**Max-Planck-Institut
für Mathematik
in den Naturwissenschaften
Leipzig**

Expressive Power of Conditional Restricted Boltzmann Machines

(revised version: July 2014)

by

Guido Montúfar, Nihat Ay, and Keyan Zahedi

Preprint no.: 16 2014

Expressive Power of Conditional Restricted Boltzmann Machines

Guido Montúfar[1], Nihat Ay[1,2,3], and Keyan Ghazi-Zahedi[1]

[1]Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, 04103 Leipzig, Germany
[2]Department of Mathematics and Computer Science, Leipzig University, PF 10 09 20, 04009 Leipzig, Germany
[3]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

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, parametrized by interaction weights and biases. We address the representational power of these models, proving results on the minimal size of universal approximators of conditional probability distributions, the minimal size of universal approximators of deterministic functions, the maximal model approximation errors, and on the dimension of the set of representable conditional distributions. We contribute new tools for investigating conditional models and obtain significant improvements over the results that can be derived directly 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 recurrent neural 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 (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. For an introduction see (Fischer and Igel 2012). An RBM defines Gibbs-Boltzmann probability distributions over the observable states of the network, depending on the interaction weights and biases. The expressive power of these probability models has attracted much attention in recent years and has been studied in numerous papers, treating, in particular, their universal approximation properties (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 2012), 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. 2012; 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 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. For example: given that there exist RBM universal approximators of probability distributions, there also exist CRBM universal approximators of conditional probability distributions. However, knowing the smallest number of hidden units that make an RBM a universal approximator only provides relatively loose bounds on the smallest number of hidden units that make a CRBM a universal approximator. 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: what is the smallest number of hidden units that suffices for obtaining a model that can 1) approximate any target conditional distribution arbitrarily well (a universal approximator); 2) approximate selected classes of conditional distributions arbitrarily well; 3) approximate any target conditional distribution without exceeding a given error tolerance? We provide non-trivial solutions to all of these problems. We focus on the case of binary units, but the main ideas of our discussion extend to the case of discrete non-binary units. Furthermore, some of the tools that we develop can be applied to study not only CRBMs but also any other conditional models defined by layered stochastic networks.

Section 2 contains formal definitions and elementary properties of CRBMs. In Section 3 we address the universal approximation problem, deriving an upper-bound on the minimal number of hidden units required for this purpose (Theorem 4). In Section 4 we study the approximation of conditional distributions with restricted supports, including bounds on the number of hidden units needed in order to approximate every deterministic conditional arbitrarily well (Theorem 22). In Section 5 we analyze the maximal approximation errors of CRBM models (assuming optimal parameters) and derive an upper-bound for the number of hidden units that suffices to approximate every conditional distribution within a given error tolerance (Theorem 28). In Section 6 we study the dimension of the sets of conditional probability distributions represented by CRBM models and show that for most combinations of input, hidden, and output layer sizes, the model has the dimension expected from counting the parameters of the model (Theorem 32). In Section 7 we offer a discussion and an outlook.

2 Definitions

A Boltzmann machine is an undirected stochastic network with binary units, some of which may be hidden. It defines probabilities for the joint states of its visible units, given by the relative frequencies at which they are observed, asymptotically, depending on the network's parameters. At each time $t \in \mathbb{N}$, this machine selects a unit at random, say unit i, and updates its state according to a Bernoulli draw with success probability $\operatorname{sigm}(\sum_j W_{ij} x_j + b_i)$, where $\operatorname{sigm}(c) = \frac{1}{1+\exp(-c)}$, $x_j \in \{0,1\}$ is the current state of unit j, $W_{ij} \in \mathbb{R}$ is an interaction weight attached to the unit pair $\{i,j\}$, and $b_i \in \mathbb{R}$ is a bias attached to unit i. In the limit of infinite time, each possible joint state $x = (x_V; x_H) = (x_1, \dots, x_{|V|}; x_{|V|+1}, \dots, x_{|V|+|H|})$ of the network's visible and hidden units occurs with a relative frequency described by the Gibbs-Boltzmann probability distribution

$$p(x) = \frac{1}{Z} \exp(-\mathcal{H}(x))$$

with energy function

$$\mathcal{H}(x) = -\sum_{i<j} W_{ij} x_i x_j - \sum_i b_i x_i$$

and normalizing function

$$Z = \sum_{x'} \exp(-\mathcal{H}(x')).$$

The probabilities of the visible states are given by marginalizing out the states of the hidden units; that is, by $p(x_V) = \sum_{x_H} p(x_V, x_H)$.

Definition 1. An RBM is a Boltzmann machine with the restriction that there are no interactions between the visible units nor interactions between the hidden units, such that $W_{ij} \neq 0$ only when unit i is visible and unit j is hidden. See Figure 1 for an illustration of this architecture. The probabilities of the visible states are given by

$$p(x_V) = \sum_{x_H} \frac{1}{Z} \exp\Big(\sum_{i \in V, j \in H} W_{ij} x_i x_j + \sum_{i \in V} b_i x_i + \sum_{j \in H} c_j x_j \Big).$$

We will denote the set of probability distributions on $\{0,1\}^n$ by

$$\Delta_n := \Big\{ (p(x))_{x \in \{0,1\}^n} \in \mathbb{R}^{2^n} : p(x) \geq 0, \sum_x p(x) = 1 \Big\}.$$

Each probability distribution $p \in \Delta_n$ is a vector of 2^n non-negative entries $p(x)$, $x \in \{0,1\}^n$, adding to one. Geometrically, the set Δ_n is a $(2^n - 1)$-dimensional simplex in $\mathbb{R}^{2^n}$.

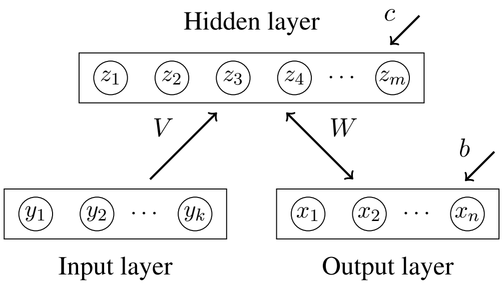


Figure 1: Architecture of a CRBM. An RBM is the special case with $k = 0$.

We will denote by Δ_n^k the $2^k(2^n - 1)$-dimensional polytope of conditional distributions, given by the row-stochastic matrices $(p(x|y))_{x,y}$ with rows $p(\cdot|y) \in \Delta_n$ for each $y \in \{0,1\}^k$. This is the 2^k-fold Cartesian product of the $(2^n - 1)$-dimensional probability simplex; namely

$$\Delta_n^k = \bigtimes_{y\in\{0,1\}^k} \Delta_n,$$

where there is one probability simplex Δ_n for each possible input state y. For example, the polytope Δ_1^k is the Cartesian product of 2^k lines (1-dimensional simplices), which is a 2^k-dimensional cube.

As any multivariate model of probability distributions, an RBM defines models of conditional distributions, according to the relation $p(x,y) = p(y)p(x|y)$. The formal definition of these conditional models is as follows:

Definition 2. The conditional restricted Boltzmann machine (CRBM) model with k input, n output, and m hidden units, denoted $\mathrm{RBM}_{n,m}^k$, is the set of all conditional distributions in Δ_n^k that can be written as

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

with normalization function

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

Here, y, x, and z are joint state (column) vectors of the k input visible units, n output visible units, and m hidden units, respectively ($x^\top$ denotes x transposed).

The model $\mathrm{RBM}_{n,m}^k$ has $(n+k)m + n + m$ parameters. A bias term $a^\top y$ for the input units does not appear in the definition, as it would cancel out with the normalization function Z. When there are no input units ($k = 0$), the model $\mathrm{RBM}_{n,m}^k$ reduces to the restricted Boltzmann machine probability model with n visible and m hidden units, which we denote by $\mathrm{RBM}_{n,m}$.

We can interpret $\mathrm{RBM}_{n,m}^k$ as a collection of 2^k RBMs with shared parameters. For each input y, $p(\cdot|y)$ is the distribution represented by $\mathrm{RBM}_{n,m}$ for the parameters $W, b, (Vy+c)$. In particular, for each choice of the parameters of $\mathrm{RBM}_{n,m}^k$, all $p(\cdot|y)$ are distributions from $\mathrm{RBM}_{n,m}$ with the same interaction weights W, the same biases b for the visible units, but with different biases $(Vy + c)$ for the hidden units. The joint behaviour of these distributions with shared parameters is not trivial.

The direct interpretation of $\mathrm{RBM}_{n,m}^k$ is that it represents block-wise normalized versions of the joint probability distributions represented by $\mathrm{RBM}_{n+k,m}$. Namely, any $p \in \mathrm{RBM}_{n+k,m} \subseteq \Delta_{n+k}$ can be written as a matrix $(p(x,y))_{y\in\{0,1\}^k, x\in\{0,1\}^n}$ with 2^k rows and 2^n columns. Conditioning p on y is equivalent to considering the normalized y-th row $p(x|y) = p(x,y)/\sum_{x'} p(x',y)$ for all $x \in \{0,1\}^n$. The tuple of normalized rows of p is a conditional distribution represented by $\mathrm{RBM}_{n,m}^k$.

3 Universal Approximation: Necessary and Sufficient Number of Hidden Units

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

Note that each conditional distribution $p(x|y)$ can be identified with the set of joint distributions of the form $r(x,y) = q(y)p(x|y)$, with strictly positive marginals $q(y)$. By fixing $q(y)$ equal to the uniform distribution over $\{0,1\}^k$, we obtain an identification of Δ_n^k with $\frac{1}{2^k}\Delta_n^k \subseteq \Delta_{n+k}$. Figure 2A illustrates this identification in the case $n = k = 1$.

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

Proposition 3. *The model* $\mathrm{RBM}_{n,m}^k$ *can approximate every conditional distribution from* Δ_n^k *arbitrarily well whenever* $m \geq \frac{1}{2}2^{k+n} - 1$.

This improves a previous result by van der Maaten (2011). On the other hand, since conditional models do not need to model the input-state distribution $q(y)$, in principle it is possible that $\mathrm{RBM}_{n,m}^k$ is a universal approximator of conditional distributions even if $\mathrm{RBM}_{n+k,m}$ is not a universal approximator of probability distributions. Therefore, we consider an improvement of Proposition 3 that does not follow from corresponding results for RBM probability models:

Theorem 4. *The model* $\mathrm{RBM}_{n,m}^k$ *can approximate every conditional distribution from* Δ_n^k *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 $\mathrm{RBM}_{n,m}^k$ *can approximate every conditional distribution from* Δ_n^k *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 15 and Proposition 17) which tend to approximately* 0.2263 *and* 0.0269, *respectively, as* r *tends to infinity.*

The bound on m decreases with increasing k. We note the following weaker but practical version of Theorem 4:

Corollary 5. *Let* $k \geq 1$. *The model* $\mathrm{RBM}_{n,m}^k$ *can approximate every conditional distribution from* Δ_n^k *arbitrarily well whenever* $m \geq \frac{1}{2}2^k(2^n-1) = \frac{1}{2}2^{k+n} - \frac{1}{2}2^k$.

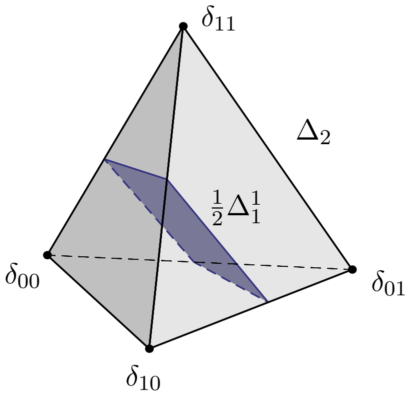


Figure 2: The polytope of conditional distributions Δ_1^1 embedded in the probability simplex Δ_2.

Theorem 4 and also its Corollary 5 represent a substantial improvement of Proposition 3. In particular, they specify universal approximation bounds that reflect the structure of the polytope Δ_n^k as a 2^k-fold product of the $(2^n - 1)$-simplex Δ_n. In contrast, the bound from Proposition 3, which is derived from a universal approximation bound for probability distributions, reflects the structure of the $(2^{n+k} - 1)$-dimensional joint probability simplex Δ_{k+n}. We think that our new results are reasonably tight, although some improvements may still be possible. The proof of Theorem 4 is supported by several lemmas and propositions that we develop in Section 3.1.

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 nodes that suffices for universal approximation can be obtained by comparing the number of parameters of the CRBM model and the dimension of the conditional polytope. This analysis shows that exponentially many hidden units are needed:

Proposition 6. *If the model* $\mathrm{RBM}_{n,m}^k$ *approximates every conditional distribution from* Δ_n^k *arbitrarily well, then necessarily* $m \geq \frac{1}{(n+k+1)}(2^k(2^n-1)-n)$.

The details of this statement are given in Section 3.2.

3.1 Details on the Sufficient Number of Hidden Units (Proof of Theorem 4)

This section contains the proof of Theorem 4 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 7. 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.

If the model $\mathrm{RBM}_{n,m}$ can represent a probability distribution p, then the model $\mathrm{RBM}_{n,m+1}$ can represent the probability distribution $p' = p * q$, where $q = \lambda' r + (1-\lambda')s$ is any mixture, with $\lambda' \in [0,1]$, of two strictly positive product distributions $r(x_1,\dots,x_n) = \prod_{i=1}^n r_i(x_i)$ and $s(x_1,\dots,x_n) = \prod_{i=1}^n s_i(x_i)$ on $\{0,1\}^n$. Choosing r equal to the uniform distribution, we see that $p' = p*q$ can be made equal to $\lambda p + (1-\lambda)p*s$, 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 8. 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 $\lambda \in [0,1]$.

We will need two more standard definitions from coding theory:

Definition 9. A radius-1 *Hamming ball* in $\{0,1\}^k$ is a set B consisting of a length-k binary vector and all its immediate neighbours; that is, $B = \{y \in \{0,1\}^k \colon d_H(y,z) \leq 1\}$ for some $z \in \{0,1\}^k$, where $d_H(y,z) := |\{i \in [k] \colon y_i \neq z_i\}|$ denotes the Hamming distance between y and z.

Definition 10. 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, $C = \{y \in \{0,1\}^k \colon y_i = z_i \text{ for all } i \in \Lambda\}$ 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, and the vectors in a radius-1 Hamming ball correspond to the vertices of a corner of a unit cube. The vectors in a radius-1 Hamming ball are affinely independent. See Figure 3A for an illustration.

In order to prove Theorem 4, 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 sharing step sequences, we will use the following technical lemma:

Lemma 11. *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 $\lambda^y \in (0,1)$ for all $y \in B \cap C$, let $\tilde{x} \in \{0,1\}^n$ and let $\delta_{\tilde{x}}$ denote the Dirac delta on $\{0,1\}^n$ assigning probability one to $\tilde{x}$. Let $p \in \Delta_{n+k}$ be a strictly positive probability distribution with conditionals $p(\cdot|y)$ and let*

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

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

Proof. We define the sharing step $p' = \lambda p + (1-\lambda)p * s$ with a product distribution s supported on $\{\tilde{x}\} \times C \subseteq \{0,1\}^{n+k}$. Note that given any distribution q on C and a radius-1 Hamming ball B in C, there is a product distribution s on C such that $(s_i)_{i\in C\cap B} \propto (q_i)_{i\in C\cap B}$. 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, note that a product distribution is a vector with entries $s(y) = \prod_{i\in[k]} s_i(y_i)$ for all $y = (y_1, \dots, y_k)$, with factor distributions s_i. Hence the restriction of s to B is given by $\Big(\prod_i s_i(0), \frac{s_1(1)}{s_1(0)} \prod_i s_i(0), \dots, \frac{s_k(1)}{s_k(0)} \prod_i s_i(0)\Big)$, where, without loss of generality, we chose B centered at $(0,\dots,0)$. Now, by choosing the factor distributions s_i appropriately, the vector $\big(\frac{s_1(1)}{s_1(0)}, \dots, \frac{s_k(1)}{s_k(0)}\big)$ can be made arbitrary in $\mathbb{R}^k_+$. □

We have the following two implications of Lemma 11:

Corollary 12. *For any $\epsilon > 0$ and any $q(\cdot|y) \in \Delta_n$ for all $y \in B\cap C$, there is an $\epsilon' > 0$ such that, for any strictly positive joint distribution $p \in \Delta_{n+k}$ with conditionals satisfying $\sum_x |p(x|y) - \delta_0(x)| \leq \epsilon'$ for all $y \in B\cap C$, there are $2^n - 1$ sharing steps taking p to a joint distribution p'' with conditionals satisfying $\sum_x |p''(x|y) - p'(x|y)| \leq \epsilon$ for all $y \in (B\cap C) \cup (\{0,1\}^k \setminus C)$, where δ_0 is the Dirac delta on $\{0,1\}^n$ assigning probability one to the vector of zeros and*

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

Note that this corollary does not make any statement about the rows $p''(\cdot|y)$ with $y \in C \setminus B$. When transforming the $(B\cap C)$-rows of p according to Lemma 11, 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 13. *For any $\epsilon > 0$, any cylinder set $C \subseteq \{0,1\}^k$, and any $\tilde{x} \in \{0,1\}^n$, any strictly positive joint distribution p can be transformed by a single probability sharing step to a joint distribution p'' with conditionals satisfying $\sum_x |p''(x|y) - p'(x|y)| \leq \epsilon$ for all $y \in \{0,1\}^k$, where*

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

The sharing step of this corollary can be defined as $p'' = \lambda p + (1-\lambda)p * s$ with s close to the uniform distribution on $\{\tilde{x}\} \times C$ and λ close to 0 (close enough depending on ϵ). We refer to such a sharing step as a *reset* of the C-rows of p.

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. We denote by *star* the intersection of a radius-1 Hamming ball and a cylinder set containing the center of the ball. For simplicity of notation, given a cylinder set $C \subseteq \{0,1\}^k$, we denote by B the largest star in C centered at the smallest element of C, whereby the set of strings $\{0,1\}^k$ is ordered lexicographically. See Figure 3A. The details of the algorithm are given in Algorithm 1.

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

```
Input: Strictly positive joint distribution $p$, target conditional distribution $q(\cdot|\cdot)$, and $\epsilon > 0$
Output: Transformation $p'$ of the input distribution with $\sum_x |p'(x|y) - q(x|y)| \leq \epsilon$ for all $y$
Initialize $\mathcal{B} \leftarrow \emptyset$;
while $\mathcal{B} \not\supseteq \{0,1\}^k$ do
    Choose (disjoint) cylinder sets $C^1, \dots, C^K$ 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 $p(\cdot|y)$ close to $\delta_0$ for all $y \in C^i$ for all $i \in [K]$ and leaving all other rows close to
    their current values, according to Corollary 13;
    for each $i \in [K]$ do
        Perform at most $2^n - 1$ sharing steps taking $p(\cdot|y)$ close to $q(\cdot|y)$ for all $y \in B^i$ and
        leaving the $(\{0,1\}^k \setminus C^i)$-rows close to their current values, according to
        Corollary 12;
    end
    $\mathcal{B} \leftarrow \mathcal{B} \cup (\cup_{i\in[K]} B^i)$;
end
```

Algorithm 1: Algorithmic illustration of the proof of Theorem 4. The algorithm performs sequential sharing steps on a strictly positive joint distribution $p \in \Delta_{n+k}$ until the resulting distribution p' has a conditional distribution $p'(\cdot|\cdot)$ satisfying $\sum_x |p'(x|y) - q(x|y)| \leq \epsilon$ for all y. Here $\mathcal{B} \subseteq \{0,1\}^k$ denotes the set of inputs y that have been readily processed in the current iteration. Given a cylinder set C^i, the set B^i is a star in C^i.

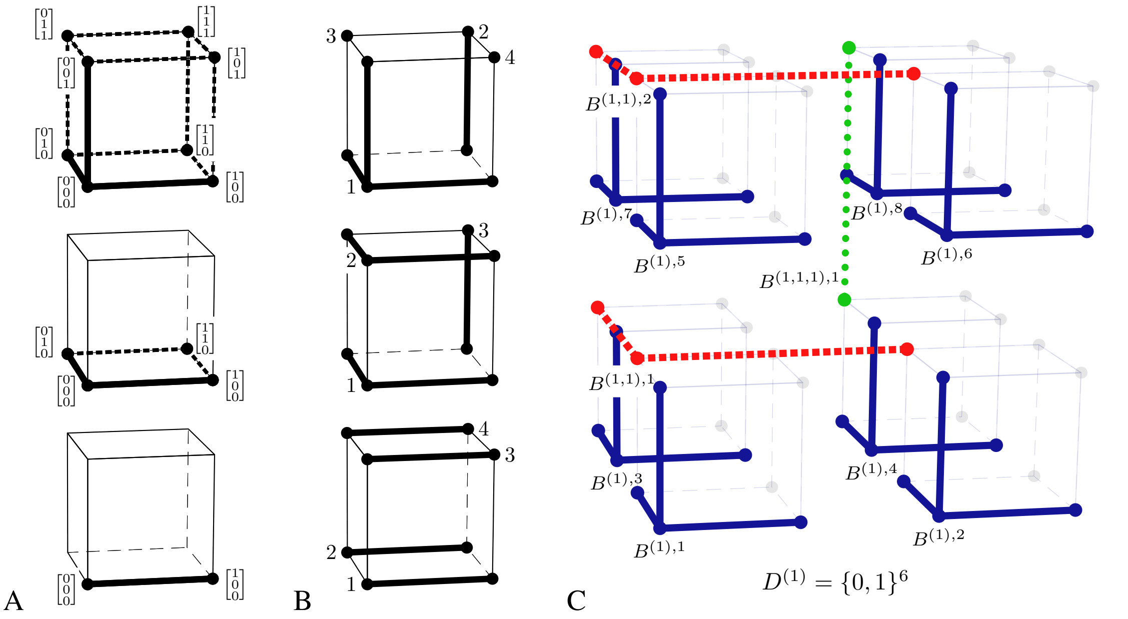
of sharing steps run by Algorithm 1. For this purpose, we investigate the combinatorics of sharing step sequences and evaluate their worst case lengths. We can choose as starting distribution some $p \in \mathrm{RBM}_{n+k,0}$ with conditional distribution satisfying $\sum_x |p(x|y) - \delta_0(x)| \leq \epsilon'$ for all $y \in \{0,1\}^k$, for some $\epsilon' > 0$ small enough depending on the target conditional $q(\cdot|\cdot)$ and the targeted approximation accuracy ϵ.

Definition 14. A sequence of stars $B^1, \dots, B^l$ packing $\{0,1\}^k$ with the property that the smallest cylinder set containing any given star 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 3B 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.

Lemma 15. *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.



$\begin{bmatrix}0\\1\\1\end{bmatrix}$
$\begin{bmatrix}1\\1\\1\end{bmatrix}$
$\begin{bmatrix}0\\0\\1\end{bmatrix}$
$\begin{bmatrix}1\\0\\1\end{bmatrix}$
$\begin{bmatrix}0\\1\\0\end{bmatrix}$
$\begin{bmatrix}1\\1\\0\end{bmatrix}$
$\begin{bmatrix}0\\0\\0\end{bmatrix}$
$\begin{bmatrix}1\\0\\0\end{bmatrix}$
$\begin{bmatrix}0\\1\\0\end{bmatrix}$
$\begin{bmatrix}1\\1\\0\end{bmatrix}$
$\begin{bmatrix}0\\0\\0\end{bmatrix}$
$\begin{bmatrix}1\\0\\0\end{bmatrix}$
$\begin{bmatrix}0\\0\\0\end{bmatrix}$
$\begin{bmatrix}1\\0\\0\end{bmatrix}$
A
3
2
4
1
3
2
1
4
3
2
1
B
$B^{(1,1),2}$
$B^{(1),7}$
$B^{(1),5}$
$B^{(1,1,1),1}$
$B^{(1),8}$
$B^{(1),6}$
$B^{(1,1),1}$
$B^{(1),3}$
$B^{(1),4}$
$B^{(1),1}$
$B^{(1),2}$
$D^{(1)} = \{0,1\}^6$
C

Figure 3: 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 15 for $\{0,1\}^6$.

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

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 $C^{(j_1),i}$, $i \in \{1,\dots,2^{S(r-1)}\}$. 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 $D^{(j_1,j_2)}$, $j_2 \in \{1,\dots,2^r-(r+1)\}$.

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 $p'(\cdot|y) \approx \delta_0$ for all $y \in \cup_{j_1} D^{(j_1,j_2)}$.

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 $C^{(j_1,j_2),i}$, $i \in \{1,\dots,2^{S(r-2)}\}$, and the corresponding stars are included in the sequence.

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 $D^{(j_1,\dots,j_i,j_{i+1})}$, $j_{i+1} \in [2^{r-(i-1)} - (r+1-(i-1))]$.

	$m_{n,k}^{(r)} =$					
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 16. Each row evaluates the universal approximation bound $m_{n,k}^{(r)}$ for a value of r.

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 3C offers an illustration of these star packing sequences. The figure 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, translated versions of the highlighted ones. □

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

Proposition 16 (Theorem 4, general bound)**.** *Let $k \geq S(r)$. The model $\mathrm{RBM}_{n,m}^k$ can approximate every conditional distribution from Δ_n^k 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 given in Lemma 15. □

In order to make the universal approximation bound more comprehensible, in Table 1 we evaluated the sequence $m_{n,k}^{(r)}$ for $r = 1, 2, 3\dots$ 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 4. 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 17 (Theorem 4, 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 that $\prod_{i=0}^{\infty}\left(1 - \frac{i-3}{2^i}\right) \approx 7.7413$. 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 that $K(r) \leq K(r'-1)\prod_{i=r'}^{r}\left(1 - \frac{i-4}{2^i}\right)$, 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. □

This concludes the proof of Theorem 4. 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, a circumstance that we will use in Sections 4 and 5:

Remark 18. We note that if we only want to model a restricted class of conditional distributions, then often it is possible to adapt Algorithm 1 to these restrictions and obtain tighter bounds for the number of hidden units needed to represent these restricted conditionals. For example:

If we only want to model the target conditionals $q(\cdot|y)$ for the inputs y from a subset $\mathcal{S} \subseteq \{0,1\}^k$ and do not care about $q(\cdot|y)$ for $y \notin \mathcal{S}$, 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 C^i the conditionals $q(\cdot|y)$ with $y \in B^i$ have a common support set $T \subseteq \{0,1\}^n$, then the C^i-rows of p can be reset to a distribution δ_x with $x \in T$, and only $|T| - 1$ sharing steps are needed to transform p in order to approximate $q(\cdot|y)$ for all $y \in B^i$ to any desired accuracy. In particular, for the class of target conditional distributions with $\operatorname{supp} q(\cdot|y) = T$ for all y, the term $2^n - 1$ in the complexity bound of Algorithm 1 is replaced by $|T| - 1$.

3.2 Details on the Necessary Number of Hidden Units (Proof of Proposition 6)

Proposition 6 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 $\mathcal{Y}$ and outputs from a finite set $\mathcal{X}$. Accordingly, we denote by $\Delta_{\mathcal{X}}$ the set of probability distributions on $\mathcal{X}$.

Lemma 19. *Let $\mathcal{Y}$, $\mathcal{X}$, 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 $\mathcal{M}' \subseteq \Delta_{\mathcal{Y}\times\mathcal{X}}$ of an exponential family $\mathcal{E} \subseteq \Delta_{\mathcal{Y}\times\mathcal{X}\times\mathcal{Z}}$. If $\mathcal{M}$ is a universal approximator of conditionals from $\Delta_{\mathcal{X}}^{\mathcal{Y}}$, then $\dim(\mathcal{E}) \geq \dim(\Delta_{\mathcal{X}}^{\mathcal{Y}}) = |\mathcal{Y}|(|\mathcal{X}|-1)$.*

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. We consider first the case of probability distributions; that is, the case with $|\mathcal{Y}| = 1$ and $\mathcal{Y}\times\mathcal{X} \cong \mathcal{X}$. 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 $\overline{\mathcal{M}} = \Delta_{\mathcal{X}}$. 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 $\dim(\mathcal{E}) < \dim(\Delta_{\mathcal{X}})$, 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 $\Delta_{\mathcal{X}}$. 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 $\Delta_{\mathcal{X}}$.

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

Proof of Proposition 6. If $\mathrm{RBM}_{n,m}^k$ is a universal approximator of conditionals from Δ_n^k, then the model consisting of all probability distributions of the form $p(x,y) = \frac{1}{Z}\sum_z \exp(z^\top Wx + z^\top Vy + b^\top x + c^\top z + f(y))$ is a universal approximator of probability distributions from Δ_{n+k}. The latter is the marginal of an exponential family of dimension $mn + mk + n + m + 2^k - 1$. Thus, by Lemma 19, $m \geq \frac{2^{n+k}-2^k-n}{(n+k+1)}$. □

4 Approximation of Special Classes of Conditional Distributions

Sometimes in applications one is interested in conditional distributions which assign positive probability only to a limited number of outputs, depending on the input. Consider a collection of subsets $\mathcal{X}_y \subseteq \{0,1\}^n$ for all $y \in \{0,1\}^k$. The subset of Δ_n^k of conditionals with $\mathrm{supp}(p(\cdot|y)) = \mathcal{X}_y$ for all $y \in \{0,1\}^k$ is the Cartesian product $\times_{y\in\{0,1\}^k}\Delta_{\mathcal{X}_y}$, where $\Delta_{\mathcal{X}_y}$ is the subset of Δ_n of distributions with support in $\mathcal{X}_y$. This restricted set of conditionals corresponds to a $\sum_y(|\mathcal{X}_y|-1)$-dimensional face of the polytope Δ_n^k.

By definition, CRBMs represent only strictly positive conditional distributions. However, they can approximate certain conditional distributions with bounded support arbitrarily well. The closure $\overline{\mathrm{RBM}_{n,m}^k}$ of $\mathrm{RBM}_{n,m}^k$ with respect to the Euclidean topology (standard closure of sets in $\mathbb{R}^N$) consists of all elements of Δ_n^k that can be approximated arbitrarily well by $\mathrm{RBM}_{n,m}^k$. Which faces of the polytope Δ_n^k are intersected by this set? How many hidden units are required to obtain a CRBM that can approximate all points in a given face arbitrarily well? Clearly, if there is a joint distribution in $\overline{\mathrm{RBM}_{n+k,m}}$ with support $\cup_y(\{y\} \times \mathcal{X}_y)$, then there is also a conditional distribution $p \in \overline{\mathrm{RBM}^k_{n,m}}$ with $\operatorname{supp}(p(\cdot|y)) = \mathcal{X}_y$ for all y.

Proposition 20. *Let p be a conditional distribution from Δ^k_n with N non-zero entries, $2^k \leq N \leq 2^{k+n}$. Then p can be approximated arbitrarily well by $\mathrm{RBM}^k_{n,m}$ whenever $m \geq N - 1$.*

Proof. This is because the joint probability model $\mathrm{RBM}_{n+k,m}$ can approximate any probability distribution with support of cardinality $m + 1$ arbitrarily well (Montúfar and Ay 2011). □

There are cases where the bound from Proposition 20 can be reduced. In particular, if there are no restrictions on the supports, Theorem 4 gives a better bound. Taking advantage of Remark 18, we obtain the following generalization of Theorem 4, addressing restricted support conditionals:

Proposition 21. *Let r be an integer with $k \geq S(r)$, and let $\mathcal{Z}$ be a subset of $\{0,1\}^n$ of cardinality $|\mathcal{Z}| = L$. The model $\mathrm{RBM}^k_{n,m}$ can approximate any conditional $p \in \Delta^k_n$ with $\operatorname{supp}(p(\cdot|y)) \subseteq \mathcal{Z}$ for all $y \in \{0,1\}^k$ arbitrarily well, whenever $m \geq 2^{k-S(r)}F(r)(L-1) + R(r)$.*

Proof. This is analogous to the proof of Proposition 16. 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 y. □

The extreme points of Δ^k_n are called *deterministic policies*. These are the conditional distributions that assign positive probability to exactly one output given each input. Each deterministic policy p in Δ^k_n corresponds to a function $f\colon \{0,1\}^k \to \{0,1\}^n$ with $p(x|y) = \delta_{f(y)}(x)$ for all $y \in \{0,1\}^k$ and $x \in \{0,1\}^n$. The name “policy” comes from the context of reinforcement learning (Sutton and Barto 1998), where conditional distributions describe action selection mechanisms.

Ay et al. (2013) show that there are two-dimensional manifolds which contain all deterministic policies of Δ^k_n for any n, k. In the case of CRBMs we cannot expect that two parameters will suffice for this purpose. Proposition 20 implies that $\mathrm{RBM}^k_{n,m}$ can approximate every deterministic policy from Δ^k_n arbitrarily well whenever $m \geq 2^k - 1$. The following result complements this sufficient condition by a necessary condition:

Theorem 22. *The model $\mathrm{RBM}^k_{n,m}$ can approximate every deterministic policy from Δ^k_n arbitrarily well if $m \geq 2^k - 1$ and only if $m \geq 2^{k/2} - \frac{(n+k)^2}{2n}$.*

Note that the approximation of all deterministic policies requires a CRBM with exponentially many hidden units, with respect to the number of input units.

Remark 23. Theorem 22 has the following direct implication for RBMs. There are subsets of $\{0,1\}^{n+k}$ of cardinality 2^k which are not the support sets of any probability distributions in $\overline{\mathrm{RBM}_{n+k,m}}$ unless $m \geq 2^{k/2} - \frac{(n+k)^2}{2n}$.

In order to prove Theorem 22, we use the following combinatorial property of the deterministic policies that can be approximated arbitrarily well by CRBMs. Recall that the Heaviside step function hs maps a real number a to 0 if $a < 0$, to $1/2$ if $a = 0$, and to 1 if $a > 0$. We have:

Lemma 24. *Consider a function $f\colon \{0,1\}^k \to \{0,1\}^n$. The model $\mathrm{RBM}^k_{n,m}$ can approximate the deterministic policy $p(x|y) = \delta_{f(y)}(x)$ arbitrarily well only if there is a choice of the model parameters W, V, b, c for which*

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

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

The equation in Lemma 24 means that for each input state $y \in \{0,1\}^k$, each of the most likely hidden states of $\mathrm{RBM}^k_{n,m}$ given y and $x = f(y)$ has $f(y)$ as its most likely output state.

Proof. Consider a choice of W, V, b, c. For each input state y, the conditional represented by $\mathrm{RBM}^k_{n,m}$ is equal to the mixture distribution $p(x|y) = \sum_z p(z|y)p(x|y,z)$, with mixture components $p(x|y,z) = p(x|z) \propto \exp((z^\top W + b)x)$ and mixture weights $p(z|y) \propto \sum_{x'} \exp((z^\top W + b^\top)x' + z^\top(Vy + c))$ 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 $\sum_x |p(x|y) - \delta_{f(y)}(x)| \le \alpha$, then $\sum_x |p(x|y,z) - \delta_{f(y)}(x)| \le \alpha/\epsilon$ for all z with $p(z|y) > \epsilon$, 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 $p(z|y) > \epsilon$, necessarily

$$(z^\top W + b^\top) f(y) \gg (z^\top W + b^\top)x, \quad \text{for all } x \ne f(y), \tag{3}$$

and hence

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

Furthermore, the probability assigned by $p(z|y)$ to all z that do not satisfy Equation (3) has to be very close to zero (upper bounded by a function that decreases with α). The probability of z given y is given by

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

In view of Equation (3), for all z with $p(z|y) > \epsilon$, if α is small enough, $p(z|y)$ is arbitrarily close to

$$\frac{1}{Z_{z|y}} \exp(z^\top(Vy + c)) \exp((z^\top W + b^\top) f(y)). \tag{4}$$

If ϵ is chosen small enough from the beginning, Equation (4) holds especially for each z that maximizes $p(z|y)$, and so

$$\begin{aligned} \mathrm{argmax}_z\, p(z|y) &= \mathrm{argmax}_z\, z^\top(Wf(y) + Vy + c) \\ &= \Big\{ z\colon\ \mathrm{sgn}([W, V] \left[\begin{smallmatrix} f(y) \\ y \end{smallmatrix}\right] + c) \doteq \mathrm{sgn}(z - \tfrac{1}{2}) \Big\}. \end{aligned}$$

Each of these z satisfies Equation (3). Here $\doteq$ signifies that the condition holds in the entries where the left hand side is non-zero. This completes the proof. □

Proof of Theorem 22. Recall that a linear threshold function with N input bits and M output bits is a function of the form $\{0,1\}^N \to \{0,1\}^M$; $x \mapsto \mathrm{hs}(Wx + b)$ with $W \in \mathbb{R}^{M\times N}$ and $b \in \mathbb{R}^M$.

Lemma 24 shows that each deterministic policy that can be approximated by $\mathrm{RBM}_{n,m}^{k}$ arbitrarily well corresponds to the x-coordinate fixed points of a map defined as the composition of two linear threshold functions $\{0,1\}^{n+k} \to \{0,1\}^{m}$; $(x,y) \mapsto \mathrm{hs}([W,V]\left[\begin{smallmatrix}x\\y\end{smallmatrix}\right] + c)$ 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 $\mathrm{RBM}_{n,m}^{k}$, 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 $\mathrm{RBM}_{n,m}^{k}$ 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 much smaller, in view of the fixed-point and shared parameter constraints. On the other hand, the number of deterministic policies in Δ_{n}^{k} is as large as $(2^n)^{2^k} = 2^{n2^k}$. The claim follows from comparing these two numbers. □

5 Approximation Errors

In the following, we define the Kullback-Leibler divergence of conditional distributions and investigate the divergence from target conditional distributions to their best approximations within CRBM models.

The Kullback-Leibler divergence or relative entropy from some p to some q in the joint probability simplex Δ_{n+k} is given by

$$D(p\|q) := \sum_{y}\sum_{x} p(y)p(x|y)\log\frac{p(y)p(x|y)}{q(y)q(x|y)} = D(p_Y\|q_Y) + \sum_{y} p(y)D(p(\cdot|y)\|q(\cdot\|y)),$$

where $p_Y = \sum_{x\in\{0,1\}^n} p(x,y)$ denotes the marginal distribution over $\{0,1\}^k$. We measure the divergence from a conditional distribution $p(\cdot|\cdot)$ to another conditional distribution $q(\cdot|\cdot)$ in Δ_{n}^{k} by

$$D(p(\cdot|\cdot)\|q(\cdot|\cdot)) := \sum_{y} u_Y(y)D(p(\cdot|y)\|q(\cdot|y)),$$

where u_Y denotes the uniform distribution over y. In other words, we measure the divergence of conditionals as the average of the divergences of their rows. We can write this as

$$D(p(\cdot|\cdot)\|q(\cdot|\cdot)) = D(u_Y p(\cdot|\cdot)\|q_Y q(\cdot|\cdot)) - D(u_Y\|q_Y), \quad \text{for all } q_Y \in \Delta_{k}^{+},$$

where Δ_{k}^{+} denotes the set of all strictly positive distributions on $\{0,1\}^k$. Note that the left hand side is independent of the marginal q_Y, and that the divergence of conditionals may vanish even if the divergence of their joints does not.

The divergence from a conditional $p(\cdot|\cdot)$ to the set of conditionals $\mathcal{M}_{n}^{k}$ represented by a joint model $\mathcal{M}_{n+k}$ is given by

$$D(p(\cdot|\cdot)\|\mathcal{M}_{n}^{k}) = \inf_{q\in\mathcal{M}_{n+k}} D(u_Y p(\cdot|\cdot)\|q) - D(u_Y\|q_Y).$$

This shows that we can estimate the approximation errors of conditional distributions by estimating the approximation errors of joint distributions with uniform marginal $p_Y = u_Y$. By the joint convexity of the divergence, we have that $\max_{p\in\Delta_{n+k}\colon p_Y=u_Y} D(p\|u) = n$. The maximizers are the distributions $p(x,y) = u(y)p(x|y)$ with deterministic conditionals $p(\cdot|\cdot)$.

The maximum over all possible target conditionals satisfies

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

Proposition 25. *Let $m \leq 2^{(n+k)-1} - 1$. The divergence from any conditional distribution $p(\cdot|\cdot) \in \Delta_n^k$ to the model $\mathrm{RBM}_{n,m}^k$ is bounded by*

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

Proof. We have that $D_{\mathrm{RBM}_{n,m}^k} \leq \max_{p\in\Delta_{n+k}\colon p_Y=u_Y} D(p\|\mathrm{RBM}_{n+k,m})$. 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). □

Proposition 25 implies the universal approximation bound given in Proposition 3, as the latter corresponds to the cases where the approximation error vanishes. However, Proposition 25 does not imply Theorem 4 in the same way. Therefore, we want to consider an improvement. We will upper bound the approximation errors of CRBMs by the approximation errors of submodels of CRBMs. First, we note the following:

Lemma 26. *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 $\mathcal{M} = \times_{y\in\{0,1\}^k}\mathcal{N} \subseteq \Delta_n^k$ for some $\mathcal{N} \subseteq \Delta_n$, then $D_{\mathcal{M}} \leq D_{\mathcal{N}}$.*

Proof. For any $p \in \Delta_n^k$, we have

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

□

Given a partition $\mathcal{Z} = \{\mathcal{X}_1,\dots,\mathcal{X}_L\}$ 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 $\{x \in \{0,1\}^n\colon x_{[l]} = z\}$ 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 $\mathrm{RBM}_{n,m}^k$ can approximate certain products of partition models arbitrarily well:

Proposition 27. *Let $\mathcal{Z} = \{0,1\}^l$ with $l \leq n$. Let r be any integer with $k \geq S(r)$. The model $\mathrm{RBM}_{n,m}^k$ can approximate any conditional distribution from the product of partition models $\mathcal{P}_{\mathcal{Z}}^k := \mathcal{P}_{\mathcal{Z}} \times \cdots \times \mathcal{P}_{\mathcal{Z}}$ arbitrarily well whenever $m \geq 2^{k-S(r)}F(r)(L-1) + R(r)$.*

Proof. This is analogous to the proof of Proposition 21, with a few differences. Each element z of $\mathcal{Z}$ corresponds to a cylinder set $\{x \in \{0,1\}^n \colon x_{[l]} = z\}$ 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. □

In turn, we obtain the following approximation error bound:

Theorem 28. *Let $l \in [n]$. The divergence from any conditional distribution in Δ_n^k to the model $\mathrm{RBM}_{n,m}^k$ is bounded from above by*

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

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

Theorem 28 implies the universal approximation result given in Theorem 4 as the special case with vanishing approximation error. We note the following weaker but practical version of Theorem 28, which is the analogue of Corollary 5:

Corollary 29. *Let $k \geq 1$ and $l \in [n]$. The divergence from any conditional distribution in Δ_n^k to the model $\mathrm{RBM}_{n,m}^k$ is bounded from above by $D_{\mathrm{RBM}_{n,m}^k} \leq n - l$, whenever $m \geq \frac{1}{2} 2^k (2^l - 1)$.*

For the theorem we used the approximability of the partition model with blocks $\{x \colon x_\lambda = z\}$, $z \in \{0,1\}^l$, where $\lambda = [l]$. In fact, $\mathrm{RBM}_{n,m}^k$ can approximate all powers of partition models with λ equal to any subset of $[n]$ of cardinality l. This larger class of conditionals does not yield improvements of the universal approximation bound, but it can be used to improve the approximation error bound. An analysis in this direction was provided by Montúfar et al. (2011) for RBM probability models.

6 Dimension

Now we study the dimension of the set of conditional distributions $\mathrm{RBM}_{n,m}^k$ (as a manifold, possibly with singularities). Since this model is defined by marginalizing out the hidden units, several choices of parameters may represent the same conditional distributions. Hence, in principle, the dimension of a CRBM model may be smaller than the number of model parameters. When the dimension is equal to the number of parameters, $(n+k)m + n + m$, or to the maximal possible dimension, $\dim(\Delta_n^k) = 2^k(2^n - 1)$, then $\mathrm{RBM}_{n,m}^k$ is said to have the *expected dimension*. We will show that $\mathrm{RBM}_{n,m}^k$ has the expected dimension for most triplets (n, k, m).

A direct way of evaluating the dimension of a parametric model is to compute the rank of the Jacobian of its parametrization. This approach leads to combinatorial problems that can be largely captured by a piece-wise linearized version of the model, called the *tropical model*. Cueto et al.

(2010) used this perspective to study the dimension of RBM probability models. Here we follow the same idea in order to study the dimension of the conditional model.

We start by recalling the following two functions from coding theory, which are useful to express some of the combinatorics involved in the analysis:

Definition 30. Let $A_2(n,d)$ denote the cardinality of the largest subset of $\{0,1\}^n$ whose elements are at least Hamming distance d apart. Let $K_2(n,d)$ denote the smallest cardinality of a subset of $\{0,1\}^n$ such that no element of $\{0,1\}^n$ is farther than Hamming distance d apart from an element of that set.

For the probability model $\mathrm{RBM}_{n,m}$, Cueto et al. (2010) obtained the following result:

- $\dim(\mathrm{RBM}_{n,m}) = nm + n + m$ for $m + 1 \leq A_2(n,3)$.
- $\dim(\mathrm{RBM}_{n,m}) = 2^n - 1$ for $m \geq K_2(n,1)$.

This shows that $\mathrm{RBM}_{n,m}$ has the expected dimension for most pairs (n,m). From the dimension of the probability model $\mathrm{RBM}_{n+k,m}$, we can lower-bound the dimension of the conditional model $\mathrm{RBM}^k_{n,m}$ in the following way:

Proposition 31. $\dim(\mathrm{RBM}^k_{n,m}) \geq \dim(\mathrm{RBM}_{n+k,m}) - (2^k - 1)$*, and*

- $\dim(\mathrm{RBM}^k_{n,m}) \geq (n+k)m + n + m + k - (2^k - 1)$ *for* $m + 1 \leq A_2(n+k,3)$.
- $\dim(\mathrm{RBM}^k_{n,m}) \geq 2^k(2^n - 1)$ *for* $m \geq K_2(n+k,1)$.

Proof. Each joint distribution of x and y has the form $p(x,y) = p(y)p(x|y)$ and the set Δ_k of all marginals $p(y)$ has dimension $2^k - 1$. This shows the first statement. The items follow directly from the corresponding items for the probability model. □

Proposition 31 shows that if the probability model $\mathrm{RBM}_{n+k,m}$ is full dimensional, with dimension equal to the dimension of the joint probability simplex $\dim(\Delta_{n+k}) = 2^{n+k} - 1$, then the conditional model is full dimensional, with dimension equal to the dimension of the conditional polytope $\dim(\Delta^k_n) = 2^k(2^n - 1)$. Otherwise, however, the lower bounds from the proposition are too loose and do not allow us to attest whether the conditional model has the expected dimension or not, even if the probability model has the expected dimension. Hence we need to study the conditional model in more detail. We prove the following result:

Theorem 32. *The conditional model* $\mathrm{RBM}^k_{n,m}$ *has the expected dimension in the following cases:*

- *When* $\{0,1\}^{n+k}$ *admits* m *disjoint radius-*1 *Hamming balls whose union does not contain any cylinder set* $[y] \subseteq \{0,1\}^{n+k}$, $y \in \{0,1\}^k$ *and has a full rank complement, then*

  $$\dim(\mathrm{RBM}^k_{n,m}) = (n+k+1)m + n.$$

  The condition is satisfied, for example, if $m + 1 \leq A_2(n+k,4)$.
- *When* $m \geq K_2(n+k,1)$*, then* $\dim(\mathrm{RBM}^k_{n,m}) = 2^k(2^n - 1)$.

Proof. The rank of the Jacobian of $\mathrm{RBM}^k_{n,m}$, with the parametrization from Definition 2, depends on the parameter $\theta = (W, V, b, c) \in \mathbb{R}^N$, $N = n + m + (n + k)m$. The maximum over all possible choices of θ is the dimension of the model. Let $h_\theta(v) := \mathrm{argmax}_{z\in\{0,1\}^m} p(z|v)$ denote the most likely hidden state of the RBM given the visible state $v = (x, y)$. 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

$$\mathcal{A}_\theta = \big[(1, x, y)|(1, x, y) \otimes h_\theta(x, y)\big]_{(x,y)\in\{0,1\}^{n+k}} , \tag{5}$$

modulo vectors whose (x, y)-th entries are independent of x given y. Here $\otimes$ is the Kronecker product defined by $(a_{ij})_{i,j} \otimes (b_{kl})_{k,l} = (a_{ij}b_{kl})_{ik,jl}$. The modulo has the effect of disregarding $p(y)$ in the joint distributions $p(x, y) = p(y)p(x|y)$. For example, from the first block of $\mathcal{A}_\theta$ we can remove the columns that correspond to y, 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 x given y 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 called the *tropical CRBM morphism*, and its image is called the *tropical CRBM model*.

Each linear region of the tropical morphism corresponds to an inference function $h_\theta \colon \{0, 1\}^{n+k} \to \{0, 1\}^m$ taking visible state vectors to the most likely hidden state vectors. Geometrically, any inference function corresponds to m slicings of the $(n + k)$-dimensional unit hypercube. Namely, every hidden unit divides the visible space $\{0, 1\}^{n+k} \subset \mathbb{R}^{n+k}$ 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 $(1, v_1, \dots, v_{n+k})$ for all $v \in \{0, 1\}^{n+k}$, 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 $A_{\{0,1\}^{n+k}\setminus\cup_{i\in[m]}C_i}$ (which is the remainder rank of the first block A). The column-rank modulo functions of y is equal to the rank minus $k + 1$ (which is the dimension of the functions of y spanned by columns of A), minus at most the number of cylinder sets $[y] = \{(x, y) \colon x \in \{0, 1\}^n\}$ for some $y \in \{0, 1\}^k$ that are contained in $\cup_{i\in[m]}C_i$. This completes the proof of the general statement in the first item.

The example given in the first item is a consequence of the following observations. Each cylinder set $[y]$ contains 2^n points. If a given cylinder set $[y]$ 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 $[y]$.

The second item is by the second item of Proposition 31; when the probability model $\mathrm{RBM}_{n+k,m}$ is full dimensional, then $\mathrm{RBM}^k_{n,m}$ is full dimensional. □

$m \backslash n+k$	2	3	4	5	6	7	8	9	10
1	✔	✔	✔	✔	✔	✔	✔	✔	✔
2	✓	✓				✔	✔	✔	✔
3	✓	✓				✔	✔	✔	✔
4	✓	✓	✓				✔	✔	✔
5	✓	✓	✓				✔	✔	✔
6	✓	✓	✓				✔	✔	✔
7	✓	✓	✓				✔	✔	✔
8	✓	✓	✓	✓					✔
9	✓	✓	✓	✓					✔
10	✓	✓	✓	✓					✔

Table 2: Illustration of Corollary 33.

In order to make the conditions on m and $(n+k)$ more explicit, in the following corollary we insert bounds on the functions $A_2(n+k,4)$ and $K_2(n+k,1)$ appearing in the theorem:

Corollary 33. *The conditional model* $\mathrm{RBM}_{n,m}^k$ *has the expected dimension in the following cases:*

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

Proof. For the maximal cardinality of distance-4 binary codes of length l it is known that $A_2(l,4) \geq 2^r$, where r is the largest integer with $2^r < \frac{2^{l+1}}{l^2-l+2}$ (Gilbert 1952; Varshamov 1957), and so $A_2(l,4) \geq 2^{l-\lfloor \log_2(l^2-l+2)\rfloor}$. On the other hand, for the minimal size of radius one covering codes of length l it is known that $K_2(l,1) \leq 2^{l-\lfloor \log_2(l+1)\rfloor}$ (Cueto et al. 2010). □

In particular, when m is large enough, the model may be full dimensional without being a universal approximator. The same observation holds for RBM probability models. In fact, $\mathrm{RBM}_{3,2}$ is an example of a full dimensional RBM model which is not a universal approximator (see Montúfar and Morton 2012). See Table 2 for an illustration of Corollary 33.

7 Conclusion

This paper gives a theoretical contribution towards understanding the representational capabilities of CRBMs. CRBMs are based on the well studied RBM probability models. However, their analysis differs substantially from one another, because the RBM tuples defined by CRBMs share parameters and cannot be treated independently. We proved an extensive series of results that generalize recent theoretical work on RBMs in a non-trivial way.

We addressed the classical problem of universal approximation, for which we showed that a CRBM with k input, n output, and m hidden units can approximate every conditional distribution of an n-bit random variable given a k-bit random variable without surpassing a Kullback-Leibler approximation error bound of the form $n+k-1-\log_2(2^{k-1}+m)$ (assuming optimal model parameters). Thus this model is a universal approximator whenever $m \geq \frac{1}{2}2^k(2^n - 1)$. In fact, for large k, larger than about 50, our approximation error upper-bound has the form $n + k - 2.1437 - \log_2(2^{k-2.3263} + m)$, implying universal approximation whenever $m \geq \frac{1}{4}2^k(2^n - 29/30)$. We think that this result is reasonably tight and that it closely describes the actual maximal model-approximation-error of CRBMs. 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 certain improvements may still be possible.

It is worth mentioning that the set of target conditionals for which the approximation error is maximal may be very small. We note that, besides upper-bounding the worst-case approximation error, our results can be plugged into certain analytic integrals (Montúfar and Rauh 2014) to produce upper-bounds for the expectation value of the approximation error when approximating conditionals drawn from product Dirichlet densities on the conditional polytope.

For future work it would be interesting to extend our (optimal-parameter) model approximation error analysis by an analysis of the CRBM training complexity or the errors resulting from non-optimal parameter choices, and to test our theoretical bounds empirically.

In addition to the error bounds, we showed that a CRBM can approximate all binary functions with k input bits and n output bits if the number of hidden units satisfies $m \geq 2^k - 1$ and only if $m \geq 2^{k/2} - (n+k)^2/2n$. This implies, in particular, 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 also studied a slightly more abstract problem, about the dimension of the set of conditional distributions that can be represented by CRBMs. We showed that for all CRBMs with up to exponentially many hidden units (in the number of inputs and outputs), the dimension of the set of representable conditional distributions is equal to the number of CRBM model parameters, $(n + k)m + m + n$. Thus, in all practical cases, CRBMs do not waste parameters, and, generically, there are only finitely many choices of the interaction and bias weights for which the model produces the same conditional distribution.

The results presented in this work represent considerable advances relating the model complexity and model accuracy of CRBMs. 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 and 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. As a final remark, we note that although we focussed on the case of binary units, the main ideas of our discussion extend to the case of discrete non-binary units. However the combinatorics of the non-binary case is more challenging, with moderate efforts it is possible to derive non-binary versions of our theorems.

Acknowledgment

We would like to acknowledge support for this project 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.

References

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

Y. Bengio. Learning deep architectures for AI. *Found. Trends Mach. Learn.*, 2(1):1–127, Jan. 2009.

M. A. Cueto, J. Morton, and B. Sturmfels. Geometry of the restricted Boltzmann machine. In M. Viana and H. Wynn, editors, *Algebraic methods in statistics and probability II, AMS Special Session*. AMS, 2010.

A. Fischer and C. 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 Berlin Heidelberg, 2012.

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

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

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

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

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

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

J. Martens, A. Chattopadhya, T. Pitassi, and R. 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.

V. Mnih, H. Larochelle, and G. E. Hinton. Conditional restricted Boltzmann machines for structured output prediction. *CoRR*, abs/1202.3748, 2012.

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

G. F. Montúfar and J. Morton. When does a mixture of products contain a product of mixtures? *arXiv preprint arXiv:1206.0387*, 2012.

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

G. F. Montúfar, J. Rauh, and N. 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.

G. F. Montúfar, J. Rauh, and N. Ay. Maximal information divergence from statistical models defined by neural networks. In F. Nielsen and F. Barbaresco, editors, *Geometric Science of Information*, LNCS 8085, pages 759–766. Springer, 2013.

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

R. Salakhutdinov, A. Mnih, and G. E. Hinton. Restricted Boltzmann machines for collaborative filtering. In *Proceedings of the 24th International Conference on Machine Learning (ICML 2007)*, pages 791–798. ACM, 2007.

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

P. 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, 1986.

I. Sutskever and G. E. Hinton. Learning multilevel distributed representations for high-dimensional sequences. In M. Meila and X. Shen, editors, *AISTATS*, volume 2 of *JMLR Proceedings*, pages 548–555. JMLR.org, 2007.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

G. W. Taylor, G. E. Hinton, and S. 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.

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

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

W. Wenzel, N. Ay, and F. Pasemann. Hyperplane arrangements separating arbitrary vertex classes in n-cubes. *Adv. Appl. Math.*, 25(3):284–306, 2000.

M. Zeiler, G. Taylor, N. Troje, and G. E. Hinton. Modeling pigeon behaviour using a conditional restricted Boltzmann machine. In *17th European Symposium on Artificial Neural Networks (ESANN)*. 2009.