boltzmann machine (CRBM) with k input units, n output units, and m hidden units, denoted $\mathrm{RBM}^k_{n,m}$, is the set of all conditional distributions in $\Delta_{k,n}$ that can be written as

$$p(y|x) = \frac{1}{Z(W,b,Vx+c)} \sum_{z\in\{0,1\}^m} \exp(z^\top Vx + z^\top Wy + b^\top y + c^\top z), \quad \forall y \in \{0,1\}^n, x \in \{0,1\}^k,$$

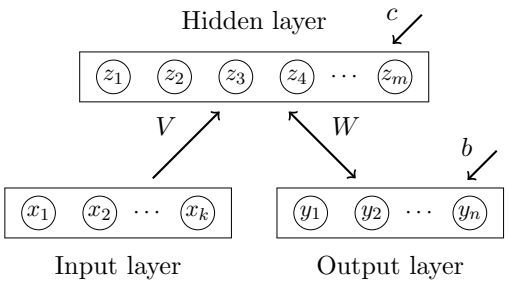

Figure 1: Architecture of a CRBM. An RBM is the special case with $k = 0$.

with normalization function

$$Z(W, b, Vx + c) = \sum_{y \in \{0,1\}^n} \sum_{z \in \{0,1\}^m} \exp(z^\top Vx + z^\top Wy + b^\top y + c^\top z), \quad \forall x \in \{0,1\}^k.$$

Here, x, y, and z are column state vectors of the k input units, n output units, and m hidden units, respectively, and $^\top$ denotes transposition. The parameters of this model are the matrices of interaction weights $V \in \mathbb{R}^{m \times k}$, $W \in \mathbb{R}^{m \times n}$ and the vectors of biases $b \in \mathbb{R}^n$, $c \in \mathbb{R}^m$. When there are no input units ($k = 0$), the model $\mathrm{RBM}^k_{n,m}$ reduces to the restricted Boltzmann machine probability model with n visible units and m hidden units, denoted $\mathrm{RBM}_{n,m}$.

We can view $\mathrm{RBM}^k_{n,m}$ as a collection of 2^k restricted Boltzmann machine probability models with shared parameters. For each input $x \in \{0,1\}^k$, the output distribution $p(\cdot|x)$ is the probability distribution represented by $\mathrm{RBM}_{n,m}$ for the parameters $W, b, (Vx + c)$. All $p(\cdot|x)$ have the same interaction weights W, the same biases b for the visible units, and differ only in the biases $(Vx + c)$ for the hidden units. The joint behavior of these distributions with shared parameters is not trivial.

The model $\mathrm{RBM}^k_{n,m}$ can also be regarded as representing block-wise normalized versions of the joint probability distributions represented by $\mathrm{RBM}_{n+k,m}$. Namely, a joint distribution $p \in \mathrm{RBM}_{n+k,m} \subseteq \Delta_{k+n}$ is an array with entries $p(x, y)$, $x \in \{0,1\}^k$, $y \in \{0,1\}^n$. Conditioning p on x is equivalent to considering the normalized x-th row $p(y|x) = p(x, y)/\sum_{y'} p(x, y')$, $y \in \{0,1\}^n$.

3. Geometry of Conditional Restricted Boltzmann Machines

In this section we investigate three basic questions about the geometry of CRBM models. First, what is the dimension of a CRBM model? Second, how many hidden units does a CRBM need in order to be able to approximate every conditional distribution arbitrarily well? Third, how accurate are the approximations of a CRBM, depending on the number of hidden units?

3.1 Dimension

The model $\mathrm{RBM}^k_{n,m}$ is defined by marginalizing out the hidden units of a graphical model. This implies that several choices of parameters may represent the same conditional distri-

butions. In turn, the dimension of the set of representable conditional distributions may be smaller than the number of model parameters, in principle.

When the dimension of $\mathrm{RBM}^k_{n,m}$ is equal to $\min\{(k+n)m+n+m, 2^k(2^n-1)\}$, which is the minimum of the number of parameters and the dimension of the ambient polytope of conditional distributions, the CRBM model is said to have the *expected dimension*. In this section we show that $\mathrm{RBM}^k_{n,m}$ has the expected dimension for most triplets (k,n,m). In particular, we show that this holds in all practical cases, where the number of hidden units m is smaller than exponential with respect to the number of visible units $k+n$.

The dimension of a parametric model is given by the maximum of the rank of the Jacobian of its parameterization (assuming mild differentiability conditions). Computing the rank of the Jacobian is not easy in general. A resort is to compute the rank only in the limit of large parameters, which corresponds to considering a piece-wise linearized version of the original model, called the *tropical model*. Cueto et al. (2010) used this approach to study the dimension of RBM probability models. Here we apply their ideas to address the dimension of CRBM conditional models.

The following functions from coding theory will be useful for phrasing the results:

Definition 2 Let $A(n,d)$ denote the cardinality of a largest subset of $\{0,1\}^n$ whose elements are at least Hamming distance d apart. Let $K(n,d)$ denote the smallest cardinality of a set such that every element of $\{0,1\}^n$ is at most Hamming distance d apart from that set.

Cueto et al. (2010) showed that $\dim(\mathrm{RBM}_{n,m}) = nm+n+m$ for $m+1 \leq A(n,3)$, and $\dim(\mathrm{RBM}_{n,m}) = 2^n-1$ for $m \geq K(n,1)$. It is known that $A(n,3) \geq 2^{n-\lceil \log_2(n+1)\rceil}$ and $K(n,1) \leq 2^{n-\lfloor \log_2(n+1)\rfloor}$. In turn, for most pairs (n,m) the probability model $\mathrm{RBM}_{n,m}$ has the expected dimension (although for many values of n there is a range of values of m where the results are inconclusive about this). Noting that $\dim(\mathrm{RBM}^k_{n,m}) \geq \dim(\mathrm{RBM}_{k+n,m}) - (2^k-1)$, these results on the dimension of RBM probability models directly imply following bounds on the dimension of CRBM models:

Proposition 3 *The conditional model* $\mathrm{RBM}^k_{n,m}$ *satisfies the following:*

- $\dim(\mathrm{RBM}^k_{n,m}) \geq (n+k)m+n+m+k-(2^k-1)$ *for* $m+1 \leq A(k+n,3)$.
- $\dim(\mathrm{RBM}^k_{n,m}) = 2^k(2^n-1)$ *for* $m \geq K(k+n,1)$.

This result shows that, when $m \geq K(k+n,1)$, the CRBM model has the maximum possible dimension, equal to the dimension of $\Delta_{k,n}$. In all other cases, however, the dimension bounds are too loose and do not allow us to conclude whether or not the CRBM model has the expected dimension. Hence we need to study the conditional model in more detail. We obtain the following result:

Theorem 4 *The conditional model* $\mathrm{RBM}^k_{n,m}$ *satisfies the following:*

- $\dim(\mathrm{RBM}^k_{n,m}) = (k+n)m+n+m$ *for* $m+1 \leq A(k+n,4)$.
- $\dim(\mathrm{RBM}^k_{n,m}) = 2^k(2^n-1)$ *for* $m \geq K(k+n,1)$.

We note the following practical version of the theorem, which results from inserting appropriate bounds on the functions A and K:

Corollary 5 *The conditional model* $\mathrm{RBM}^k_{n,m}$ *has the expected dimension in the following cases:*

- $\dim(\mathrm{RBM}^k_{n,m}) = (n+k)m + n + m$ *for* $m \leq 2^{(k+n)-\lfloor \log_2((k+n)^2-(k+n)+2)\rfloor}$.
- $\dim(\mathrm{RBM}^k_{n,m}) = 2^k(2^n-1)$ *for* $m \geq 2^{(k+n)-\lfloor \log_2(k+n+1)\rfloor}$.

These results show that, in all cases of practical interest, where m is less than exponential in $k+n$, the dimension of the CRBM model is indeed equal to the number of model parameters. In all these cases, almost every conditional distribution that can be represented by the model is represented by at most finitely many different choices of parameters. We should note that there is an interval of exponentially large values of m where the results remain inconclusive, namely the interval $A(k+n,4) \leq m < K(k+n,1)$. This is similar to the gap already mentioned above for RBM probability models and poses interesting theoretical problems (see also Montúfar and Morton, 2015).

On the other hand, the dimension alone is not very informative about the ability of a model to approximate target distributions. In particular, it may be that a high dimensional model covers only a tiny fraction of the set of all conditional distributions, or also that a low dimensional model can approximate any target conditional relatively well. We address the minimal dimension and number of parameters of a universal approximator in the next section. In the subsequent section we address the approximation errors depending on the number of parameters.

3.2 Universal Approximation

In this section we ask for the smallest number of hidden units m for which the model $\mathrm{RBM}^k_{n,m}$ can approximate every conditional distribution from $\Delta_{k,n}$ arbitrarily well.

Note that each conditional distribution $p(y|x)$ can be identified with the set of joint distributions of the form $r(x,y) = q(x)p(y|x)$, with strictly positive marginals $q(x)$. In particular, by fixing a marginal distribution, we obtain an identification of $\Delta_{k,n}$ with a subset of Δ_{k+n}. Figure 2 illustrates this identification in the case $n = k = 1$ and $q(0) = q(1) = \frac{1}{2}$.

This implies that universal approximators of joint probability distributions define universal approximators of conditional distributions. We know that $\mathrm{RBM}_{n+k,m}$ is a universal approximator whenever $m \geq \frac{1}{2}2^{k+n} - 1$ (see Montúfar and Ay, 2011), and therefore:

Proposition 6 *The model* $\mathrm{RBM}^k_{n,m}$ *can approximate every conditional distribution from* $\Delta_{k,n}$ *arbitrarily well whenever* $m \geq \frac{1}{2}2^{k+n} - 1$.

This improves previous results by Younes (1996) and van der Maaten (2011). On the other hand, since conditional models do not need to model the input distributions, in principle it is possible that $\mathrm{RBM}^k_{n,m}$ is a universal approximator even if $\mathrm{RBM}_{n+k,m}$ is not a universal approximator. In fact, we obtain the following improvement of Proposition 6, which does not follow from corresponding results for RBM probability models:

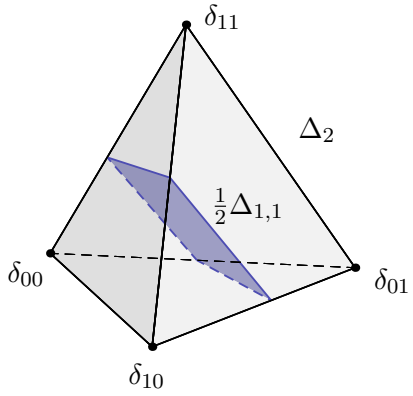

Figure 2: The polytope of conditional distributions $\Delta_{1,1}$ embedded in the simplex of probability distributions Δ_2.

Theorem 7 *The model* $\mathrm{RBM}^k_{n,m}$ *can approximate every conditional distribution from* $\Delta_{k,n}$ *arbitrarily well whenever*

$$m \geq \begin{cases} \frac{1}{2}2^k(2^n-1), & \text{if } k \geq 1 \\ \frac{3}{8}2^k(2^n-1)+1, & \text{if } k \geq 3 \\ \frac{1}{4}2^k(2^n-1+1/30), & \text{if } k \geq 21 \end{cases}.$$

In fact, the model $\mathrm{RBM}^k_{n,m}$ *can approximate every conditional distribution from* $\Delta_{k,n}$ *arbitrarily well whenever* $m \geq 2^k K(r)(2^n-1)+2^{S(r)}P(r)$*, where* r *is any natural number satisfying* $k \geq 1+\cdots+r =: S(r)$*, and* K *and* P *are functions (defined in Lemma 30 and Proposition 32) which tend to approximately* 0.2263 *and* 0.0269*, respectively, as* r *tends to infinity.*

We note the following weaker but practical version of Theorem 7:

Corollary 8 *Let* $k \geq 1$*. The model* $\mathrm{RBM}^k_{n,m}$ *can approximate every conditional distribution from* $\Delta_{k,n}$ *arbitrarily well whenever* $m \geq \frac{1}{2}2^k(2^n-1) = \frac{1}{2}2^{k+n} - \frac{1}{2}2^k$*.*

These results are significant, because they reduce the bounds following from universal approximation results for probability models by an additive term of order 2^k, which corresponds precisely to the order of parameters needed to model the input distributions.

As expected, the asymptotic behavior of the theorem's bound is exponential in the number of input and output units. This lies in the nature of the universal approximation property. A crude lower bound on the number of hidden units that suffices for universal approximation can be obtained by comparing the number of parameters of the model and the dimension of the conditional polytope:

Proposition 9 *If the model* $\mathrm{RBM}^k_{n,m}$ *can approximate every conditional distribution from* $\Delta_{k,n}$ *arbitrarily well, then necessarily* $m \geq \frac{1}{(n+k+1)}(2^k(2^n-1)-n)$*.*

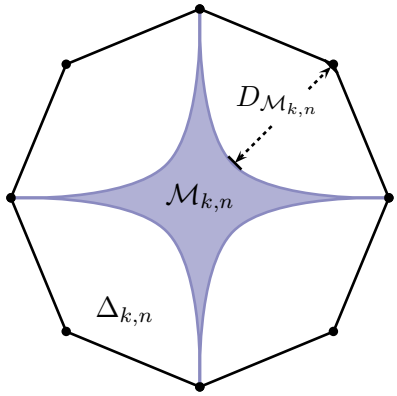

Figure 3: Schematic illustration of the maximal approximation error of a model of conditional distributions $\mathcal{M}_{k,n} \subseteq \Delta_{k,n}$.

The results presented above highlight the fact that CRBM universal approximation may be possible with a drastically smaller number of hidden units than RBM universal approximation, for the same number of visible units. However, even with these reductions the universal approximation property requires an enormous number of hidden units. In order to provide a more informative description of the approximation capabilities of CRBMs, in the next section we investigate how the maximal approximation error decreases as hidden units are added to the model.

3.3 Maximal Approximation Errors

From a practical perspective it is not necessary to approximate conditional distributions arbitrarily well, but fair approximations suffice. This can be especially important if the number of required hidden units grows disproportionately with the quality of the approximation. In this section we investigate the maximal approximation errors of CRBMs depending on the number of hidden units. Figure 3 gives a schematic illustration of the maximal approximation error of a conditional model.

The Kullback-Leibler divergence of two probability distributions p and q in Δ_{k+n} is given by

$$\begin{aligned} D(p\|q) :=& \sum_x \sum_y p(x)p(y|x) \log \frac{p(x)p(y|x)}{q(x)q(y|x)} \\ =& D(p_X\|q_X) + \sum_x p(x) D(p(\cdot|x)\|q(\cdot\|x)), \end{aligned}$$

where $p_X = \sum_{y\in\{0,1\}^n} p(x,y)$ denotes the marginal distribution over $x \in \{0,1\}^k$. The divergence of two conditional distributions $p(\cdot|\cdot)$ and $q(\cdot|\cdot)$ in $\Delta_{k,n}$ is given by

$$D(p(\cdot|\cdot)\|q(\cdot|\cdot)) := \sum_x u_X(x) D(p(\cdot|x)\|q(\cdot|x)),$$

where u_X denotes the uniform distribution over x. Even if the divergence between two joint distributions does not vanish, the divergence between their conditional distributions may vanish.

Consider a model $\mathcal{M}_{k+n} \subseteq \Delta_{k+n}$ of joint probability distributions and a corresponding model $\mathcal{M}_{k,n} \subseteq \Delta_{k,n}$ of conditional distributions. More precisely, $\mathcal{M}_{k,n}$ consists of all conditional distributions of the form $q(y|x) = q(x,y)/\sum_{y' \in \{0,1\}^n} q(x,y')$, for all $y \in \{0,1\}^n$ and $x \in \{0,1\}^k$, where $q(\cdot,\cdot)$ is any joint probability distribution from $\mathcal{M}_{k+n}$ satisfying $\sum_{y' \in \{0,1\}^n} q(x,y') > 0$, for all $x \in \{0,1\}^k$. The divergence from a conditional distribution $p(\cdot|\cdot) \in \Delta_{k,n}$ to the model $\mathcal{M}_{k,n}$ is given by

$$D(p(\cdot|\cdot)\|\mathcal{M}_{k,n}) := \inf_{q \in \mathcal{M}_{k,n}} D(p(\cdot|\cdot)\|q(\cdot|\cdot)) = \inf_{q \in \mathcal{M}_{k+n}} D(u_X p(\cdot|\cdot)\|q) - D(u_X\|q_X).$$

In turn, the maximum of the divergence from a conditional distribution to $\mathcal{M}_{k,n}$ satisfies

$$D_{\mathcal{M}_{k,n}} := \max_{p(\cdot|\cdot) \in \Delta_{k,n}} D(p(\cdot|\cdot)\|\mathcal{M}_{k,n}) \leq \max_{p \in \Delta_{k+n}} D(p\|\mathcal{M}_{k+n}) =: D_{\mathcal{M}_{k+n}}.$$

Hence we can bound the maximal divergence of a CRBM by the maximal divergence of an RBM (studied by Montúfar et al., 2011) and obtain the following:

Proposition 10 *If $m \leq 2^{(n+k)-1} - 1$, then the divergence from any conditional distribution $p(\cdot|\cdot) \in \Delta_{k,n}$ to the model $\mathrm{RBM}^k_{n,m}$ is bounded by*

$$D_{\mathrm{RBM}^k_{n,m}} \leq D_{\mathrm{RBM}_{k+n,m}} \leq (n+k) - \lfloor \log_2(m+1) \rfloor - \frac{m+1}{2^{\lfloor \log_2(m+1) \rfloor}}.$$

This proposition implies the universal approximation result from Proposition 6 as the special case with vanishing approximation error, but it does not imply Theorem 7 in the same way. Taking more specific properties of the conditional model into account, we can improve the proposition and obtain the following:

Theorem 11 *Let $l \in [n]$. The divergence from any conditional distribution in $\Delta_{k,n}$ to the model $\mathrm{RBM}^k_{n,m}$ is bounded from above by*

$$D_{\mathrm{RBM}^k_{n,m}} \leq n - l, \quad \text{whenever } m \geq \begin{cases} \frac{1}{2}2^k(2^l - 1), & \text{if } k \geq 1 \\ \frac{3}{8}2^k(2^l - 1) + 1, & \text{if } k \geq 3 \\ \frac{1}{4}2^k(2^l - 1 + 1/30), & \text{if } k \geq 21 \end{cases}.$$

In fact, the divergence from any conditional distribution in $\Delta_{k,n}$ to $\mathrm{RBM}^k_{n,m}$ is bounded from above by $D_{\mathrm{RBM}^k_{n,m}} \leq n - l$, where l is the largest integer with $m \geq 2^{k - S(r)}F(r)(2^l - 1) + R(r)$.

In plain terms, this theorem shows that the worst case approximation errors of CRBMs decrease at least with the logarithm of the number of hidden units. Given an error tolerance, we can use these bounds to find a sufficient number of hidden units that guarantees approximations within this error tolerance. Furthermore, the result implies the universal approximation result from Theorem 7 as the special case with vanishing approximation error. We note the following weaker but practical version of Theorem 11 (analogue to Corollary 8):

Corollary 12 *Let $k \geq 1$ and $l \in [n]$. The divergence from any conditional distribution in $\Delta_{k,n}$ to the model $\mathrm{RBM}^k_{n,m}$ is bounded from above by $D_{\mathrm{RBM}^k_{n,m}} \leq n - l$, whenever $m \geq \frac{1}{2}2^k(2^l - 1)$.*

In this section we have discussed the worst case approximation errors of CRBMs. On the other hand, in practice one is not interested in approximating all possible conditional distributions, but only special classes. One can expect that CRBMs can approximate certain classes of conditional distributions better than others. This is the subject of the next section.

4. Representation of Special Classes of Conditional Models

In this section we ask about the classes of conditional distributions that can be compactly represented by CRBMs and whether CRBMs can approximate interesting conditional distributions using only a moderate number of hidden units.

The first part of the question is about familiar classes of conditional distributions that can be expressed in terms of CRBMs, which in turn would allow us to compare CRBMs with other models and to develop a more intuitive picture of Definition 1.

The second part of the question clearly depends on the specific problem at hand. Nonetheless, some classes of conditional distributions may be considered generally interesting, as they contain solutions to all instances of certain classes of problems. An example is the class of deterministic conditional distributions, which suffices to solve any Markov decision problem in an optimal way.

4.1 Representation of Conditional Markov Random Fields

In this section we discuss the ability of CRBMs to represent conditional Markov random fields, depending on the number of hidden units that they have. The main idea is that each hidden unit of an RBM can be used to model the pure interaction of a group of visible units. This idea appeared in previous work by Younes (1996), in the context of universal approximation.

Definition 13 *Consider a simplicial complex I on $[N]$; that is, a collection of subsets of $[N] = \{1, \dots, N\}$ such that $A \in I$ implies $B \in I$ for all $B \subseteq A$ (in particular $\emptyset \in I$). The random field $\mathcal{E}_I \subseteq \Delta_N$ with interactions I is the set of probability distributions of the form*

$$p(x) = \frac{1}{Z} \exp\Big(\sum_{A \in I} \theta_A \prod_{i \in A} x_i \Big), \quad \text{for all } x = (x_1, \dots, x_N) \in \{0,1\}^N,$$

with normalization $Z = \sum_{x' \in \{0,1\}^N} \exp(\sum_{A \in I} \theta_A \prod_{i \in A} x'_i)$ and parameters $\theta_A \in \mathbb{R}$, $A \in I$.

Given a set S, we will denote the set of all subsets of S by 2^S. We obtain the following result:

Theorem 14 *Let I be a simplicial complex on $[k+n]$ and let $J = 2^{[k]} \cup \{\{k+1\}, \dots, \{k+n\}\}$. If $m \geq |I \setminus J|$, then the model $\mathrm{RBM}^k_{n,m}$ can represent every conditional distribution of $(x_{k+1}, \dots, x_{k+n})$, given $(x_1, \dots, x_k)$, that can be represented by $\mathcal{E}_I \subseteq \Delta_{k+n}$.*

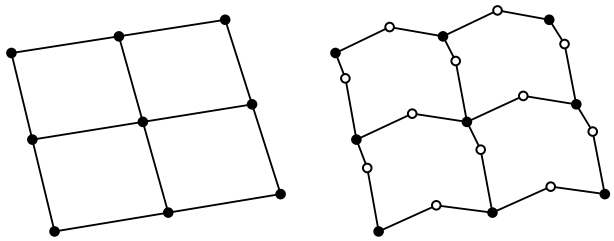

Figure 4: Example of a Markov random field and a corresponding RBM architecture that can represent it. Visible units are depicted in black and hidden units in white.

An interesting special case is when each output distribution can be chosen arbitrarily from a given Markov random field:

Corollary 15 *Let I be a simplicial complex on $[n]$ and for each $x \in \{0,1\}^n$ let p^x be some probability distribution from $\mathcal{E}_I \subseteq \Delta_n$. If $m \geq 2^k(|I| - 1) - |\{A \in I \colon |A| = 1\}|$, then the model $\mathrm{RBM}^k_{n,m}$ can represent the conditional distribution defined by $q(y|x) = p^x(y)$, for all $y \in \{0,1\}^n$, for all $x \in \{0,1\}^k$.*

We note the following direct implication for RBM probability models:

Corollary 16 *Let I be a simplicial complex on $[n]$. If $m \geq |\{A \in I \colon |A| > 1\}|$, then $\mathrm{RBM}_{n,m}$ can represent any probability distribution p from $\mathcal{E}_I$.*

Figure 4 illustrates a Markov random field and an RBM that can represent it.

4.2 Approximation of Conditional Distributions with Restricted Supports

In this section we continue the discussion about the classes of conditional distributions that can be represented by CRBMs, depending on the number of hidden units. Here we focus on a hierarchy of conditional distributions defined by the total number of input-output pairs with positive probability.

Definition 17 *For any k, n, and $0 \leq d \leq 2^k(2^n - 1)$, let $C_{k,n}(d) \subseteq \Delta_{k,n}$ denote the union of all d-dimensional faces of $\Delta_{k,n}$; that is, the set of conditional distributions that have a total of $2^k + d$ or fewer non-zero entries, $C_{k,n}(d) := \{p(\cdot|\cdot) \in \Delta_{k,n} \colon |\{(x,y) \colon p(y|x) > 0\}| \leq 2^k + d\}$.*

Note that $C_{k,n}(2^k(2^n - 1)) = \Delta_{k,n}$. The vertices (zero-dimensional faces) of $\Delta_{k,n}$ are the conditional distributions which assign positive probability to only one output, given each input, and are called *deterministic*. By Carathéodory's theorem, every element of $C_{k,n}(d)$ is a convex combination of $(d+1)$ or fewer deterministic conditional distributions.

The sets $C_{k,n}(d)$ arise naturally in the context of reinforcement learning and partially observable Markov decision processes (POMDPs). Namely, every finite POMDP has an associated effective dimension d, which is the dimension of the set of all state processes that can be generated by stationary stochastic policies. Montúfar et al. (2015) showed that

the policies represented by conditional distributions from the set $C_{k,n}(d)$ are sufficient to generate all the processes that can be generated by $\Delta_{k,n}$. In general, the effective dimension d is relative small, such that $C_{k,n}(d)$ is a much smaller policy search space than $\Delta_{k,n}$.

We have the following result:

Proposition 18 *If $m \geq 2^k + d - 1$, then the model $\mathrm{RBM}^k_{n,m}$ can approximate every element from $C_{k,n}(d)$ arbitrarily well.*

This result shows the intuitive fact that each hidden unit can be used to model the probability of an input-output pair. Since each conditional distribution has 2^k input-output probabilities that are completely determined by the other probabilities (due to normalization), it is interesting to ask whether the amount of hidden units indicated in Proposition 18 is strictly necessary. Further below, Theorem 21 will show that, indeed, hidden units are required for modeling the positions of the positive probability input-output pairs, even if their specific values do not need to be modeled.

We note that certain structures of positive probability input-output pairs can be modeled with fewer hidden units than stated in Proposition 18. An simple example is the following direct generalization of Corollary 8:

Proposition 19 *If d is divisible by 2^k and $m \geq d/2$, then the model $\mathrm{RBM}^k_{n,m}$ can approximate every element from $C_{k,n}(d)$ arbitrarily well, when the set of positive-probability outputs is the same for all inputs.*

In the following we will focus on deterministic conditional distributions. This is a particularly interesting and simple class of conditional distributions with restricted supports. It is well known that any finite Markov decision processes (MDPs) has an optimal policy defined by a stationary deterministic conditional distribution (see Bellman, 1957; Ross, 1983). Furthermore, Ay et al. (2013) showed that it is always possible to define simple two-dimensional manifolds that approximate all deterministic conditional distributions arbitrarily well.

Certain classes of conditional distributions (in particular deterministic conditionals) coming from feedforward networks can be approximated arbitrarily well by CRBMs. We use the following definitions. A *linear threshold unit* with inputs $x \in \{0,1\}^k$ is a function that outputs 1 when $\sum_j V_{ij} x_j + c_i > 0$, and outputs 0 otherwise. A *sigmoid belief unit* with inputs $z \in \{0,1\}^m$ is a stochastic function that outputs 1 with probability $p(y_i = 1|z) = \sigma(\sum_j W_{ij} z_j + b_i)$, where $\sigma(s) = \frac{1}{1+\exp(-s)}$, and outputs 0 with complementary probability.

Theorem 20 *The model $\mathrm{RBM}^k_{n,m}$ can approximate every conditional distribution arbitrarily well, which can be represented by a feedforward network with k input units, a hidden layer of m linear threshold units, and an output layer of n sigmoid belief units. In particular, the model $\mathrm{RBM}^k_{n,m}$ can approximate every deterministic conditional distribution from $\Delta_{k,n}$ arbitrarily well, which can be represented by a feedforward linear threshold network with k input, m hidden, and n output units.*

The representational power of feedforward linear threshold networks has been studied intensively in the literature. For example, Wenzel et al. (2000) showed that a feedforward linear threshold network with $k \geq 1$ input, m hidden, and $n = 1$ output units, can represent the following:

- Any Boolean function $f\colon \{0,1\}^k \to \{0,1\}$, when $m \geq 3 \cdot 2^{k-1-\lfloor \log_2(k+1) \rfloor}$; e.g., when $m \geq \frac{3}{k+2} 2^k$.
- The parity function $f_{\text{parity}}\colon \{0,1\}^k \to \{0,1\};\ x \mapsto \sum_i x_i \mod 2$, when $m \geq k$.
- The indicator function of any union of m linearly separable subsets of $\{0,1\}^k$.

Although CRBMs can approximate this rich class of deterministic conditional distributions arbitrarily well, the next result shows that the number of hidden units required for universal approximation of deterministic conditional distributions is rather large:

Theorem 21 *The model* $\text{RBM}^k_{n,m}$ *can approximate every deterministic conditional distribution from* $\Delta_{k,n}$ *arbitrarily well if* $m \geq \min\left\{2^k - 1, \frac{3n}{k+2} 2^k\right\}$ *and only if* $m \geq 2^{k/2} - \frac{(n+k)^2}{2n}$.

This theorem refines the statement of Proposition 18 in the special case $d = 0$. By this theorem, in order to approximate all deterministic conditional distributions arbitrarily well, a CRBM requires exponentially many hidden units, with respect to the number of input units.

5. Conclusion

This paper gives a theoretical description of the representational capabilities of conditional restricted Boltzmann machines (CRBMs) relating model complexity and model accuracy. CRBMs are based on the well studied restricted Boltzmann machine (RBM) probability models. We proved an extensive series of results that generalize recent theoretical work on the representational power of RBMs in a non-trivial way.

We studied the problem of parameter identifiability. We showed that every CRBM with up to exponentially many hidden units (in the number of input and output units) represent a set of conditional distributions of dimension equal to the number of model parameters. This implies that in all practical cases, CRBMs do not waste parameters, and, generically, only finitely many choices of the interaction weights and biases produce the same conditional distribution.

We addressed the classical problems of universal approximation and approximation quality. Our results show that a CRBM with m hidden units can approximate every conditional distribution of n output units, given k input units, without surpassing a Kullback-Leibler approximation error of the form $n - \log_2(m/2^{k-1} + 1)$ (assuming optimal parameters). Thus this model is a universal approximator whenever $m \geq \frac{1}{2} 2^k (2^n - 1)$. In fact we provided tighter bounds depending on k. For instance, if $k \geq 21$, then the universal approximation property is attained whenever $m \geq \frac{1}{4} 2^k (2^n - 29/30)$. Our proof is based on an upper bound for the complexity of an algorithm that packs Boolean cubes with sequences of non-overlapping stars, for which improvements may be possible. It is worth mentioning that the set of conditional distributions for which the approximation error is maximal may be very small. This is a largely open and difficult problem. We note that our results can be plugged into certain analytic integrals (see Montúfar and Rauh, 2014) to produce upper-bounds for the expectation value of the approximation error when approximating conditional distributions drawn from a product Dirichlet density on the polytope of all conditional distributions.

For future work it would be interesting to extend our (optimal-parameter) considerations by an analysis of the CRBM training complexity and the errors resulting from non-optimal parameter choices.

We also studied specific classes of conditional distributions that can be represented by CRBMs, depending on the number of hidden units. We showed that CRBMs can represent conditional Markov random fields by using each hidden unit to model the interaction of a group of visible variables. Furthermore, we showed that CRBMs can approximate all binary functions with k input bits and n output bits arbitrarily well if $m \geq 2^k - 1$ or $m \geq \frac{3n}{k+2} 2^k$ and only if $m \geq 2^{k/2} - (n+k)^2/2n$. In particular, this implies that there are exponentially many deterministic conditional distributions which can only be approximated arbitrarily well by a CRBM if the number of hidden units is exponential in the number of input units. This aligns with well known examples of functions that cannot be compactly represented by shallow feedforward networks, and reveals some of the intrinsic constraints of CRBM models that may prevent them from grossly over-fitting.

We think that the developed techniques can be used for studying other conditional probability models as well. In particular, for future work it would be interesting to compare the representational power of CRBMs and of combinations of CRBMs with feedforward nets (combined models of this kind include CRBMs with retroactive connections and recurrent temporal RBMs). Also, it would be interesting to apply our techniques to study stacks of CRBMs and other multilayer conditional models. Finally, although our analysis focuses on the case of binary units, the main ideas can be extended to the case of discrete non-binary units.

Acknowledgments

We are grateful to anonymous reviewers for helpful comments. We acknowledge support from the DFG Priority Program Autonomous Learning (DFG-SPP 1527). G. M. and K. G.-Z. would like to thank the Santa Fe Institute for hosting them during the initial work on this article.

Appendix A. Details on the Dimension

Proof of Proposition 3 Each joint distribution of x and y has the form $p(x,y) = p(x)p(y|x)$ and the set Δ_k of all marginals $p(x)$ has dimension $2^k - 1$. The items follow directly from the corresponding statements for the probability model (Cueto et al., 2010). ■

We will need two standard definitions from coding theory:

Definition 22 Let r and k be two natural numbers with $r \leq k$. A radius-r *Hamming ball* in $\{0,1\}^k$ is a set B consisting of a length-k binary vector, together with all other length-k binary vectors that are at most Hamming distance r apart from that vector; that is, $B = \{x \in \{0,1\}^k \colon d_H(x,z) \leq r\}$ for some $z \in \{0,1\}^k$, where $d_H(x,z) := |\{i \in [k] \colon x_i \neq z_i\}|$ denotes the Hamming distance between x and z. Here $[k] := \{1, \ldots, k\}$.

Definition 23 An r-dimensional *cylinder set* in $\{0,1\}^k$ is a set C of length-k binary vectors with arbitrary values in r coordinates and fixed values in the other coordinates; that is, $C = \{x \in \{0,1\}^k \colon x_i = z_i \text{ for all } i \in \Lambda\}$ for some $z \in \{0,1\}^k$ and some $\Lambda \subseteq [k]$ with $k - |\Lambda| = r$.

The geometric intuition is simple: a cylinder set corresponds to the vertices of a face of a unit cube. A radius-1 Hamming ball corresponds to the vertices of a corner of a unit cube. The vectors in a radius-1 Hamming ball are affinely independent. See Figure 5A for an illustration.

Proof of Theorem 4 The proof is based on the ideas developed by Cueto et al. (2010) for studying the RBM probability model. We prove a stronger (more technical) statement than the one given in the theorem: The set $\{0,1\}^{k+n}$ contains m disjoint radius-1 Hamming balls whose union does not contain any set of the form $[x] := \{(x,y) \in \{0,1\}^{k+n} \colon y \in \{0,1\}^n\}$ for $x \in \{0,1\}^k$, and whose complement has full affine rank, as a subset of $\mathbb{R}^{k+n}$.

We consider the Jacobian of $\mathrm{RBM}^k_{n,m}$ for the parameterization given in Definition 1. The dimension of $\mathrm{RBM}^k_{n,m}$ is the maximum rank of the Jacobian over all possible choices of $\theta = (W, V, b, c) \in \mathbb{R}^N$, $N = n + m + (n+k)m$. Let $h_\theta(v) := \operatorname{argmax}_{z \in \{0,1\}^m} p(z|v)$ denote the most likely hidden state of $\mathrm{RBM}_{k+n,m}$ given the visible state $v = (x,y)$, depending on the parameter θ. After a few direct algebraic manipulations, we find that the maximum rank of the Jacobian is bounded from below by the maximum over θ of the dimension of the column-span of the matrix $\mathcal{A}_\theta$ with rows

$$\left((1, x^\top, y^\top), (1, x^\top, y^\top) \otimes h_\theta(x,y)^\top\right), \quad \text{for all } (x,y) \in \{0,1\}^{k+n},$$

modulo vectors whose (x,y)-th entries are independent of y given x. Here $\otimes$ is the Kronecker product, which is defined by $(a_{ij})_{i,j} \otimes (b_{kl})_{k,l} = (a_{ij} b_{kl})_{ik,jl}$. The modulo operation has the effect of disregarding the input distribution $p(x)$ in the joint distribution $p(x,y) = p(x)p(y|x)$ represented by the RBM. For example, from the first block of $\mathcal{A}_\theta$ we can remove the columns that correspond to x, without affecting the mentioned column-span. Summarizing, the maximal column-rank of $\mathcal{A}_\theta$ modulo the vectors whose (x,y)-th entries are independent of y given x is a lower bound for the dimension of $\mathrm{RBM}^k_{n,m}$.

Note that $\mathcal{A}_\theta$ depends on θ in a discrete way: the parameter space $\mathbb{R}^N$ is partitioned in finitely many regions where $\mathcal{A}_\theta$ is constant. The piece-wise linear map thus emerging, with linear pieces represented by the $\mathcal{A}_\theta$, is the tropical CRBM morphism, and its image is the tropical CRBM model.

Each linear region of the tropical morphism corresponds to a function $h_\theta \colon \{0,1\}^{k+n} \to \{0,1\}^m$ taking visible state vectors to the most likely hidden state vectors. Geometrically, such an inference function corresponds to m slicings of the $(k+n)$-dimensional unit hypercube. Namely, every hidden unit divides the visible space $\{0,1\}^{k+n} \subset \mathbb{R}^{k+n}$ in two halfspaces, according to its preferred state.

Each of these m slicings defines a column block of the matrix $\mathcal{A}_\theta$. More precisely,

$$\mathcal{A}_\theta = (A, A_{C_1}, \cdots, A_{C_m}),$$

where A is the matrix with rows $(1, v_1, \ldots, v_{k+n})$ for all $v \in \{0,1\}^{k+n}$, and A_C is the same matrix, with rows multiplied by the indicator function of the set C of points v classified as positive by a linear classifier (slicing).

If we consider only linear classifiers that select rows of A corresponding to disjoint Hamming balls of radius one (that is, such that the C_i are disjoint radius-one Hamming balls), then the rank of $\mathcal{A}_\theta$ is equal to the number of such classifiers times $(n+k+1)$ (which is the rank of each block A_{C_i}), plus the rank of $A_{\{0,1\}^{k+n}\setminus\cup_{i\in[m]}C_i}$ (which is the remainder rank of the first block A). The column-rank modulo functions of x is equal to the rank minus $k+1$ (which is the dimension of the functions of x spanned by columns of A), minus at most the number of cylinder sets $[x] = \{(x,y)\colon y \in \{0,1\}^n\}$ for some $x \in \{0,1\}^k$ that are contained in $\cup_{i\in[m]}C_i$. This completes the proof of the claim.

The bound given in the first item is a consequence of the following observations. Each cylinder set $[x]$ contains 2^n points. If a given cylinder set $[x]$ intersects a radius-1 Hamming ball B but is not contained in it, then it also intersects the radius-2 Hamming sphere around B. Choosing the radius-1 Hamming ball slicings $C_1, \dots, C_m$ to have centers at least Hamming distance 4 apart, we can ensure that their union does not contain any cylinder set $[x]$.

The second item is by the second item of Proposition 3; when the probability model $\mathrm{RBM}_{n+k,m}$ is full dimensional, then $\mathrm{RBM}^k_{n,m}$ is full dimensional. ■

Proof of Corollary 5 For the maximal cardinality of distance-4 binary codes of length l it is known that $A(l,4) \geq 2^r$, where r is the largest integer with $2^r < \frac{2^l}{1+(l-1)+(l-1)(l-2)/2}$ (Gilbert, 1952; Varshamov, 1957), and so $A_2(l,4) \geq 2^{l-\lfloor \log_2(l^2-l+2)\rfloor}$. Furthermore, for the minimal size of radius-one covering codes of length l it is known that $K(l,1) \leq 2^{l-\lfloor \log_2(l+1)\rfloor}$ (Cueto et al., 2010). ■

Appendix B. Details on Universal Approximation

In the following two subsections we address the minimal sufficient and the necessary number of hidden units for universal approximation.

B.1 Sufficient Number of Hidden Units

This subsection contains the proof of Theorem 7 about the minimal size of CRBM universal approximators. The proof is constructive: given any target conditional distribution, it proceeds by adjusting the weights of the hidden units successively until obtaining the desired approximation. The idea of the proof is that each hidden unit can be used to model the probability of an output vector, for several different input vectors. The probability of a given output vector can be adjusted at will by a single hidden unit, jointly for several input vectors, when these input vectors are in general position. This comes at the cost of generating dependent output probabilities for all other inputs in the same affine space. The main difficulty of the proof lies in the construction of sequences of successively conflict-free groups of affinely independent inputs, and in estimating the shortest possible length of such sequences exhausting all possible inputs. The proof is composed of several lemmas and propositions. We start with a few definitions:

Definition 24 Given two probability distributions p and q on a finite set $\mathcal{X}$, the *Hadamard product* or renormalized entry-wise product $p*q$ is the probability distribution on $\mathcal{X}$ defined by $(p*q)(x) = p(x)q(x)/\sum_{x'} p(x')q(x')$ for all $x \in \mathcal{X}$. When building this product, we assume that the supports of p and q are not disjoint, such that the normalization term does not vanish.

The probability distributions that can be represented by RBMs can be described in terms of Hadamard products. Namely, for every probability distribution p that can be represented by $\mathrm{RBM}_{n,m}$, the model $\mathrm{RBM}_{n,m+1}$ with one additional hidden unit can represent precisely the probability distribution of the form $p' = p*q$, where $q = \lambda' r + (1-\lambda')s$ is a mixture, with $\lambda' \in [0,1]$, of two strictly positive product distributions $r(x) = \prod_{i\in[n]} r_i(x_i)$ and $s(x) = \prod_{i\in[n]} s_i(x_i)$. For clarity, the notations are $x = (x_1, \ldots, x_n) \in \{0,1\}^n$, $r, s \in \Delta_n$, and $r_i, s_i \in \Delta_1$ for all $i \in [n] = \{1, \ldots, n\}$. In other words, each additional hidden unit amounts to Hadamard-multiplying the distributions representable by an RBM with the distributions representable as mixtures of product distributions. The same result is obtained by considering only the Hadamard products with mixtures where r is equal to the uniform distribution. In this case, the distributions $p' = p*q$ are of the form $p' = \lambda p + (1-\lambda)p*s$, where s is any strictly positive product distribution and $\lambda = \frac{\lambda'}{\lambda' + 2^n(1-\lambda')\sum_x p(x)s(x)}$ is any weight in $[0,1]$.

Definition 25 A *probability sharing step* is a transformation taking a probability distribution p to $p' = \lambda p + (1-\lambda)p*s$, for some strictly positive product distribution s and some $\lambda \in [0,1]$.

In order to prove Theorem 7, for each $k \in \mathbb{N}$ and $n \in \mathbb{N}$ we want to find an $m_{k,n} \in \mathbb{N}$ such that: for any given strictly positive conditional distribution $q(\cdot|\cdot)$, there exists $p \in \mathrm{RBM}_{n+k,0}$ and $m_{k,n}$ probability sharing steps taking p to a strictly positive joint distribution p' with $p'(\cdot|\cdot) = q(\cdot|\cdot)$. The idea is that the starting distribution is represented by an RBM with no hidden units, and each sharing step is realized by adding a hidden unit to the RBM. In order to obtain these sequences of sharing steps, we will use the following technical lemma:

Lemma 26 *Let B be a radius-1 Hamming ball in $\{0,1\}^k$ and let C be a cylinder subset of $\{0,1\}^k$ containing the center of B. Let $\lambda^x \in (0,1)$ for all $x \in B \cap C$, let $\tilde{y} \in \{0,1\}^n$ and let $\delta_{\tilde{y}}$ denote the Dirac delta on $\{0,1\}^n$ assigning probability one to $\tilde{y}$. Let $p \in \Delta_{k+n}$ be a strictly positive probability distribution with conditionals $p(\cdot|x)$ and let*

$$p'(\cdot|x) := \begin{cases} \lambda^x p(\cdot|x) + (1-\lambda^x)\delta_{\tilde{y}}, & \text{for all } x \in B \cap C \\ p(\cdot|x), & \text{for all } x \in \{0,1\}^k \setminus C \end{cases}.$$

Then, for any $\epsilon > 0$, there is a probability sharing step taking p to a joint distribution p'' with conditionals satisfying $\sum_y |p''(y|x) - p'(y|x)| \leq \epsilon$ for all $x \in (B \cap C) \cup (\{0,1\}^k \setminus C)$.

Proof We define the sharing step $p' = \lambda p + (1-\lambda)p*s$ with a product distribution s supported on $C \times \{\tilde{y}\} \subseteq \{0,1\}^{k+n}$. Note that given any distribution q on C and a radius-1 Hamming ball B whose center is contained in C, there is a product distribution s on C such that $s|_{C\cap B} \propto q|_{C\cap B}$. In other words, the restriction of a product distribution s to a radius-1

Hamming ball B can be made proportional to any non-negative vector of length $|B|$. To see this, recall that a product distribution is a vector with entries $s(x) = \prod_{i\in[k]} s_i(x_i)$, $x = (x_1, \ldots, x_k) \in \{0,1\}^k$. Without loss of generality let B be centered at $(0, \ldots, 0)$; that is, $B = \{x \in \{0,1\}^k \colon \sum_{i\in[k]} x_i \leq 1\}$. The restriction of s to B is given by

$$\begin{aligned} s|_B &= \Big(\prod_{i\in[k]} s_i(0),\ s_1(1) \prod_{i\in[k]\setminus\{1\}} s_i(0),\ s_2(1) \prod_{i\in[k]\setminus\{2\}} s_i(0),\ \ldots,\ s_k(1) \prod_{i\in[k]\setminus\{k\}} s_i(0) \Big) \\ &= \Big(\prod_{i\in[k]} s_i(0),\ \frac{s_1(1)}{s_1(0)} \prod_{i\in[k]} s_i(0),\ \frac{s_2(1)}{s_2(0)} \prod_{i\in[k]} s_i(0),\ \ldots,\ \frac{s_k(1)}{s_k(0)} \prod_{i\in[k]} s_i(0) \Big) \\ &\propto \Big(1,\ \frac{s_1(1)}{s_1(0)},\ \frac{s_2(1)}{s_2(0)},\ \ldots,\ \frac{s_k(1)}{s_k(0)}\Big). \end{aligned}$$

Now, by choosing the factor distributions $s_i = (s_i(0), s_i(1)) \in \Delta_1$ appropriately, the vector $\big(\frac{s_1(1)}{s_1(0)}, \ldots, \frac{s_k(1)}{s_k(0)}\big)$ can be made arbitrary in $\mathbb{R}^k_+$. ■

We have the following two implications of Lemma 26:

Corollary 27 *For any $\epsilon > 0$ and $q(\cdot|x) \in \Delta_n$ for all $x \in B \cap C$, there is an $\epsilon' > 0$ such that, for any strictly positive joint distribution $p \in \Delta_{k+n}$ with conditionals satisfying $\sum_y |p(y|x) - \delta_0(y)| \leq \epsilon'$ for all $x \in B \cap C$, there are $2^n - 1$ sharing steps taking p to a joint distribution p'' with conditionals satisfying $\sum_y |p''(y|x) - p'(y|x)| \leq \epsilon$ for all $x \in (B \cap C) \cup (\{0,1\}^k \setminus C)$, where δ_0 is the Dirac delta on $\{0,1\}^n$ assigning probability one to the vector of zeros and*

$$p'(\cdot|x) := \begin{cases} q(\cdot|x), & \text{for all } x \in B \cap C \\ p(\cdot|x), & \text{for all } x \in \{0,1\}^k \setminus C \end{cases}.$$

Proof Consider any $x \in B \cap C$. We will show that the probability distribution $q(\cdot|x) \in \Delta_n$ can be written as the transformation of a Dirac delta by $2^n - 1$ sharing steps. Then the claim follows from Lemma 26. Let $\sigma \colon \{0,1\}^n \to \{0, \ldots, 2^n - 1\}$ be an enumeration of $\{0,1\}^n$. Let $p^{(0)}(y|x) = \delta_{\sigma^{-1}(0)}(y)$ be the starting distribution (the Dirac delta concentrated at the state $\tilde{y} \in \{0,1\}^n$ with $\sigma(\tilde{y}) = 0$) and let the t-th sharing step be defined by $p^{(t)}(y) = \lambda^x_{\sigma^{-1}(t)} p^{(t-1)}(y|x) + (1 - \lambda^x_{\sigma^{-1}(t)}) \delta_{\sigma^{-1}(t)}(y)$, for some weight $\lambda^x_{\sigma^{-1}(t)} \in [0,1]$. After $2^n - 1$ sharing steps, we obtain the distribution

$$p^{(2^n-1)}(y|x) = \sum_{\tilde{y}} \Big(\prod_{\tilde{y}' \colon \sigma(\tilde{y}') > \sigma(\tilde{y})} \lambda^x_{\tilde{y}'} \Big) (1 - \lambda^x_{\tilde{y}}) \delta_{\tilde{y}}(y), \quad \text{for all } y \in \{0,1\}^n,$$

whereby $\lambda^x_{\tilde{y}} := 0$ for $\sigma(\tilde{y}) = 0$. This distribution is equal to $q(\cdot|x)$ for the following choice of weights:

$$\lambda^x_{\tilde{y}} := 1 - \frac{q(\tilde{y}|x)}{1 - \sum_{\tilde{y}' \colon \sigma(\tilde{y}') > \sigma(\tilde{y})} q(\tilde{y}'|x)}, \quad \text{for all } \tilde{y} \in \{0,1\}^n.$$

It is easy to verify that these weights satisfy the condition $\lambda^x_{\tilde{y}} \in [0,1]$ for all $\tilde{y} \in \{0,1\}^n$, and $\lambda^x_{\tilde{y}} = 0$ for that $\tilde{y}$ with $\sigma(\tilde{y}) = 0$, independently of the specific choice of σ. ■

Note that this corollary does not make any statement about the rows $p''(\cdot|x)$ with $x \in C \setminus B$. When transforming the $(B \cap C)$-rows of p according to Lemma 26, the $(C \setminus B)$-rows get transformed as well, in a non-trivial dependent way. Fortunately, there is a sharing step that allows us to "reset" exactly certain rows to a desired point measure, without introducing new non-trivial dependencies:

Corollary 28 *For any $\epsilon > 0$, any cylinder set $C \subseteq \{0,1\}^k$, and any $\tilde{y} \in \{0,1\}^n$, any strictly positive joint distribution p can be transformed by a probability sharing step to a joint distribution p'' with conditionals satisfying $\sum_y |p''(y|x) - p'(y|x)| \leq \epsilon$ for all $x \in \{0,1\}^k$, where*

$$p'(\cdot|x) := \begin{cases} \delta_{\tilde{y}}, & \text{for all } x \in C \\ p(\cdot|x), & \text{for all } x \in \{0,1\}^k \setminus C \end{cases}.$$

Proof The sharing step can be defined as $p'' = \lambda p + (1-\lambda) p * s$ with s close to the uniform distribution on $C \times \{\tilde{y}\}$ and λ close to 0 (close enough depending on ϵ). ■

We will refer to a sharing step as described in Corollary 28 as a *reset* of the C-rows of p. Furthermore, we will denote by *star* the intersection of a radius-1 Hamming ball and a cylinder set containing the center of the ball. See Figure 5A.

With all the observations made above, we can construct an algorithm that generates an arbitrarily accurate approximation of any given conditional distribution by applying a sequence of sharing steps to any given strictly positive joint distribution. The details are given in Algorithm 1. The algorithm performs sequential sharing steps on a strictly positive joint distribution $p \in \Delta_{k+n}$ until the resulting distribution p' has a conditional distribution $p'(\cdot|\cdot)$ satisfying $\sum_y |p'(y|x) - q(y|x)| \leq \epsilon$ for all x.

In order to obtain a bound on the number m of hidden units for which $\mathrm{RBM}^k_{n,m}$ can approximate a given target conditional distribution arbitrarily well, we just need to evaluate the number of sharing steps run by Algorithm 1. For this purpose, we investigate the combinatorics of sharing step sequences and evaluate their worst case lengths. We can choose as starting distribution some $p \in \mathrm{RBM}_{n+k,0}$ with conditionals satisfying $\sum_y |p(y|x) - \delta_0(y)| \leq \epsilon'$ for all $x \in \{0,1\}^k$, for some $\epsilon' > 0$ small enough depending on the target conditional $q(\cdot|\cdot)$ and the targeted approximation accuracy ϵ.

Definition 29 A sequence of stars $B^1, \ldots, B^l$ packing $\{0,1\}^k$ with the property that the smallest cylinder set containing any of the stars in the sequence does not intersect any previous star in the sequence is called a *star packing sequence* for $\{0,1\}^k$.

The number of sharing steps run by Algorithm 1 is bounded from above by $(2^n - 1)$ times the length of a star packing sequence for the set of inputs $\{0,1\}^k$. Note that the choices of stars and the lengths of the possible star packing sequences are not unique. Figure 5B gives an example showing that starting a sequence with large stars is not necessarily the best strategy to produce a short sequence. The next lemma states that there is a class of star packing sequences of a certain length, depending on the size of the input space. Thereby, this lemma upper-bounds the worst case complexity of Algorithm 1.

Algorithm 1 Algorithmic illustration of the proof of Theorem 7.

Input: Strictly positive joint distribution p, target conditional distribution $q(\cdot|\cdot)$, and $\epsilon > 0$
Output: Transformation p' of the input p with $\sum_y |p'(y|x) - q(y|x)| \leq \epsilon$ for all x

Initialize $\mathcal{B} \leftarrow \emptyset$ {Here $\mathcal{B} \subseteq \{0,1\}^k$ denotes the set of inputs x that have been readily processed in the current iteration}
while $\mathcal{B} \not\supseteq \{0,1\}^k$ **do**
  Choose (disjoint) cylinder sets $C^1, \ldots, C^K$ packing $\{0,1\}^k \setminus \mathcal{B}$
  If needed, perform at most K sharing steps resetting the C^i rows of p for all $i \in [K]$, taking $p(\cdot|x)$ close to δ_0 for all $x \in C^i$ for all $i \in [K]$ and leaving all other rows close to their current values, according to Corollary 28
  for each $i \in [K]$ **do**
    Perform at most $2^n - 1$ sharing steps taking $p(\cdot|x)$ close to $q(\cdot|x)$ for all $x \in B^i$, where B^i is some star contained in C^i, and leaving the $(\{0,1\}^k \setminus C^i)$-rows close to their current values, according to Corollary 27
  end for
  $\mathcal{B} \leftarrow \mathcal{B} \cup (\cup_{i\in[K]} B^i)$
end while

Lemma 30 *Let $r \in \mathbb{N}$, $S(r) := 1 + 2 + \cdots + r$, $k \geq S(r)$, $f_i(z) := 2^{S(i-1)} + (2^i - (i+1))z$, and $F(r) := f_r(f_{r-1}(\cdots f_2(f_1)))$. There is a star packing sequence for $\{0,1\}^k$ of length $2^{k-S(r)}F(r)$. Furthermore, for this sequence, Algorithm 1 requires at most $R(r) := \prod_{i=2}^r (2^i - (i+1))$ resets.*

Proof The star packing sequence is constructed by the following procedure. In each step, we define a set of cylinder sets packing all sites of $\{0,1\}^k$ that have not been covered by stars so far, and include a sub-star of each of these cylinder sets in the sequence.

- As an initialization step, we split $\{0,1\}^k$ into $2^{k-S(r)}$ $S(r)$-dimensional cylinder sets, denoted $D^{(j_1)}$, $j_1 \in \{1, \ldots, 2^{k-S(r)}\}$.

- In the first step, for each j_1, the $S(r)$-dimensional cylinder set $D^{(j_1)}$ is packed by $2^{S(r-1)}$ r-dimensional cylinder sets $C^{(j_1),i}$, $i \in \{1, \ldots, 2^{S(r-1)}\}$. For each i, we define the star $B^{(j_1),i}$ as the radius-1 Hamming ball within $C^{(j_1),i}$ centered at the smallest element of $C^{(j_1),i}$ (with respect to the lexicographic order of $\{0,1\}^k$), and include it in the sequence.

- At this point, the sites in $D^{(j_1)}$ that have not yet been covered by stars is $D^{(j_1)} \setminus (\cup_i B^{(j_1),i})$. This set is split into $2^r - (r+1)$ $S(r-1)$-dimensional cylinder sets, which we denote by $D^{(j_1,j_2)}$, $j_2 \in \{1, \ldots, 2^r - (r+1)\}$.

- Note that $\cup_{j_1} D^{(j_1,j_2)}$ is a cylinder set, and hence, for each j_2, the $(\cup_{j_1} D^{(j_1,j_2)})$-rows of a conditional distribution being processed by Algorithm 1 can be jointly reset by one single sharing step to achieve $p'(\cdot|x) \approx \delta_0$ for all $x \in \cup_{j_1} D^{(j_1,j_2)}$.

- In the second step, for each j_2, the cylinder set $D^{(j_1,j_2)}$ is packed by $2^{S(r-2)}$ $(r-1)$-dimensional cylinder sets $C^{(j_1,j_2),i}$, $i \in \{1, \ldots, 2^{S(r-2)}\}$, and the corresponding stars are included in the sequence.

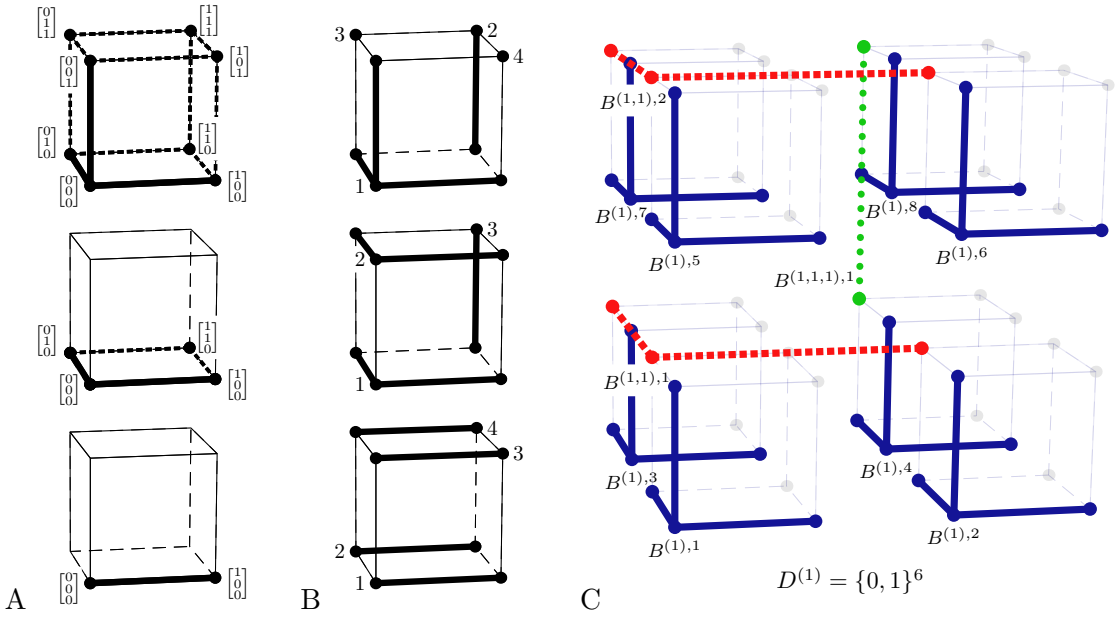

Figure 5: A) Examples of radius-1 Hamming balls in cylinder sets of dimension 3, 2, and 1. The cylinder sets are shown as bold vertices connected by dashed edges, and the nested Hamming balls (stars) as bold vertices connected by solid edges. B) Three examples of star packing sequences for $\{0,1\}^3$. C) Illustration of the star packing sequence constructed in Lemma 30 for $\{0,1\}^6$.

- The procedure is iterated until the r-th step. In this step, each $D^{(j_1,\dots,j_r)}$ is a 1-dimensional cylinder set and is packed by a single 1-dimensional cylinder set $C^{(j_1,\dots,j_r),1} = B^{(j_1,\dots,j_r),1}$. Hence, at this point, all of $\{0,1\}^k$ has been exhausted and the procedure terminates.

Summarizing, the procedure is initialized by creating the branches $D^{(j_1)}$, $j_1 \in [2^{k-S(r)}]$. In the first step, each branch $D^{(j_1)}$ produces $2^{S(r-1)}$ stars and splits into the branches $D^{(j_1,j_2)}$, $j_2 \in [2^r - (r+1)]$. More generally, in the i-th step, each branch $D^{(j_1,\dots,j_i)}$ produces $2^{S(r-i)}$ stars, and splits into the branches $D^{(j_1,\dots,j_i,j_{i+1})}$, $j_{i+1} \in [2^{r-(i-1)} - (r+1-(i-1))]$.

The total number of stars $D^{(j_1,\dots,j_r)}$ is given precisely by $2^{k-S(r)}$ times the value of the iterative function $F(r) = f_r(f_{r-1}(\cdots f_2(f_1)))$, whereby $f_1 = 1$. The total number of resets is given by the number of branches created from the first step on, which is precisely $R(r) = \prod_{i\in[r]}(2^i - (i+1))$.

Figure 5C offers an illustration of these star packing sequences. It shows the case $k = S(3) = 6$. In this case there is only one initial branch $D^{(1)} = \{0,1\}^6$. The stars $B^{(1),i}$, $i \in [2^{S(2)}] = [8]$ are shown in solid blue, $B^{(1,1),i}$, $i \in [2^{S(1)}] = [2]$ in dashed red, and $B^{(1,1,1),1}$ in dotted green. For clarity, only these stars are highlighted. The stars $B^{(1,j_2),i}$ and $B^{(1,j_2,1),1}$ resulting from split branches are similar to those highlighted. ■

With this, we obtain the general bound of the theorem:

	$m_{n,k}^{(r)} =$					
r	2^k	$2^{-S(r)}$	$F(r)$	(2^n-1)	$+$	$R(r)$
1	2^k	2^{-1}	1	(2^n-1)	$+$	0
2	2^k	2^{-3}	3	(2^n-1)	$+$	1
3	2^k	2^{-6}	20	(2^n-1)	$+$	4
4	2^k	2^{-10}	284	(2^n-1)	$+$	44
5	2^k	2^{-15}	8408	(2^n-1)	$+$	1144
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
> 17	2^k	0.2263		(2^n-1)	$+$	$2^{S(r)}0.0269$

Table 1: Numerical evaluation of the bounds from Proposition 31. Each row evaluates the universal approximation bound $m_{n,k}^{(r)}$ for a value of r.

Proposition 31 (Theorem 7, general bound) *Let $k \geq S(r)$. The model $\mathrm{RBM}_{n,m}^k$ can approximate every conditional distribution from $\Delta_{k,n}$ arbitrarily well whenever $m \geq m_{k,n}^{(r)}$, where $m_{k,n}^{(r)} := 2^{k-S(r)}F(r)(2^n-1) + R(r)$.*

Proof This is in view of the complexity of Algorithm 1 for the sequence described in Lemma 30. ■

In order to make the universal approximation bound more comprehensible, in Table 1 we evaluated the sequence $m_{n,k}^{(r)}$ for $r = 1, 2, 3 \ldots$ and $k \geq S(r)$. Furthermore, the next proposition gives an explicit expression for the coefficients $2^{-S(r)}F(r)$ and $R(r)$ appearing in the bound. This yields the second part of Theorem 7. In general, the bound $m_{n,k}^{(r)}$ decreases with increasing r, except possibly for a few values of k when n is small. For a pair (k, n), any $m_{n,k}^{(r)}$ with $k \geq S(r)$ is a sufficient number of hidden units for obtaining a universal approximator.

Proposition 32 (Theorem 7, explicit bounds) *The function $K(r) := 2^{-S(r)}F(r)$ is bounded from below and above as $K(6)\prod_{i=7}^{r}\left(1-\frac{i-3}{2^i}\right) \leq K(r) \leq K(6)\prod_{i=7}^{r}\left(1-\frac{i-4}{2^i}\right)$ for all $r \geq 6$. Furthermore, $K(6) \approx 0.2442$ and $K(\infty) \approx 0.2263$. Moreover, $R(r) := \prod_{i=2}^{r}(2^i-(i+1)) = 2^{S(r)}P(r)$, where $P(r) := \frac{1}{2}\prod_{i=2}^{r}(1-\frac{(i+1)}{2^i})$, and $P(\infty) \approx 0.0269$.*

Proof From the definition of $S(r)$ and $F(r)$, we obtain that

$$K(r) = 2^{-r} + K(r-1)(1-2^{-r}(r+1)). \tag{1}$$

Note that $K(1) = \frac{1}{2}$, and that $K(r)$ decreases monotonically.

Now, note that if $K(r-1) \leq \frac{1}{c}$, then the left hand side of Equation (1) is bounded from below as $K(r) \geq K(r-1)(1-2^{-r}(r+1-c))$. For a given c, let r^c be the first r for which $K(r-1) \leq \frac{1}{c}$, assuming that such an r exists. Then

$$K(r) \geq K(r^c-1)\prod_{i=r^c}^{r}\left(1-\frac{i+1-c}{2^i}\right), \quad \text{for all } r \geq r^c. \tag{2}$$

Similarly, if $K(r) > \frac{1}{d}$ for all $r \geq r^b$, then

$$K(r) \leq K(r^b - 1) \prod_{i=r^b}^{r} \left(1 - \frac{i+1-b}{2^i}\right), \quad \text{for any } r \geq r^b.$$

Direct computations show that $K(6) \approx 0.2445 \leq \frac{1}{4}$. On the other hand, using the computational engine Wolfram—Alpha(access June 01, 2014) we obtain $\prod_{i=0}^{\infty}\left(1 - \frac{i-3}{2^i}\right) \approx 7.7413$. Plugging both terms into Equation (2) yields that $K(r)$ is always bounded from below by 0.2259.

Since $K(r)$ is never smaller than or equal to $\frac{1}{5}$, we obtain $K(r) \leq K(r'-1)\prod_{i=r'}^{r}\left(1 - \frac{i-4}{2^i}\right)$, for any r' and $r \geq r'$. Using $r' = 7$, the right hand side evaluates in the limit of large r to approximately 0.2293.

Numerical evaluation of $K(r)$ from Equation (1) for r up to one million (using Matlab R2013b) indicates that, indeed, $K(r)$ tends to approximately 0.2263 for large r. ■

We close this subsection with the remark that the proof strategy can be used not only to study universal approximation, but also approximability of selected classes of conditional distributions:

Remark 33 If we only want to model a restricted class of conditional distributions, then adapting Algorithm 1 to these restrictions may yield tighter bounds for the number of hidden units that suffices to represent these restricted conditionals. For example:

If we only want to model the target conditionals $q(\cdot|x)$ for the inputs x from a subset $\mathcal{S} \subseteq \{0,1\}^k$ and do not care about $q(\cdot|x)$ for $x \notin \mathcal{S}$, then in the algorithm we just need to replace $\{0,1\}^k$ by $\mathcal{S}$. In this case, a cylinder set packing of $\mathcal{S}\setminus\mathcal{B}$ is understood as a collection of disjoint cylinder sets $C^1, \ldots, C^K \subseteq \{0,1\}^k$ with $\cup_{i\in[K]} C^i \supseteq \mathcal{S}\setminus\mathcal{B}$ and $(\cup_{i\in[K]} C^i) \cap \mathcal{B} = \emptyset$.

Furthermore, if for some cylinder set C^i and a corresponding star $B^i \subseteq C^i$ the conditionals $q(\cdot|x)$ with $x \in B^i$ have a common support set $T \subseteq \{0,1\}^n$, then the C^i-rows of p can be reset to a distribution δ_y with $y \in T$, and only $|T| - 1$ sharing steps are needed to transform p to a distribution whose conditionals approximate $q(\cdot|x)$ for all $x \in B^i$ to any desired accuracy. In particular, for the class of target conditional distributions with $\operatorname{supp} q(\cdot|x) = T$ for all x, the term $2^n - 1$ in the complexity bound of Algorithm 1 is replaced by $|T| - 1$.

B.2 Necessary Number of Hidden Units

Proposition 9 follows from simple parameter counting arguments. In order to make this rigorous, first we make the observation that universal approximation of (conditional) probability distributions by Boltzmann machines or any other models based on exponential families, with or without hidden variables, requires the number of model parameters to be as large as the dimension of the set being approximated. We denote by $\Delta_{\mathcal{X},\mathcal{Y}}$ the set of conditionals with inputs form a finite set $\mathcal{X}$ and outputs from a finite set $\mathcal{Y}$. Accordingly, we denote by $\Delta_{\mathcal{Y}}$ the set of probability distributions on $\mathcal{Y}$.

Lemma 34 *Let $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ be some finite sets. Let $\mathcal{M} \subseteq \Delta_{\mathcal{X},\mathcal{Y}}$ be defined as the set of conditionals of the marginal $\mathcal{M}' \subseteq \Delta_{\mathcal{X}\times\mathcal{Y}}$ of an exponential family $\mathcal{E} \subseteq \Delta_{\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}}$. If*

$\mathcal{M}$ is a universal approximator of conditionals from $\Delta_{\mathcal{X},\mathcal{Y}}$, then $\dim(\mathcal{E}) \geq \dim(\Delta_{\mathcal{X},\mathcal{Y}}) = |\mathcal{X}|(|\mathcal{Y}|-1)$.

The intuition of this lemma is that, for models defined by marginals of exponential families, the set of conditionals that can be approximated arbitrarily well is essentially equal to the set of conditionals that can be represented exactly, implying that there are no low-dimensional universal approximators of this type.

Proof of Lemma 34 We consider first the case of probability distributions; that is, the case with $|\mathcal{X}| = 1$ and $\mathcal{X} \times \mathcal{Y} \cong \mathcal{Y}$. Let $\mathcal{M}$ be the image of the exponential family $\mathcal{E}$ by a differentiable map f (for example, the marginal map). The closure $\overline{\mathcal{E}}$, which consists of all distributions that can be approximated arbitrarily well by $\mathcal{E}$, is a compact set. Since f is continuous, the image of $\overline{\mathcal{E}}$ is also compact, and $\overline{\mathcal{M}} = \overline{f(\mathcal{E})} = f(\overline{\mathcal{E}})$. The model $\mathcal{M}$ is a universal approximator if and only if $\overline{\mathcal{M}} = \Delta_{\mathcal{Y}}$. The set $\overline{\mathcal{E}}$ is a finite union of exponential families; one exponential family $\mathcal{E}_F$ for each possible support set F of distributions from $\overline{\mathcal{E}}$. When $\dim(\mathcal{E}) < \dim(\Delta_{\mathcal{Y}})$, each point of each $\mathcal{E}_F$ is a critical point of f (the Jacobian is not surjective at that point). By Sard's theorem, each $\mathcal{E}_F$ is mapped by f to a set of measure zero in $\Delta_{\mathcal{Y}}$. Hence the finite union $\cup_F f(\mathcal{E}_F) = f(\cup_F \mathcal{E}_F) = f(\overline{\mathcal{E}}) = \overline{\mathcal{M}}$ has measure zero in $\Delta_{\mathcal{Y}}$.

For the general case, with $|\mathcal{X}| \geq 1$, note that $\mathcal{M} \subseteq \Delta_{\mathcal{X},\mathcal{Y}}$ is a universal approximator if and only if the joint model $\Delta_{\mathcal{X}}\mathcal{M} = \{p(x)q(y|x)\colon p \in \Delta_{\mathcal{X}}, q \in \mathcal{M}\} \subseteq \Delta_{\mathcal{X}\times\mathcal{Y}}$ is a universal approximator. The latter is the marginal of the exponential family $\Delta_{\mathcal{X}} * \mathcal{E} = \{p * q\colon p \in \Delta_{\mathcal{X}}, q \in \mathcal{E}\} \subseteq \Delta_{\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}}$. Hence the claim follows from the first part. ■

Proof of Proposition 9 If $\mathrm{RBM}^k_{n,m}$ is a universal approximator of conditionals from $\Delta_{k,n}$, then the model consisting of all probability distributions of the form $p(x,y) = \frac{1}{Z}\sum_z \exp(z^\top W y + z^\top V x + b^\top y + c^\top z + f(x))$ is a universal approximator of probability distributions from Δ_{k+n}. The latter is the marginal of an exponential family of dimension $mn + mk + n + m + 2^k - 1$. Thus, by Lemma 34, $m \geq \frac{2^{k+n}-2^k-n}{(n+k+1)}$. ■

Appendix C. Details on the Maximal Approximation Errors

Proof of Proposition 10 We have that $D_{\mathrm{RBM}^k_{n,m}} \leq \max_{p\in\Delta_{k+n}\colon p_X = u_X} D(p\|\,\mathrm{RBM}_{n+k,m})$. The right hand side is bounded by n, since the RBM model contains the uniform distribution. It is also bounded by the maximal divergence $D_{\mathrm{RBM}_{n+k,m}} \leq (n+k) - \lfloor \log_2(m+1)\rfloor - \frac{m+1}{2^{\lfloor \log_2(m+1)\rfloor}}$ (Montúfar et al., 2013). ■

In order to prove Theorem 11, we will upper bound the approximation errors of CRBMs by the approximation errors of submodels of CRBMs. First, we note the following:

Lemma 35 *The maximal divergence of a conditional model that is a Cartesian product of a probability model is bounded from above by the maximal divergence of that probability model: if $\mathcal{M} = \times_{x\in\{0,1\}^k}\mathcal{N} \subseteq \Delta_{k,n}$ for some $\mathcal{N} \subseteq \Delta_n$, then $D_{\mathcal{M}} \leq D_{\mathcal{N}}$.*

Proof For any $p \in \Delta_{k,n}$, we have

$$\begin{aligned} D(p\|\mathcal{M}) &= \inf_{q\in\mathcal{M}} \frac{1}{2^k}\sum_x D(p(\cdot|x)\|q(\cdot|x)) \\ &= \frac{1}{2^k}\sum_x \inf_{q(\cdot|x)\in\mathcal{N}} D(p(\cdot|x)\|q(\cdot|x)) \\ &\leq \frac{1}{2^k}\sum_x D_{\mathcal{N}} = D_{\mathcal{N}}. \end{aligned}$$

This proves the claim. ■

Definition 36 *Given a partition $\mathcal{Z} = \{\mathcal{Y}_1, \ldots, \mathcal{Y}_L\}$ of $\{0,1\}^n$, the* partition model $\mathcal{P}_{\mathcal{Z}} \subseteq \Delta_n$ *is the set of all probability distributions on $\{0,1\}^n$ with constant value on each partition block.*

The set $\{0,1\}^l$, $l \leq n$ naturally defines a partition of $\{0,1\}^n$ into cylinder sets $\{y \in \{0,1\}^n \colon y_{[l]} = z\}$ for all $z \in \{0,1\}^l$. The divergence from $\mathcal{P}_{\mathcal{Z}}$ is bounded from above by $D_{\mathcal{P}_{\mathcal{Z}}} \leq l - n$. Now, the model $\mathrm{RBM}^k_{n,m}$ can approximate certain products of partition models arbitrarily well:

Proposition 37 *Let $\mathcal{Z} = \{0,1\}^l$ with $l \leq n$. Let r be any integer with $k \geq S(r)$. The model $\mathrm{RBM}^k_{n,m}$ can approximate any conditional distribution from the product of partition models $\mathcal{P}^k_{\mathcal{Z}} := \mathcal{P}_{\mathcal{Z}} \times \cdots \times \mathcal{P}_{\mathcal{Z}}$ arbitrarily well whenever $m \geq 2^{k-S(r)}F(r)(|\mathcal{Z}| - 1) + R(r)$.*

Proof This is analogous to the proof of Proposition 19, with a few differences. Each element z of $\mathcal{Z}$ corresponds to a cylinder set $\{y \in \{0,1\}^n \colon y_{[l]} = z\}$ and the collection of cylinder sets for all $z \in \mathcal{Z}$ is a partition of $\{0,1\}^n$. Now we can run Algorithm 1 in a slightly different way, with sharing steps defined by $p' = \lambda p + (1-\lambda)u_z$, where u_z is the uniform distribution on the cylinder set corresponding to z. ■

Proof of Theorem 11 This follows directly from Lemma 35 and Proposition 37. ■

Appendix D. Details on the Representation of Conditional Distributions from Markov Random Fields

The proof of Theorem 14 is based on ideas from Younes (1996), who discussed the universal approximation property of Boltzmann machines. We will use the following:

Lemma 38 (Younes 1996, Lemma 1) *Let ϱ be a real number. Consider a fixed integer N and binary variables $x_1, \ldots, x_N$. There are real numbers w and b such that:*

- *If $\varrho \geq 0$,* $\log\left(1 + \exp(w(x_1 + \cdots + x_N) + b)\right) = \varrho \prod_i x_i + Q(x_1, \ldots, x_N)$.
- *If $\varrho \leq 0$,* $\log\left(1 + \exp(w(x_1 + \cdots + x_{N-1} - x_N) + b)\right) = \varrho \prod_i x_i + Q(x_1, \ldots, x_N)$.

Where Q is in each case a polynomial of degree less than $N-1$ in $x_1, \dots, x_N$.

The following is a generalization of another result from the same work:

Lemma 39 *Let I and J be two simplicial complexes on $[n]$ with $J \subseteq I$. If p is any distribution from $\mathcal{E}_I$ and $m \geq |\{A \in I \setminus J \colon |A| > 1\}|$, then there is a distribution $p' \in \mathcal{E}_J$, such that $p * p'$ is contained in* $\mathrm{RBM}_{n,m}$.

Proof The proof follows closely the arguments presented by Younes (1996, Lemma 2). Let $K = \{A \in I \setminus J \colon |A| > 1\}$. Consider an RBM with n visible units and $m = |K|$ hidden units. Consider a joint distribution $q(x,u) = \frac{1}{Z}\exp(H(x,u))$ of the fully observable RBM, defined as follows. We label the hidden units by subsets $A \in K$. For each $A \in K$, let $s(A)$ denote the largest element of A, and let

$$H(x,u) = \sum_{A \in K} u_A \left(w_A S_A^{\epsilon_A}(x_A) + b_A \right) + \sum_{s \in [n]} b_s x_s,$$

where

$$S_A^{\epsilon_A}(x_A) = \Big(\sum_{s \in A, s < s(A)} x_s \Big) + \epsilon_A x_{s(A)},$$

for some $\epsilon_A \in \{-1, +1\}$, $w_A, b_A, b_s \in \mathbb{R}$ that we will specify further below.

Denote the log probabilities of $p(x)$ and $p'(x)$ by

$$E(x) = \sum_{A \in I} \theta_A \prod_{i \in A} x_i \quad \text{and} \quad E'(x) = \sum_{A \in J} \vartheta_A \prod_{i \in A} x_i.$$

We obtain the desired equality $(p * p')(x) = \sum_u q(x,u)$ when

$$E(x) = \log \left(\sum_u \exp(H(x,u)) \right) - \sum_{A \in J} \vartheta_A \prod_{i \in A} x_i, \tag{3}$$

for some choice of ϑ_A, for $A \in J$, some choice of ϵ_A, w_A, b_A, for $A \in K$, and some choice of b_s, for $s \in [n]$. We have

$$\begin{aligned}
\log \left(\sum_u \exp(H(x,u)) \right) &= \log \left(\sum_u \exp \Big(\sum_A u_A (w_A S_A^{\epsilon_A}(x_A) + b_A) + \sum_{s \in [n]} b_s x_s \Big) \right) \\
&= \log \left(\Big(\sum_u \prod_A \exp(u_A (w_A S_A^{\epsilon_A}(x_A) + b_A)) \Big) \exp \Big(\sum_{s \in [n]} b_s x_s \Big) \right) \\
&= \log \left(\Big(\prod_A \sum_{u_A} \exp(u_A (w_A S_A^{\epsilon_A}(x_A) + b_A)) \Big) \exp \Big(\sum_{s \in [n]} b_s x_s \Big) \right) \\
&= \sum_A \log(1 + \exp(w_A S_A^{\epsilon_A}(x_A) + b_A)) + \sum_{s \in [n]} b_s x_s.
\end{aligned}$$

The terms $\phi_A^{\epsilon_A}(x_A) := \log\big(1+\exp(w_A S_A^{\epsilon_A}(x_A)+b_A)\big)$ are of the same form as the functions from Lemma 38. To solve Equation (3), we first apply Lemma 38 on $\phi_A^{\epsilon_A}$ to cancel the terms $\theta_A \prod_{i\in A} x_i$ of $E(x)$ for which A is a maximal element of $I\setminus J$ of cardinality more than one. This involves choosing appropriate $\epsilon_A \in \{-1,+1\}$, w_A and b_A, for the corresponding A. The remaining polynomial consists of terms with strictly smaller monomials. We apply lemma 38 repeatedly on this polynomial, until only monomials with $A \in J$ or $|A|=1$ remain. These terms are canceled with $\vartheta_A \prod_{i\in A} x_i$, $A\in J$, or with $b_s x_s$, $s\in[n]$. ■

Proof of Theorem 14 By Lemma 39, there is a $p' \in \mathcal{E}_J$, $J = 2^{[k]}$, such that $p * p'$ is in $\mathrm{RBM}_{k+n,m}$. Now, the conditionals distribution $(p*p')(y|x)$ of the last n units, given the first k units, are independent of p', since this is independent of y. ■

Proof of Corollary 15 The statement follows from Theorem 14, considering the simplicial complex $I = 2^{[k]} \times J$ and a joint probability distribution $p \in \mathcal{E}_I \subseteq \Delta_{k+n}$ with the desired conditionals $p(\cdot|x) = p^x$. ■

Appendix E. Details on the Approximation of Conditional Distributions with Restricted Supports

Proof of Proposition 18 This follows from the fact that $\mathrm{RBM}_{n+k,m}$ can approximate any probability distribution with support of cardinality $m+1$ arbitrarily well (Montúfar and Ay, 2011). ■

Proof of Proposition 19 This is analogous to the proof of Proposition 31. The complexity of Algorithm 1 as evaluated there does not depend on the specific structure of the support sets, but only on their cardinality, as long as they are the same for all x. ■

The following lemma states that a CRBM can compute all deterministic conditionals that can be computed by a feedforward linear threshold network with the same number of hidden units. Recall that the Heaviside step function, here denoted hs, maps a real number a to 0 if $a<0$, to $1/2$ if $a=0$, and to 1 if $a>0$. A linear threshold function with N input bits and M output bits is just a function of the form $\{0,1\}^N \to \{0,1\}^M$; $y \mapsto \mathrm{hs}(Wy+b)$ with a generic choice of $W\in\mathbb{R}^{M\times N}$ and $b\in\mathbb{R}^M$.

Lemma 40 *Consider a function $f\colon \{0,1\}^k \to \{0,1\}^n$. The model $\mathrm{RBM}_{n,m}^k$ can approximate the deterministic policy $p(y|x)=\delta_{f(x)}(y)$ arbitrarily well, whenever this can be represented by a feedforward linear threshold network with m hidden units; that is, when*

$$f(x) = \mathrm{hs}(W^\top(\mathrm{hs}(Vx+c))+b), \quad \textit{for all } x\in\{0,1\}^k,$$

for some generic choice of W,V,b,c.

Proof Consider the conditional distribution $p(\cdot|x)$. This is the visible marginal of $p(y,z|x) = \frac{1}{Z}\exp((Vx+c)^\top z + b^\top y + z^\top W y)$. Consider weights α and β, with α large enough, such that

$\operatorname{argmax}_z(\alpha Vx+\alpha c)^\top z = \operatorname{argmax}_z(\alpha Vx+\alpha c)^\top z + (\beta W^\top z+\beta b)^\top y$ for all $y \in \{0,1\}^n$. Note that for generic choices of V and c, the set $\operatorname{argmax}_z(\alpha V+\alpha c)^\top z$ consists of a single point $z^* = \operatorname{hs}(Vx+c)$. We have $\operatorname{argmax}_{(y,z)}(\alpha Vx+\alpha c)^\top z+(\beta W^\top z+\beta b)^\top y = (z^*, \operatorname{argmax}_y(\beta W^\top z^* + \beta b)^\top y)$. Here, again, for generic choices of V and b, the set $\operatorname{argmax}_y(\beta W^\top z^* + \beta b)^\top y$ consists of a single point $y^* = \operatorname{hs}(W^\top z^* + b)$. The joint distribution $p(y,z|x)$ with parameters $t\beta W, t\alpha V, t\beta b, t\alpha c$ tends to the point measure $\delta_{(y^*,z^*)}(y,z)$ as $t \to \infty$. In this case $p(y|x)$ tends to $\delta_{y^*}(y)$ as $t \to \infty$, where $y^* = \operatorname{hs}(W^\top z^* + b) = \operatorname{hs}(W^\top \operatorname{hs}(Vx + c) + b)$, for all $x \in \{0,1\}^k$. ■

Proof of Theorem 20 The second statement is precisely Lemma 40. For the more general statement the arguments are as follows. Note that the conditional distribution $p(y|z)$ of the output units, given the hidden units, is the same for a CRBM and for its feedforward network version. Furthermore, for each input x, the CRBM output distribution is $p(y|x) = \sum_z (q(z|x) * p(z))p(y|z)$, where

$$q(z|x) = \frac{\exp(z^\top Vx + c^\top z)}{\sum_{z'} \exp(z'^\top Vx + c^\top z')}$$

is the conditional distribution represented by the first layer,

$$p(y,z) = \frac{\exp(z^\top Wy + b^\top y)}{\sum_{y',z'} \exp(z'^\top Wy' + b^\top y')}$$

is the distribution represented by the RBM with parameters $W, b, 0$, and

$$q(z|x) * p(z) = \frac{q(z|x)p(z)}{\sum_{z'} q(z'|x)p(z')}, \quad \text{for all } z,$$

is the renormalized entry-wise product of the conditioned distribution $q(\cdot|x)$ and the RBM hidden marginal distribution

$$p(z) = \sum_y p(y,z).$$

Now, if q is deterministic, then $q(z|x) * p(z)$ is the same as $q(z|x)$, regardless of $p(z)$ (strictly positive). ■

The proof of Theorem 21 builds on the following lemma, which describes a combinatorial property of the deterministic policies that can be approximated arbitrarily well by CRBMs.

Lemma 41 *Consider a function $f\colon \{0,1\}^k \to \{0,1\}^n$. The model $\mathrm{RBM}^k_{n,m}$ can approximate the deterministic policy $p(y|x) = \delta_{f(x)}(y)$ arbitrarily well only if there is a choice of the model parameters W, V, b, c for which*

$$f(x) = \operatorname{hs}(W^\top \operatorname{hs}([W,V]\left[\begin{smallmatrix} f(x) \\ x \end{smallmatrix}\right] + c) + b), \quad \textit{for all } x \in \{0,1\}^k,$$

where the Heaviside function hs *is applied entry-wise to its argument.*

Proof Consider a choice of W, V, b, c. For each input state x, the conditional represented by $\mathrm{RBM}^k_{n,m}$ is equal to the mixture distribution $p(y|x) = \sum_z p(z|x)p(y|x,z)$, with mixture components $p(y|x,z) = p(y|z) \propto \exp((z^\top W + b^\top)y)$ and mixture weights $p(z|x) \propto \sum_{y'} \exp((z^\top W + b^\top)y' + z^\top(Vx+c))$ for all $z \in \{0,1\}^m$. The support of a mixture distribution is equal to the union of the supports of the mixture components with non-zero mixture weights. In the present case, if $\sum_y |p(y|x) - \delta_{f(x)}(y)| \leq \alpha$, then $\sum_y |p(y|x,z) - \delta_{f(x)}(y)| \leq \alpha/\epsilon$ for all z with $p(z|x) > \epsilon$, for any $\epsilon > 0$. Choosing α small enough, α/ϵ can be made arbitrarily small for any fixed $\epsilon > 0$. In this case, for every z with $p(z|x) > \epsilon$, necessarily

$$(z^\top W + b^\top)f(x) \gg (z^\top W + b^\top)y, \quad \text{for all } y \neq f(x), \tag{4}$$

and hence

$$\operatorname{sgn}(z^\top W + b^\top) = \operatorname{sgn}(f(x) - \tfrac{1}{2}).$$

Furthermore, the probability assigned by $p(z|x)$ to all z that do not satisfy Equation (4) has to be very close to zero (upper bounded by a function that decreases with α). The probability of z given x is given by

$$p(z|x) = \frac{1}{Z_{z|x}} \exp(z^\top(Vx+c)) \sum_{y'} \exp((z^\top W + b^\top)y').$$

In view of Equation (4), for all z with $p(z|x) > \epsilon$, if α is small enough, $p(z|x)$ is arbitrarily close to

$$\frac{1}{Z_{z|x}} \exp(z^\top(Vx+c)) \exp((z^\top W + b^\top)f(x)).$$

This holds, in particular, for every z that maximizes $p(z|x)$. Therefore,

$$\operatorname{argmax}_z p(z|x) = \operatorname{argmax}_z z^\top(Wf(x) + Vx + c).$$

Each of these z must satisfy Equation (4). This completes the proof. ■

Proof of Theorem 21 We start with the sufficient condition. The bound $2^k - 1$ follows directly from Proposition 18. For the second bound, note that any function $f\colon \{0,1\}^k \to \{0,1\}^n;\ x \mapsto y$ can be computed by a parallel composition of the functions $f_i\colon x \mapsto y_i$, for all $i \in [n]$. Hence the bound follows from Lemma 40 and the fact that a feedforward linear threshold network with $\frac{3}{k+2}2^k$ hidden units can compute any Boolean function.

We proceed with the necessary condition. Lemma 41 shows that each deterministic policy that can be approximated by $\mathrm{RBM}^k_{n,m}$ arbitrarily well corresponds to the y-coordinate fixed points of a map defined as the composition of two linear threshold functions $\{0,1\}^{k+n} \to \{0,1\}^m;\ (x,y) \mapsto \mathrm{hs}([W,V]\left[\begin{smallmatrix} y \\ x \end{smallmatrix}\right] + c)$ and $\{0,1\}^m \to \{0,1\}^n;\ z \mapsto \mathrm{hs}(W^\top z + b)$. In particular, we can upper bound the number of deterministic policies that can be approximated arbitrarily well by $\mathrm{RBM}^k_{n,m}$, by the total number of compositions of two linear threshold functions; one with $n+k$ inputs and m outputs and the other with m inputs and n outputs.

Let $\mathrm{LTF}(N,M)$ be the number of linear threshold functions with N inputs and M outputs. It is known that (Ojha, 2000; Wenzel et al., 2000)

$$\mathrm{LTF}(N,M) \leq 2^{N^2 M}.$$

The number of deterministic policies that can be approximated arbitrarily well by $\mathrm{RBM}^k_{n,m}$ is thus bounded above by $\mathrm{LTF}(n+k,m)\cdot\mathrm{LTF}(m,n)\leq 2^{m(n+k)^2+nm^2}$. The actual number may be smaller, in view of the fixed-point and shared parameter constraints. On the other hand, the number of deterministic policies in $\Delta_{k,n}$ is as large as $(2^n)^{2^k}=2^{n2^k}$. The claim follows from comparing these two numbers. ■

References

Nihat Ay, Guido Montúfar, and Johannes Rauh. Selection criteria for neuromanifolds of stochastic dynamics. In Y. Yamaguchi, editor, *Advances in Cognitive Neurodynamics (III)*, pages 147–154. Springer, 2013.

Richard E. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NY, 1957.

Yoshua Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127, January 2009.

Maria A. Cueto, Jason Morton, and Bernd Sturmfels. Geometry of the restricted Boltzmann machine. In M. Viana and H. Wynn, editors, *Algebraic Methods in Statistics and Probability II, AMS Special Session*, volume 2. AMS, 2010.

Asja Fischer and Christian Igel. An introduction to restricted Boltzmann machines. In L. Alvarez, M. Mejail, L. Gomez, and J. Jacobo, editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, volume 7441 of *Lecture Notes in Computer Science*, pages 14–36. Springer, 2012.

Yoav Freund and David Haussler. *Unsupervised Learning of Distributions of Binary Vectors Using Two Layer Networks*. Technical report. Computer Research Laboratory, University of California, Santa Cruz, 1994.

Edgar N. Gilbert. A comparison of signalling alphabets. *Bell System Technical Journal*, 31:504–522, 1952.

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

Geoffrey E. Hinton. A practical guide to training restricted Boltzmann machines. In G. Montavon, G. Orr, and K.-R. Müller, editors, *Neural Networks: Tricks of the Trade*, volume 7700 of *Lecture Notes in Computer Science*, pages 599–619. Springer, 2012.

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

Hugo Larochelle and Yoshua Bengio. Classification using discriminative restricted Boltzmann machines. In W. Cohen, A. McCallum, and S. Roweis, editors, *Proceedings of the 25th International Conference on Machine Learning*, pages 536–543. ACM, 2008.

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

James Martens, Arkadev Chattopadhya, Toni Pitassi, and Richard Zemel. On the expressive power of restricted Boltzmann machines. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2877–2885. Curran Associates, Inc., 2013.

Volodymyr Mnih, Hugo Larochelle, and Geoffrey E. Hinton. Conditional restricted Boltzmann machines for structured output prediction. In F. Cozman and A. Pfeffer, editors, *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, pages 514–522. AUAI Press, 2011.

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

Guido Montúfar and Jason Morton. Discrete restricted Boltzmann machines. *Journal of Machine Learning Research*, 16:653–672, 2015.

Guido Montúfar and Jason Morton. When does a mixture of products contain a product of mixtures? *SIAM Journal on Discrete Mathematics*, 29:321–347, 2015.

Guido Montúfar and Johannes Rauh. Scaling of model approximation errors and expected entropy distances. *Kybernetika*, 50(2):234–245, 2014.

Guido Montúfar, Johannes Rauh, and Nihat Ay. Expressive power and approximation errors of restricted Boltzmann machines. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 415–423. Curran Associates, Inc., 2011.

Guido Montúfar, Johannes Rauh, and Nihat Ay. Maximal information divergence from statistical models defined by neural networks. In F. Nielsen and F. Barbaresco, editors, *Geometric Science of Information*, volume 8085 of *Lecture Notes in Computer Science*, pages 759–766. Springer, 2013.

Guido Montúfar, Keyan Ghazi-Zahedi, and Nihat Ay. A theory of cheap control in embodied systems. *PLoS Comput Biol*, 11(9):e1004427, 09 2015.

Piyush C. Ojha. Enumeration of linear threshold functions from the lattice of hyperplane intersections. *IEEE Transactions on Neural Networks*, 11(4):839–850, Jul 2000.

Sheldon M. Ross. *Introduction to Stochastic Dynamic Programming*. Academic Press, Inc., Orlando, FL, USA, 1983.

Ruslan Salakhutdinov, Andriy Mnih, and Geoffrey E. Hinton. Restricted Boltzmann machines for collaborative filtering. In *Proceedings of the 24th International Conference on Machine Learning*, pages 791–798. ACM, 2007.

Brian Sallans and Geoffrey E. Hinton. Reinforcement learning with factored states and actions. *Journal of Machine Learning Research*, 5:1063–1088, 2004.

Paul Smolensky. Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1. chapter Information Processing in Dynamical Systems: Foundations of Harmony Theory, pages 194–281. MIT Press, Cambridge, MA, USA, 1986.

Ilya Sutskever and Geoffrey E. Hinton. Learning multilevel distributed representations for high-dimensional sequences. In M. Meila and X. Shen, editors, *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*, pages 548–555. Journal of Machine Learning Research, 2007.

Graham W. Taylor, Geoffrey E. Hinton, and Sam T. Roweis. Modeling human motion using binary latent variables. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1345–1352. MIT Press, 2007.

Laurens van der Maaten. Discriminative restricted Boltzmann machines are universal approximators for discrete data. Technical Report EWI-PRB TR 2011001, Delft University of Technology, 2011.

Rom R. Varshamov. Estimate of the number of signals in error correcting codes. *Doklady Akad. Nauk SSSR*, 117:739–741, 1957.

Walter Wenzel, Nihat Ay, and Frank Pasemann. Hyperplane arrangements separating arbitrary vertex classes in n-cubes. *Advances in Applied Mathematics*, 25(3):284–306, 2000.

Laurent Younes. Synchronous Boltzmann machines can be universal approximators. *Applied Mathematics Letters*, 9(3):109 – 113, 1996.

Matthew Zeiler, Graham Taylor, Niko Troje, and Geoffrey E. Hinton. Modeling pigeon behaviour using a conditional restricted Boltzmann machine. In *17th European Symposium on Artificial Neural Networks*, 2009.