Max-Planck-Institut
für Mathematik
in den Naturwissenschaften
Leipzig

Discrete Restricted Boltzmann Machines

by

Guido Montúfar and Jason Morton

Preprint no.: 106 2014

Discrete Restricted Boltzmann Machines

Guido Montúfar MONTUFAR@MIS.MPG.DE
Max Planck Institute for Mathematics in the Sciences
Inselstrasse 22, 04103 Leipzig, Germany

Department of Mathematics
Pennsylvania State University
University Park, PA 16802, USA

Jason Morton MORTON@MATH.PSU.EDU
Department of Mathematics
Pennsylvania State University
University Park, PA 16802, USA

Editor: Representation Learning

Abstract

We describe discrete restricted Boltzmann machines: probabilistic graphical models with bipartite interactions between visible and hidden discrete variables. Examples are binary restricted Boltzmann machines and discrete naïve Bayes models. We detail the inference functions and distributed representations arising in these models in terms of configurations of projected products of simplices and normal fans of products of simplices. We bound the number of hidden variables, depending on the cardinalities of their state spaces, for which these models can approximate any probability distribution on their visible states to any given accuracy. In addition, we use algebraic methods and coding theory to compute their dimension.

Keywords: Restricted Boltzmann Machine, Naïve Bayes Model, Representational Power, Distributed Representation, Expected Dimension

1. Introduction

A restricted Boltzmann machine (RBM) is a probabilistic graphical model with bipartite interactions between an observed set and a hidden set of units (see Smolensky, 1986; Freund and Haussler, 1991; Hinton, 2002, 2010). A characterizing property of these models is that the observed units are independent given the states of the hidden units and vice versa. This is a consequence of the bipartiteness of the interaction graph and does not depend on the units' state spaces. Typically RBMs are defined with binary units, but other types of units have also been considered, including continuous, discrete, and mixed type units (see Welling et al., 2005; Marks and Movellan, 2001; Salakhutdinov et al., 2007; Dahl et al., 2012; Tran et al., 2011). We study discrete RBMs, also called multinomial or softmax RBMs, which are special types of exponential family harmoniums (Welling et al., 2005). While each unit X_i of a binary RBM has the state space $\{0, 1\}$, the state space of each unit X_i of a discrete RBM is a finite set $\mathcal{X}_i = \{0, 1, \dots, r_i - 1\}$. Like binary RBMs, discrete RBMs can be trained using contrastive divergence (CD) (Hinton, 1999, 2002; Carreira-Perpiñán and Hinton, 2005) or expectation-maximization (EM) (Dempster et al., 1977) and can be used to train the parameters of deep systems layer by layer (Hinton et al., 2006; Bengio et al., 2007).

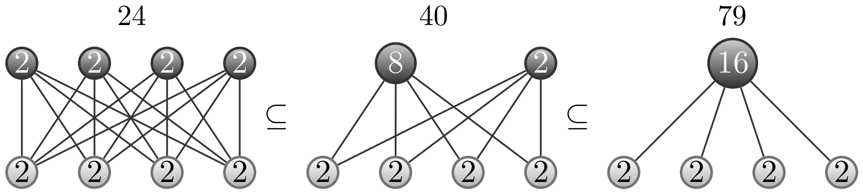

Figure 1: Examples of probability models treated in this paper, in the special case of binary visible variables. The light (dark) nodes represent visible (hidden) variables with the indicated number of states. The total parameter count of each model is indicated at the top. From left to right: a binary RBM; a discrete RBM with one 8-valued and one binary hidden units; and a binary naïve Bayes model with 16 hidden classes.

Non-binary visible units are natural because they can directly encode non-binary features. The situation with hidden units is more subtle. States that appear in different hidden units can be activated by the same visible vector, but states that appear in the same hidden unit are mutually exclusive. Non-binary hidden units thus allow one to explicitly represent complex exclusive relationships. For example, a discrete RBM topic model would allow some topics to be mutually exclusive and other topics to be mixed together freely. This provides a better match to the semantics of several learning problems, although the learnability of such representations is mostly open. The practical need to represent mutually exclusive properties is evidenced by the common approach of adding activation sparsity parameters to binary RBM hidden states, which artificially create mutually exclusive non-binary states by penalizing models which have more than a certain percentage of hidden units active.

A discrete RBM is a *product of experts* (Hinton, 1999); each hidden unit represents an expert which is a mixture model of product distributions, or naïve Bayes model. Hence discrete RBMs capture both naïve Bayes models and binary RBMs, and interpolate between non-distributed mixture representations and distributed mixture representations (Bengio, 2009; Montúfar and Morton, 2014). See Figure 1. Naïve Bayes models have been studied across many disciplines. In machine learning they are most commonly used for classification and clustering, but have also been considered for probabilistic modeling (Lowd and Domingos, 2005; Montúfar, 2013). Theoretical work on binary RBM models includes results on universal approximation (Freund and Haussler, 1991; Le Roux and Bengio, 2008; Montúfar and Ay, 2011), dimension and parameter identifiability (Cueto et al., 2010), Bayesian learning coefficients (Aoyagi, 2010), complexity (Long and Servedio, 2010), approximation errors (Montúfar et al., 2011). In this paper we generalize some of these theoretical results to discrete RBMs.

Probability models with more general interactions than strictly bipartite have also been considered, including semi-restricted Boltzmann machines and higher-order interaction Boltzmann machines (see Sejnowski, 1986; Memisevic and Hinton, 2010; Osindero and Hinton, 2008; Ranzato et al., 2010). The techniques that we develop in this paper also serve to treat a general class of RBM-like models allowing within-layer interactions, a generalization that will be carried out in a forthcoming work (Montúfar and Morton, 2013), as well as narrow deep belief network models (Montúfar, 2014).

Section 2 collects basic facts about independence models, naïve Bayes models, and binary RBMs, including an overview on the aforementioned theoretical results. Section 3 defines discrete RBMs formally and describes them as (i) products of mixtures of product distributions (Proposition 7) and (ii) as restricted mixtures of product distributions. Section 4 elaborates on distributed representations and inference functions represented by discrete RBMs (Proposition 11, Lemma 12, and Proposition 14). Section 5 addresses the expressive power of discrete RBMs by describing explicit submodels (Theorem 15) and provides results on their maximal approximation errors and universal approximation properties (Theorem 16). Section 6 treats the dimension of discrete RBM models (Proposition 17 and Theorem 19). Section 7 contains an algebraic-combinatorial discussion of tropical discrete RBM models (Theorem 21) with consequences for their dimension collected in Propositions 24, 25, and 26.

2. Preliminaries

2.1 Independence models

Consider a system of $n < \infty$ random variables $X_1, \dots, X_n$. Assume that X_i takes states x_i in a finite set $\mathcal{X}_i = \{0, 1, \dots, r_i - 1\}$ for all $i \in \{1, \dots, n\} =: [n]$. The state space of this system is $\mathcal{X} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$. We write $x_\lambda = (x_i)_{i \in \lambda}$ for a joint state of the variables with index $i \in \lambda$ for any $\lambda \subseteq [n]$, and $x = (x_1, \dots, x_n)$ for a joint state of all variables. We denote by $\Delta(\mathcal{X})$ the set of all probability distributions on $\mathcal{X}$. We write $\langle a, b \rangle$ for the inner product $a^\top b$.

The *independence model* of the variables $X_1, \dots, X_n$ is the set of product distributions $p(x) = \prod_{i \in [n]} p_i(x_i)$ for all $x \in \mathcal{X}$, where p_i is a probability distribution with state space $\mathcal{X}_i$ for all $i \in [n]$. This model is the closure $\overline{\mathcal{E}_\mathcal{X}}$ (in the Euclidean topology) of the exponential family

$$\mathcal{E}_\mathcal{X} := \left\{ \frac{1}{Z(\theta)} \exp(\langle \theta, A^{(\mathcal{X})} \rangle) \colon \theta \in \mathbb{R}^{d_\mathcal{X}} \right\}, \tag{1}$$

where $A^{(\mathcal{X})} \in \mathbb{R}^{d_\mathcal{X} \times \mathcal{X}}$ is a matrix of sufficient statistics; with rows equal to the indicator functions $\mathbb{1}_\mathcal{X}$ and $\mathbb{1}_{\{x\colon\, x_i = y_i\}}$ for all $y_i \in \mathcal{X}_i \setminus \{0\}$ for all $i \in [n]$. The partition function $Z(\theta)$ normalizes the distributions. The convex support of $\mathcal{E}_\mathcal{X}$ is the convex hull $Q_\mathcal{X} := \operatorname{conv}(\{A^{(\mathcal{X})}_x\}_{x \in \mathcal{X}})$ of the columns of $A^{(\mathcal{X})}$, which is a Cartesian product of simplices with $Q_\mathcal{X} \cong \Delta(\mathcal{X}_1) \times \cdots \times \Delta(\mathcal{X}_n)$.

Example 1 The sufficient statistics of the independence models $\mathcal{E}_\mathcal{X}$ and $\mathcal{E}_{\mathcal{X}'}$ with state spaces $\mathcal{X} = \{0,1\}^3$ and $\mathcal{X}' = \{0,1,2\} \times \{0,1\}$ are, with rows labeled by indicator functions,

$$A^{(\mathcal{X})} = \begin{array}{c l} \begin{array}{cccccccc} \left[\begin{smallmatrix}1\\1\\1\end{smallmatrix}\right] & \left[\begin{smallmatrix}1\\1\\0\end{smallmatrix}\right] & \left[\begin{smallmatrix}1\\0\\1\end{smallmatrix}\right] & \left[\begin{smallmatrix}1\\0\\0\end{smallmatrix}\right] & \left[\begin{smallmatrix}0\\1\\1\end{smallmatrix}\right] & \left[\begin{smallmatrix}0\\1\\0\end{smallmatrix}\right] & \left[\begin{smallmatrix}0\\0\\1\end{smallmatrix}\right] & \left[\begin{smallmatrix}0\\0\\0\end{smallmatrix}\right] \end{array} & \\ \left(\begin{array}{cccccccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ \hline 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{array}\right) & \begin{array}{l} \\ x_3 = 1 \\ x_2 = 1 \\ x_1 = 1 \end{array} \end{array} \qquad A^{(\mathcal{X}')} = \begin{array}{c l} \begin{array}{cccccc} \left[\begin{smallmatrix}1\\2\end{smallmatrix}\right] & \left[\begin{smallmatrix}1\\1\end{smallmatrix}\right] & \left[\begin{smallmatrix}1\\0\end{smallmatrix}\right] & \left[\begin{smallmatrix}0\\2\end{smallmatrix}\right] & \left[\begin{smallmatrix}0\\1\end{smallmatrix}\right] & \left[\begin{smallmatrix}0\\0\end{smallmatrix}\right] \end{array} & \\ \left(\begin{array}{ccc|ccc} 1 & 1 & 1 & 1 & 1 & 1 \\ \hline 1 & 1 & 1 & 0 & 0 & 0 \\ \hline 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \end{array}\right) & \begin{array}{l} \\ x_2 = 1 \\ x_1 = 2 \\ x_1 = 1 \end{array} \end{array}.$$

In the first case the convex support is a cube and in the second it is a prism. Both convex supports are three-dimensional polytopes, but the prism has fewer vertices and is more similar to a simplex, meaning that its vertex set is affinely more independent than that of the cube. See Figure 2.

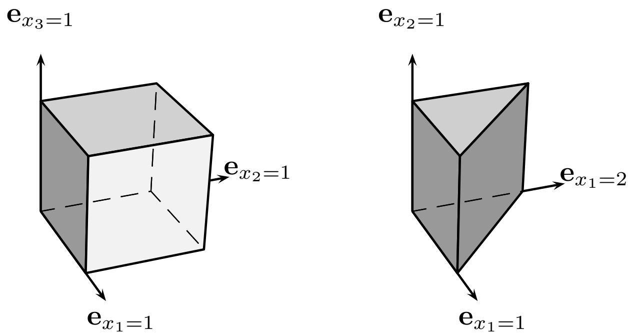

Figure 2: The convex support of the independence model of three binary variables (left) and of a binary-ternary pair of variables (right) discussed in Example 1.

2.2 Naïve Bayes models

Let $k \in \mathbb{N}$. The *k-mixture* of the independence model, or *naïve Bayes model* with k hidden classes, with visible variables $X_1, \dots, X_n$ is the set of all probability distributions expressible as convex combinations of k points in $\mathcal{E}_{\mathcal{X}}$:

$$\mathcal{M}_{\mathcal{X},k} := \Big\{ \sum_{i \in [k]} \lambda_i p^{(i)} \colon p^{(i)} \in \mathcal{E}_{\mathcal{X}},\ \lambda_i \geq 0, \text{ for all } i \in [k], \text{ and } \sum_{i \in [k]} \lambda_i = 1 \Big\}. \tag{2}$$

We write $\mathcal{M}_{n,k}$ for the k-mixture of the independence model of n binary variables. The dimensions of mixtures of binary independence models are known:

Theorem 2 (Catalisano et al. (2011)) *The mixtures of binary independence models $\mathcal{M}_{n,k}$ have the dimension expected from counting parameters,* $\min\{nk + (k-1), 2^n - 1\}$*, except for $\mathcal{M}_{4,3}$, which has dimension* 13 *instead of* 14*.*

Let $\mathfrak{A}_{\mathcal{X}}(d)$ denote the maximal cardinality of a subset $\mathcal{X}' \subseteq \mathcal{X}$ of minimum Hamming distance at least d, i.e., the maximal cardinality of a subset $\mathcal{X}' \subseteq \mathcal{X}$ with $d_H(x,y) \geq d$ for all distinct points $x, y \in \mathcal{X}'$, where $d_H(x,y) := |\{i \in [n] \colon x_i \neq y_i\}|$ denotes the Hamming distance between x and y. The function $\mathfrak{A}_{\mathcal{X}}$ is familiar in coding theory. The k-mixtures of independence models are universal approximators when k is large enough. This can be made precise in terms of $\mathfrak{A}_{\mathcal{X}}(2)$:

Theorem 3 (Montúfar (2013)) *The mixture model $\mathcal{M}_{\mathcal{X},k}$ can approximate any probability distribution on $\mathcal{X}$ arbitrarily well if $k \geq |\mathcal{X}|/\max_{i \in [n]} |\mathcal{X}_i|$ and only if $k \geq \mathfrak{A}_{\mathcal{X}}(2)$.*

By results from (Gilbert, 1952; Varshamov, 1957), when q is a power of a prime number and $\mathcal{X} = \{0, 1, \dots, q-1\}^n$, then $\mathfrak{A}_{\mathcal{X}} = q^{n-1}$. In these cases the previous theorem shows that $\mathcal{M}_{\mathcal{X},k}$ is a universal approximator of distributions on $\mathcal{X}$ if and only if $k \geq q^{n-1}$. In particular, the smallest naïve Bayes model universal approximator of distributions on $\{0,1\}^n$ has $2^{n-1}(n+1) - 1$ parameters.

Some of the distributions not representable by a given naïve Bayes model can be characterized in terms of their modes. A state $x \in \mathcal{X}$ is a *mode* of a distribution $p \in \Delta(\mathcal{X})$ if $p(x) > p(y)$ for all y with $d_H(x,y) = 1$ and it is a *strong mode* if $p(x) > \sum_{y \colon d_H(x,y)=1} p(y)$.

Lemma 4 (Montúfar and Morton (2014)) *If a mixture of product distributions $p = \sum_i \lambda_i p^{(i)}$ has strong modes $\mathcal{C} \subseteq \mathcal{X}$, then there is a mixture component $p^{(i)}$ with mode x for each $x \in \mathcal{C}$.*

2.3 Binary restricted Boltzmann machines

The binary RBM model with n visible and m hidden units, denoted $\mathrm{RBM}_{n,m}$, is the set of distributions on $\{0,1\}^n$ of the form

$$p(x) = \frac{1}{Z(W,B,C)} \sum_{h\in\{0,1\}^m} \exp(h^\top Wx + B^\top x + C^\top h) \quad \text{for all } x \in \{0,1\}^n, \tag{3}$$

where x denotes states of the visible units, h denotes states of the hidden units, $W = (W_{ji})_{ji} \in \mathbb{R}^{m\times n}$ is a matrix of interaction weights, $B \in \mathbb{R}^n$ and $C \in \mathbb{R}^m$ are vectors of bias weights, and $Z(W,B,C) = \sum_{x\in\{0,1\}^n} \sum_{h\in\{0,1\}^m} \exp(h^\top Wx + B^\top x + C^\top h)$ is the normalizing partition function.

It is known that these models have the expected dimension for many choices of n and m:

Theorem 5 (Cueto et al. (2010)) *The dimension of the model $\mathrm{RBM}_{n,m}$ is equal to $nm + n + m$ when $m + 1 \leq 2^{n-\lceil \log_2(n+1)\rceil}$ and it is equal to $2^n - 1$ when $m \geq 2^{n-\lfloor \log_2(n+1)\rfloor}$.*

It is also known that with enough hidden units, binary RBMs are universal approximators:

Theorem 6 (Montúfar and Ay (2011)) *The model $\mathrm{RBM}_{n,m}$ can approximate any distribution on $\{0,1\}^n$ arbitrarily well whenever $m \geq 2^{n-1} - 1$.*

A previous result by Le Roux and Bengio (2008, Theorem 2) shows that $\mathrm{RBM}_{n,m}$ is a universal approximator whenever $m \geq 2^n+1$. It is not known whether the bounds from Theorem 6 are always tight, but they show that for any given n, the smallest RBM universal approximator of distributions on $\{0,1\}^n$ has at most $2^{n-1}(n+1) - 1$ parameters and hence not more than the smallest naïve Bayes model universal approximator (Theorem 3).

3. Discrete restricted Boltzmann machines

Let $\mathcal{X}_i = \{0,1,\dots,r_i-1\}$ for all $i \in [n]$ and $\mathcal{Y}_j = \{0,1,\dots,s_j-1\}$ for all $j \in [m]$. The graphical model with full bipartite interactions $\{\{i,j\} \colon i \in [n], j \in [m]\}$ on $\mathcal{X} \times \mathcal{Y}$ is the exponential family

$$\mathcal{E}_{\mathcal{X},\mathcal{Y}} := \left\{ \frac{1}{Z(\theta)} \exp(\langle \theta, A^{(\mathcal{X},\mathcal{Y})}\rangle) \colon \theta \in \mathbb{R}^{d_{\mathcal{X}} d_{\mathcal{Y}}} \right\}, \tag{4}$$

with sufficient statistics matrix equal to the Kronecker product $A^{(\mathcal{X},\mathcal{Y})} = A^{(\mathcal{X})} \otimes A^{(\mathcal{Y})}$ of the sufficient statistics matrices $A^{(\mathcal{X})}$ and $A^{(\mathcal{Y})}$ of the independence models $\mathcal{E}_{\mathcal{X}}$ and $\mathcal{E}_{\mathcal{Y}}$. The matrix $A^{(\mathcal{X},\mathcal{Y})}$ has $d_{\mathcal{X}} d_{\mathcal{Y}} = \left(\sum_{i\in[n]}(|\mathcal{X}_i| - 1) + 1\right) \left(\sum_{j\in[m]}(|\mathcal{Y}_i| - 1) + 1\right)$ linearly independent rows and $|\mathcal{X} \times \mathcal{Y}|$ columns, each column corresponding to a joint state (x,y) of all variables. Disregarding the entry of θ that is multiplied with the constant row of $A^{(\mathcal{X},\mathcal{Y})}$, which cancels out with the normalization function $Z(\theta)$, this parametrization of $\mathcal{E}_{\mathcal{X},\mathcal{Y}}$ is one-to-one. In particular, this model has dimension $\dim(\mathcal{E}_{\mathcal{X},\mathcal{Y}}) = d_{\mathcal{X}} d_{\mathcal{Y}} - 1$.

The discrete RBM model $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$ is the following set of marginal distributions:

$$\mathrm{RBM}_{\mathcal{X},\mathcal{Y}} := \Big\{ q(x) = \sum_{y\in\mathcal{Y}} p(x,y) \text{ for all } x \in \mathcal{X} \colon p \in \mathcal{E}_{\mathcal{X},\mathcal{Y}} \Big\}. \tag{5}$$

In the case of one single hidden unit, this model is the naïve Bayes model on $\mathcal{X}$ with $|\mathcal{Y}_1|$ hidden classes. When all units are binary, $\mathcal{X} = \{0,1\}^n$ and $\mathcal{Y} = \{0,1\}^m$, this model is $\mathrm{RBM}_{n,m}$. Note that the exponent in eq. (3) can be written as $(h^\top W x + B^\top x + C^\top h) = \langle \theta, A^{(\mathcal{X},\mathcal{Y})}_{(x,h)} \rangle$, taking for θ the column-by-column vectorization of the matrix $\left(\begin{smallmatrix} 0 & B^\top \\ C & W \end{smallmatrix}\right)$.

Conditional distributions

The conditional distributions of discrete RBMs can be described in the following way. Consider a vector $\theta \in \mathbb{R}^{d_\mathcal{X} d_\mathcal{Y}}$ parametrizing $\mathcal{E}_{\mathcal{X},\mathcal{Y}}$, and the matrix $\Theta \in \mathbb{R}^{d_\mathcal{Y} \times d_\mathcal{X}}$ with column-by-column vectorization equal to θ. A lemma by Roth (1934) shows that $\theta^\top (A^{(\mathcal{X})} \otimes A^{(\mathcal{Y})})_{(x,y)} = (A^{(\mathcal{X})}_x)^\top \Theta^\top A^{(\mathcal{Y})}_y$ for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and hence

$$\Big\langle \theta, A^{(\mathcal{X},\mathcal{Y})}_{(x,y)} \Big\rangle = \Big\langle \Theta A^{(\mathcal{X})}_x, A^{(\mathcal{Y})}_y \Big\rangle = \Big\langle \Theta^\top A^{(\mathcal{Y})}_y, A^{(\mathcal{X})}_x \Big\rangle \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}. \tag{6}$$

The inner product in eq. (6) describes following probability distributions:

$$p_\theta(\cdot,\cdot) = \frac{1}{Z(\theta)} \exp\big(\langle \theta, A^{(\mathcal{X},\mathcal{Y})} \rangle\big), \tag{7}$$

$$p_\theta(\cdot|x) = \frac{1}{Z\big(\Theta A^{(\mathcal{X})}_x\big)} \exp\big(\langle \Theta A^{(\mathcal{X})}_x, A^{(\mathcal{Y})} \rangle\big), \text{ and} \tag{8}$$

$$p_\theta(\cdot|y) = \frac{1}{Z\big(\Theta^\top A^{(\mathcal{Y})}_y\big)} \exp\big(\langle \Theta^\top A^{(\mathcal{Y})}_y, A^{(\mathcal{X})} \rangle\big). \tag{9}$$

Geometrically, $\Theta A^{(\mathcal{X})}$ is a linear projection of the columns of the sufficient statistics matrix $A^{(\mathcal{X})}$ into the parameter space of $\mathcal{E}_\mathcal{Y}$, and similarly, $\Theta^\top A^{(\mathcal{Y})}$ is a linear projection of the columns of $A^{(\mathcal{Y})}$ into the parameter space of $\mathcal{E}_\mathcal{X}$.

Polynomial parametrization

Discrete RBMs can be parametrized not only in the exponential way discussed above, but also by simple polynomials. The exponential family $\mathcal{E}_{\mathcal{X},\mathcal{Y}}$ can be parametrized by square free monomials:

$$p(v,h) = \frac{1}{Z} \prod_{\substack{\{j,i\} \in [m]\times[n], \\ (y'_j, x'_i) \in \mathcal{Y}_j \times \mathcal{X}_i}} \big(\gamma_{\{j,i\},(y'_j,x'_i)}\big)^{\delta_{y'_j}(h_j)\delta_{x'_i}(v_i)} \quad \text{for all } (v,h) \in \mathcal{Y} \times \mathcal{X}, \tag{10}$$

where $\gamma_{\{j,i\},(y'_j,x'_i)}$ are positive reals. The probability distributions in $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$ can be written as

$$p(v) = \frac{1}{Z} \prod_{j\in[m]} \Big(\sum_{h_j \in \mathcal{Y}_j} \gamma_{\{j,1\},(h_j,v_1)} \cdots \gamma_{\{j,n\},(h_j,v_n)} \Big) \quad \text{for all } v \in \mathcal{X}. \tag{11}$$

The parameters $\gamma_{\{j,i\},(y'_j,x'_i)}$ correspond to $\exp(\theta_{\{j,i\},(y'_j,x'_i)})$ in the parametrization given in eq. (4).

Products of mixtures and mixtures of products

In the following we describe discrete RBMs from two complementary perspectives: (i) as products of experts, where each expert is a mixture of products, and (ii) as restricted mixtures of product distributions. The renormalized entry-wise (Hadamard) product of two probability distributions p and q on $\mathcal{X}$ is defined as $p\circ q := (p(x)q(x))_{x\in\mathcal{X}}/\sum_{y\in\mathcal{X}}p(y)q(y)$. Here we assume that p and q have overlapping supports, such that the definition makes sense.

Proposition 7 *The model* $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$ *is a Hadamard product of mixtures of product distributions:*

$$\mathrm{RBM}_{\mathcal{X},\mathcal{Y}} = \mathcal{M}_{\mathcal{X},|\mathcal{Y}_1|}\circ\cdots\circ\mathcal{M}_{\mathcal{X},|\mathcal{Y}_m|}\,.$$

Proof The statement can be seen directly by considering the parametrization from eq. (11). To make this explicit, one can use a *homogeneous* version of the matrix $A^{(\mathcal{X},\mathcal{Y})}$ which we denote by A and which defines the same model. Each row of A is indexed by an edge $\{i,j\}$ of the bipartite graph and a joint state (x_i,h_j) of the visible and hidden units connected by this edge. Such a row has a one in any column when these states agree with the global state, and zero otherwise. For any $j\in[m]$ let $A_{j,:}$ denote the matrix containing the rows of A with indices $(\{i,j\},(x_i,h_j))$ for all $x_i\in\mathcal{X}_i$ for all $i\in[n]$ for all $h_j\in\mathcal{Y}_j$, and let $A(x,h)$ denote the (x,h)-column of A. We have

$$\begin{aligned}
p(x) =&\frac{1}{Z}\sum_h \exp(\langle\theta, A(x,h)\rangle)\\
=&\frac{1}{Z}\sum_h \exp(\langle\theta_{1,:}, A_{1,:}(x,h)\rangle)\exp(\langle\theta_{2,:}, A_{2,:}(x,h)\rangle)\cdots\exp(\langle\theta_{m,:}, A_{m,:}(x,h)\rangle)\\
=&\frac{1}{Z}\Big(\sum_{h_1}\exp(\langle\theta_{1,:}, A_{1,:}(x,h_1)\rangle)\Big)\cdots\Big(\sum_{h_m}\exp(\langle\theta_{m,:}, A_{m,:}(x,h_m)\rangle)\Big)\\
=&\frac{1}{Z}(Z_1p^{(1)}(x))\cdots(Z_mp^{(m)}(x)) = \frac{1}{Z'}p^{(1)}(x)\cdots p^{(m)}(x),
\end{aligned}$$

where $p^{(j)}\in\mathcal{M}_{\mathcal{X},|\mathcal{Y}_j|}$ and $Z_j=\sum_{x\in\mathcal{X}}\sum_{h_j\in\mathcal{Y}_j}\exp(\langle\theta_{j,:},A_{j,:}(x,h_j)\rangle)$ for all $j\in[m]$. Since the vectors $\theta_{j,:}$ can be chosen arbitrarily, the factors $p^{(j)}$ can be made arbitrary within $\mathcal{M}_{\mathcal{X},|\mathcal{Y}_j|}$. ■

Of course, every distribution in $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$ is a mixture distribution $p(x)=\sum_{h\in\mathcal{Y}}p(x|h)q(h)$. The mixture weights are given by the marginals $q(h)$ on $\mathcal{Y}$ of distributions from $\mathcal{E}_{\mathcal{X},\mathcal{Y}}$, and the mixture components can be described as follows.

Proposition 8 *The set of conditional distributions* $p(\cdot|h)$, $h\in\mathcal{Y}$ *of a distribution in* $\mathcal{E}_{\mathcal{X},\mathcal{Y}}$ *is the set of product distributions in* $\mathcal{E}_{\mathcal{X}}$ *with parameters* $\theta_h=\Theta^\top A^{(\mathcal{Y})}_h$, $h\in\mathcal{Y}$ *equal to a linear projection of the vertices* $\{A^{(\mathcal{Y})}_h\colon h\in\mathcal{Y}\}$ *of the Cartesian product of simplices* $Q_{\mathcal{Y}}\cong\Delta(\mathcal{Y}_1)\times\cdots\times\Delta(\mathcal{Y}_m)$.

Proof This is by eq. (6). ■

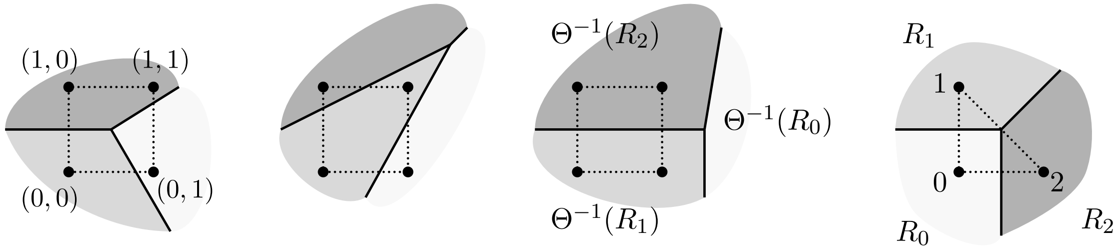

Figure 3: Three slicings of a square by the normal fan of a triangle with maximal cones R_0, R_1, and R_2, corresponding to three possible inference functions of $\mathrm{RBM}_{\{0,1\}^2,\{0,1,2\}}$.

4. Products of simplices and their normal fans

Binary RBMs have been analyzed by considering each of the m hidden units as defining a hyperplane H_j slicing the n-cube into two regions. To generalize the results provided by this analysis, in this section we replace the n-cube with a general product of simplices $Q_{\mathcal{X}}$, and replace the two regions defined by the hyperplane H_j by the $|\mathcal{Y}_j|$ regions defined by the maximal cones of the normal fan of the simplex $\Delta(\mathcal{Y}_j)$.

Subdivisions of independence models

The *normal cone* of a polytope $Q \subset \mathbb{R}^d$ at a point $x \in Q$ is the set of all vectors $v \in \mathbb{R}^d$ with $\langle v, (x-y)\rangle \geq 0$ for all $y \in Q$. We denote by R_x the normal cone of the product of simplices $Q_{\mathcal{X}} = \mathrm{conv}\{A_x^{(\mathcal{X})}\}_{x\in\mathcal{X}}$ at the vertex $A_x^{(\mathcal{X})}$. The normal fan $\mathcal{F}_{\mathcal{X}}$ is the set of all normal cones of $Q_{\mathcal{X}}$. The product distributions $p_\theta = \frac{1}{Z(\theta)}\exp(\langle\theta, A^{(\mathcal{X})}\rangle) \in \mathcal{E}_{\mathcal{X}}$ strictly maximized at $x \in \mathcal{X}$, with $p_\theta(x) > p_\theta(y)$ for all $y \in \mathcal{X}\setminus\{x\}$, are those with parameter vector θ in the relative interior of R_x. Hence the normal fan $\mathcal{F}_{\mathcal{X}}$ partitions the parameter space of the independence model into regions of distributions with maxima at different inputs.

Inference functions and slicings

For any choice of parameters of the model $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$, there is an *inference function* $\pi\colon \mathcal{X}\to\mathcal{Y}$, (or more generally $\pi\colon \mathcal{X}\to 2^{\mathcal{Y}}$), which computes the most likely hidden state given a visible state. These functions are not necessarily injective nor surjective. For a visible state x, the conditional distribution on the hidden states is a product distribution $p(y|X=x) = \frac{1}{Z}\exp(\langle\Theta A_x^{(\mathcal{X})}, A_y^{(\mathcal{Y})}\rangle)$ which is maximized at the state y for which $\Theta A_x^{(\mathcal{X})} \in R_y$. The preimages of the cones R_y by the map Θ partition the input space $\mathbb{R}^{d_\mathcal{X}}$ and are called *inference regions*. See Figure 3 and Example 10.

Definition 9 A $\mathcal{Y}$-*slicing* of a finite set $\mathcal{Z}\subset\mathbb{R}^{d_\mathcal{X}}$ is a partition of $\mathcal{Z}$ into the preimages of the cones R_y, $y\in\mathcal{Y}$ by a linear map $\Theta\colon \mathbb{R}^{d_\mathcal{X}}\to\mathbb{R}^{d_\mathcal{Y}}$. We assume that Θ is generic, such that it maps each element of $\mathcal{Z}$ into the interior of some R_y.

For example, when $\mathcal{Y}=\{0,1\}$, the fan $\mathcal{F}_\mathcal{Y}$ consists of a hyperplane and the two closed halfspaces defined by that hyperplane. A $\mathcal{Y}$-slicing is in this case a standard slicing by a hyperplane.

Example 10 Let $\mathcal{X} = \{0,1,2\} \times \{0,1\}$ and $\mathcal{Y} = \{0,1\}^4$. The maximal cones R_y, $y \in \mathcal{Y}$ of the normal fan of the 4-cube with vertices $\{0,1\}^4$ are the closed orthants of $\mathbb{R}^4$. The 6 vertices $\{A^{(\mathcal{X})}_x \colon x \in \mathcal{X}\}$ of the prism $\Delta(\{0,1,2\}) \times \Delta(\{0,1\})$ can be mapped into 6 distinct orthants of $\mathbb{R}^4$, each orthant with an even number of positive coordinates:

$$\underset{\Theta}{\begin{pmatrix} 3 & -2 & -2 & -2 \\ 1 & 2 & -2 & -2 \\ 1 & -2 & -2 & 2 \\ 1 & -2 & 2 & -2 \end{pmatrix}} \underset{A^{(\mathcal{X})}}{\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \end{pmatrix}} = \begin{pmatrix} -1 & -1 & 1 & 1 & 1 & 3 \\ 1 & 1 & 3 & -1 & -1 & 1 \\ -3 & 1 & -1 & -1 & 3 & 1 \\ 1 & -3 & -1 & 3 & -1 & 1 \end{pmatrix}. \qquad (12)$$

Even in the case of one single hidden unit the slicings can be complex, but the following simple type of slicing is always available.

Proposition 11 *Any slicing by $k-1$ parallel hyperplanes is a $\{1,2,\dots,k\}$-slicing.*

Proof We show that there is a line $\mathcal{L} = \{\lambda r - b \colon \lambda \in \mathbb{R}\}$, $r, b \in \mathbb{R}^k$ intersecting all cells of $\mathcal{F}_\mathcal{Y}$, $\mathcal{Y} = \{1,\dots,k\}$. We need to show that there is a choice of r and b such that for every $y \in \mathcal{Y}$ the set $I_y \subseteq \mathbb{R}$ of all λ with $\langle \lambda r - b, (\mathbf{e}_y - \mathbf{e}_z)\rangle > 0$ for all $z \in \mathcal{Y} \setminus \{y\}$ has a non-empty interior. Now, I_y is the set of λ with

$$\lambda(r_y - r_z) > b_y - b_z \qquad \text{for all } z \neq y. \qquad (13)$$

Choosing $b_1 < \cdots < b_k$ and $r_y = f(b_y)$, where f is a strictly increasing and strictly concave function, we get $I_1 = (-\infty, \frac{b_2 - b_1}{r_2 - r_1})$, $I_y = (\frac{b_y - b_{y-1}}{r_y - r_{y-1}}, \frac{b_{y+1} - b_y}{r_{y+1} - r_y})$ for $y = 2, 3, \dots, k-1$, and $I_k = (\frac{b_k - b_{k-1}}{r_k - r_{k-1}}, \infty)$. The lengths $\infty, l_2, \dots, l_{k-1}, \infty$ of the intervals $I_1, \dots, I_k$ can be adjusted arbitrarily by choosing suitable differences $r_{j+1} - r_j$ for all $j = 1, \dots, k-1$. ■

Strong modes

Recall the definition of strong modes given in page 4.

Lemma 12 *Let $\mathcal{C} \subseteq \mathcal{X}$ be a set of arrays which are pairwise different in at least two entries (a code of minimum distance two).*

- *If $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$ contains a probability distribution with strong modes $\mathcal{C}$, then there is a linear map Θ of $\{A^{(\mathcal{Y})}_y \colon y \in \mathcal{Y}\}$ into the $\mathcal{C}$-cells of $\mathcal{F}_\mathcal{X}$ (the cones R_x above the codewords $x \in \mathcal{C}$) sending at least one vertex into each cell.*
- *If there is a linear map Θ of $\{A^{(\mathcal{Y})}_y \colon y \in \mathcal{Y}\}$ into the $\mathcal{C}$-cells of $\mathcal{F}_\mathcal{X}$, with $\max_x\{\langle \Theta^\top A^{(\mathcal{Y})}_y, A^{(\mathcal{X})}_x\rangle\} = c$ for all $y \in \mathcal{Y}$, then $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$ contains a probability distribution with strong modes $\mathcal{C}$.*

Proof This is by Proposition 8 and Lemma 4. ■

A simple consequence of the previous lemma is that if the model $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$ is a universal approximator of distributions on $\mathcal{X}$, then necessarily the number of hidden states is at least as large as the maximum code of visible states of minimum distance two, $|\mathcal{Y}| \geq \mathfrak{A}_\mathcal{X}(2)$. Hence discrete RBMs may not be universal approximators even when their parameter count surpasses the dimension of the ambient probability simplex.

Example 13 Let $\mathcal{X} = \{0,1,2\}^n$ and $\mathcal{Y} = \{0,1,\dots,4\}^m$. In this case $\mathfrak{A}_{\mathcal{X}}(2) = 3^{n-1}$. If $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$ is a universal approximator with $n = 3$ and $n = 4$, then $m \geq 2$ and $m \geq 3$, respectively, although the smallest m for which $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$ has $3^n - 1$ parameters is $m = 1$ and $m = 2$, respectively.

Using Lemma 12 and the analysis of (Montúfar and Morton, 2014) gives the following.

Proposition 14 *If $4\lceil m/3 \rceil \leq n$, then $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$ contains distributions with 2^m strong modes.*

5. Representational power and approximation errors

In this section we describe submodels of discrete RBMs and use them to provide bounds on the model approximation errors depending on the number of units and their state spaces. Universal approximation results follow as special cases with vanishing approximation error.

Theorem 15 *The model $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$ can approximate the following arbitrarily well:*

- *Any mixture of $d_{\mathcal{Y}} = 1 + \sum_{j=1}^m (|\mathcal{Y}_j| - 1)$ product distributions with disjoint supports.*
- *When $d_{\mathcal{Y}} \geq (\prod_{i\in[k]} |\mathcal{X}_i|)/\max_{j\in[k]} |\mathcal{X}_j|$ for some $k \leq n$, any distribution from the model $\mathcal{P}$ of distributions with constant value on each block $\{x_1\} \times \cdots \times \{x_k\} \times \mathcal{X}_{k+1} \times \cdots \times \mathcal{X}_n$ for all $x_i \in \mathcal{X}_i$, for all $i \in [k]$.*
- *Any probability distribution with support contained in the union of $d_{\mathcal{Y}}$ sets of the form $\{x_1\} \times \cdots \times \{x_{k-1}\} \times \mathcal{X}_k \times \{x_{k+1}\} \times \cdots \times \{x_n\}$.*

Proof By Proposition 7 the model $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$ contains any Hadamard product $p^{(1)} \circ \cdots \circ p^{(m)}$ with mixtures of products as factors, $p^{(j)} \in \mathcal{M}_{\mathcal{X},|\mathcal{Y}_j|}$ for all $j \in [m]$. In particular, it contains $p = p^{(0)} \circ (\mathbb{1} + \tilde{\lambda}_1 \tilde{p}^{(1)}) \circ \cdots \circ (\mathbb{1} + \tilde{\lambda}_m \tilde{p}^{(m)})$, where $p^{(0)} \in \mathcal{E}_{\mathcal{X}}$, $\tilde{p}^{(j)} \in \mathcal{M}_{\mathcal{X},|\mathcal{Y}_j|-1}$, and $\tilde{\lambda}_j \in \mathbb{R}_+$. Choosing the factors $\tilde{p}^{(j)}$ with pairwise disjoint supports shows that $p = \sum_{j=0}^m \lambda_j p^{(j)}$, whereby $p^{(0)}$ can be any product distribution and $p^{(j)}$ can be any distribution from $\mathcal{M}_{\mathcal{X},|\mathcal{Y}_j|-1}$ for all $j \in [m]$, as long as $\mathrm{supp}(p^{(j)}) \cap \mathrm{supp}(p^{(j')})$ for all $j \neq j'$. This proves the first item.

For the second item: Any point in the set $\mathcal{P}$ is a mixture of uniform distributions supported on the disjoint blocks $\{x_1\} \times \cdots \times \{x_k\} \times \mathcal{X}_{k+1} \times \cdots \times \mathcal{X}_n$ for all $(x_1,\dots,x_k) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_k$. Each of these uniform distributions is a product distribution, since it factorizes as $p_{x_1,\dots,x_k} = \prod_{i\in[k]} \delta_{x_i} \prod_{i\in[n]\setminus[k]} u_i$, where u_i denotes the uniform distribution on $\mathcal{X}_i$. For any $j \in [k]$ any mixture $\sum_{x_j \in \mathcal{X}_j} \lambda_{x_j} p_{x_1,\dots,x_k}$ is also a product distribution, since it factorizes as

$$\Big(\sum_{x_j \in \mathcal{X}_j} \lambda_{x_j} \delta_{x_j}\Big) \prod_{i\in[k]\setminus\{j\}} \delta_{x_i} \prod_{i\in[n]\setminus[k]} u_i. \tag{14}$$

Hence any distribution from the set $\mathcal{P}$ is a mixture of $(\prod_{i\in[k]} |\mathcal{X}_i|)/\max_{j\in[k]} |\mathcal{X}_j|$ product distributions with disjoint supports. The claim now follows from the first item.

For the third item: The model $\overline{\mathcal{E}_{\mathcal{X}}}$ contains any distribution with support of the form $\{x_1\} \times \cdots \times \{x_{k-1}\} \times \mathcal{X}_k \times \{x_{k+1}\} \times \cdots \times \{x_n\}$. Hence, by the first item, the RBM model can approximate any distribution arbitrarily well whose support can be covered by $d_{\mathcal{Y}}$ sets of that form. ■

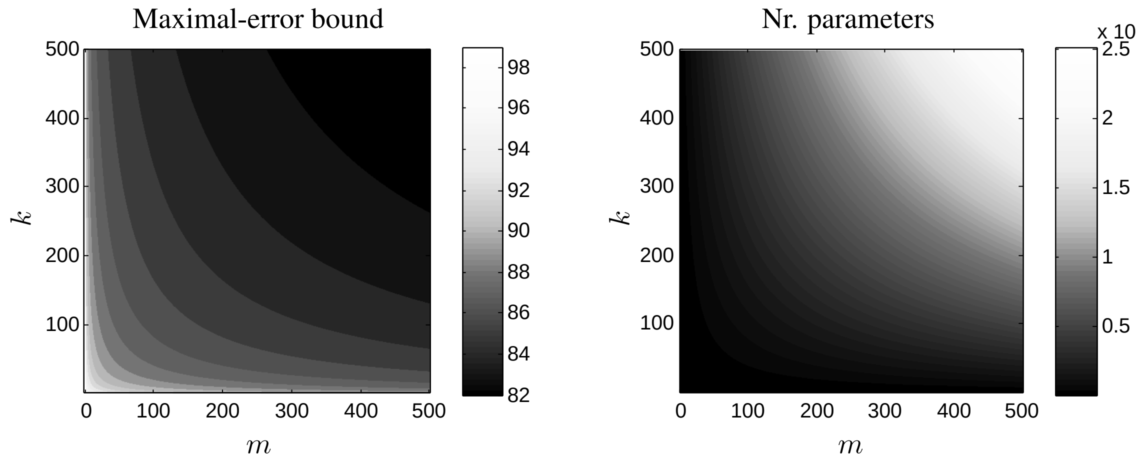

Figure 4: Illustration of Theorem 16. The left panel shows a heat map of the upper bound on the Kullback-Leibler approximation errors of discrete RBMs with 100 visible binary units and the right panel shows a map of the total number of model parameters, both depending on the number of hidden units m and their possible states $k = |\mathcal{Y}_j|$ for all $j \in [m]$.

We now analyze the RBM model approximation errors. Let p and q be two probability distributions on $\mathcal{X}$. The Kullback-Leibler divergence from p to q is defined as $D(p\|q) := \sum_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$ when $\operatorname{supp}(p) \subseteq \operatorname{supp}(q)$ and $D(p\|q) := \infty$ otherwise. The divergence from p to a model $\mathcal{M} \subseteq \Delta(\mathcal{X})$ is defined as $D(p\|\mathcal{M}) := \inf_{q\in\mathcal{M}} D(p\|q)$ and the maximal approximation error of $\mathcal{M}$ is $\sup_{p\in\Delta(\mathcal{X})} D(p\|\mathcal{M})$.

It is known that the maximal approximation error of the independence model $\mathcal{E}_\mathcal{X}$ satisfies $\sup_{p\in\Delta(\mathcal{X})} D(p\|\mathcal{E}_\mathcal{X}) \leq |\mathcal{X}|/\max_{i\in[n]}|\mathcal{X}_i|$, with equality when all units have the same number of states (see Ay and Knauf, 2006, Corollary 4.10).

Theorem 16 *If $\prod_{i\in[n]\setminus\Lambda}|\mathcal{X}_i| \leq 1 + \sum_{j\in[m]}(|\mathcal{Y}_j|-1) = d_\mathcal{Y}$ for some $\Lambda \subseteq [n]$, then the Kullback-Leibler divergence from any distribution p on $\mathcal{X}$ to the model $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$ is bounded by*

$$D(p\|\,\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}) \leq \log\frac{\prod_{i\in\Lambda}|\mathcal{X}_i|}{\max_{i\in\Lambda}|\mathcal{X}_i|}.$$

In particular, the model $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$ is a universal approximator whenever $d_\mathcal{Y} \geq |\mathcal{X}|/\max_{i\in[n]}|\mathcal{X}_i|$.

Proof The submodel $\mathcal{P}$ of $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$ described in the second item of Theorem 15 is a *partition model*. The maximal divergence from such a model is equal to the logarithm of the cardinality of the largest block with constant values (see Matúš and Ay, 2003). Thus $\max_p D(p\|\,\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}) \leq \max_p D(p\|\mathcal{P}) = \log\big((\prod_{i\in\Lambda}|\mathcal{X}_i|)/\max_{i\in\Lambda}|\mathcal{X}_i|\big)$, as was claimed. ■

Theorem 16 shows that, on a large scale, the maximal model approximation error of $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$ is smaller than that of the independence model $\mathcal{E}_\mathcal{X}$ by at least $\log(1+\sum_{j\in[m]}(|\mathcal{Y}_j|-1))$, or vanishes. The theorem is illustrated in Figure 4. The line $k = 2$ shows bounds on the approximation error of binary RBMs with m hidden units, previously treated in (Montúfar et al., 2011, Theorem 5.1), and the line $m = 1$ shows bounds for naïve Bayes models with k hidden classes.

6. Dimension

In this section we study the dimension of the model $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$. One reason RBMs are attractive is that they have a large learning capacity, e.g. may be built with millions of parameters. Dimension calculations show whether those parameters are wasted, or translate into higher-dimensional spaces of representable distributions. Our analysis builds on previous work by Cueto, Morton, and Sturmfels (2010), where binary RBMs are treated. The idea is to bound the dimension from below by the dimension of a related max-plus model, called the tropical RBM model (Pachter and Sturmfels, 2004), and from above by the dimension expected from counting parameters.

The dimension of a discrete RBM model can be bounded from above not only by its expected dimension, but also by a function of the dimension of its Hadamard factors:

Proposition 17 *The dimension of* $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$ *is bounded as*

$$\dim(\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}) \leq \dim(\mathcal{M}_{\mathcal{X},|\mathcal{Y}_i|}) + \sum_{j\in[m]\setminus\{i\}} \dim(\mathcal{M}_{\mathcal{X},|\mathcal{Y}_j|-1}) + (m-1) \quad \textit{for all } i \in [m]. \tag{15}$$

Proof Let u denote the uniform distribution. Note that $\mathcal{E}_{\mathcal{X}} \circ \mathcal{E}_{\mathcal{X}} = \mathcal{E}_{\mathcal{X}}$ and also $\mathcal{E}_{\mathcal{X}} \circ \mathcal{M}_{\mathcal{X},k} = \mathcal{M}_{\mathcal{X},k}$. This observation, together with Proposition 7, shows that the RBM model can be factorized as

$$\mathrm{RBM}_{\mathcal{X},\mathcal{Y}} = (\mathcal{M}_{\mathcal{X},|\mathcal{Y}_1|}) \circ (\lambda_1 u + (1-\lambda_1)\mathcal{M}_{\mathcal{X},|\mathcal{Y}_1|}) \circ \cdots \circ (\lambda_m u + (1-\lambda_m)\mathcal{M}_{\mathcal{X},|\mathcal{Y}_m|-1}),$$

from which the claim follows. ■

By the previous proposition, the model $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$ can have the expected dimension only if (i) the right hand side of eq. (15) equals $|\mathcal{X}|-1$, or (ii) each mixture model $\mathcal{M}_{\mathcal{X},k}$ has the expected dimension for all $k \leq \max_{j\in[m]} |\mathcal{Y}_j|$. Sometimes none of both conditions is satisfied and the models 'waste' parameters:

Example 18 The k-mixture of the independence model on $\mathcal{X}_1 \times \mathcal{X}_2$ is a subset of the set of $|\mathcal{X}_1| \times |\mathcal{X}_2|$ matrices with non-negative entries and rank at most k. It is known that the set of $M \times N$ matrices of rank at most k has dimension $k(M+N-k)$ for all $1 \leq k < \min\{M,N\}$. Hence the model $\mathcal{M}_{\mathcal{X}_1\times\mathcal{X}_2,k}$ has dimension smaller than its parameter count whenever $1 < k < \min\{|\mathcal{X}_1|,|\mathcal{X}_2|\}$. By Proposition 17 if $(\sum_{j\in[m]}(|\mathcal{Y}_j|-1)+1)(|\mathcal{X}_1|+|\mathcal{X}_2|-1) \leq |\mathcal{X}_1\times\mathcal{X}_2|$ and $1 < |\mathcal{Y}_j| \leq \min\{|\mathcal{X}_1|,|\mathcal{X}_2|\}$ for some $j \in [m]$, then $\mathrm{RBM}_{\mathcal{X}_1\times\mathcal{X}_2,\mathcal{Y}}$ does not have the expected dimension.

The next theorem indicates choices of $\mathcal{X}$ and $\mathcal{Y}$ for which the model $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$ has the expected dimension. Given a sufficient statistics matrix $A^{(\mathcal{X})}$, we say that a set $\mathcal{Z} \subseteq \mathcal{X}$ has full rank when the matrix with columns $\{A^{(\mathcal{X})}_x : x \in \mathcal{Z}\}$ has full rank.

Theorem 19 *When* $\mathcal{X}$ *contains* m *disjoint Hamming balls of radii* $2(|\mathcal{Y}_j|-1)-1$, $j \in [m]$ *and the subset of* $\mathcal{X}$ *not intersected by these balls has full rank, then the model* $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$ *has dimension equal to the number of model parameters,*

$$\dim(\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}) = (1 + \sum_{i\in[n]}(|\mathcal{X}_i|-1))(1+\sum_{j\in[m]}(|\mathcal{Y}_j|-1)) - 1.$$

On the other hand, if m *Hamming balls of radius one cover* $\mathcal{X}$*, then*

$$\dim(\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}) = |\mathcal{X}| - 1.$$

In order to prove this theorem we will need two main tools: slicings by normal fans of simplices, described in Section 4, and the tropical RBM model, described in Section 7. The theorem will follow from the analysis contained in Section 7.

7. Tropical model

Definition 20 The tropical model $\mathrm{RBM}^{\text{tropical}}_{\mathcal{X},\mathcal{Y}}$ is the image of the tropical morphism

$$\mathbb{R}^{d_{\mathcal{X}}d_{\mathcal{Y}}} \ni \theta \;\mapsto\; \Phi(v;\theta) = \max\{\langle \theta, A^{(\mathcal{X},\mathcal{Y})}_{(v,h)}\rangle \colon h \in \mathcal{Y}\} \quad \text{for all } v \in \mathcal{X}, \tag{16}$$

which evaluates $\log(\frac{1}{Z(\theta)}\sum_{h\in\mathcal{Y}}\exp(\langle\theta, A^{(\mathcal{X},\mathcal{Y})}_{(v,h)}\rangle))$ for all $v \in \mathcal{X}$ for each θ within the max-plus algebra (addition becomes $a + b = \max\{a, b\}$) up to additive constants independent of v (i.e., disregarding the normalization factor $Z(\theta)$).

The idea behind this definition is that $\log(\exp(a) + \exp(b)) \approx \max\{a, b\}$ when a and b have different order of magnitude. The tropical model captures important properties of the original model. Of particular interest is following consequence of the Bieri-Groves theorem (see Draisma, 2008), which gives us a tool to estimate the dimension of $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$:

$$\dim(\mathrm{RBM}^{\text{tropical}}_{\mathcal{X},\mathcal{Y}}) \le \dim(\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}) \le \min\{\dim(\mathcal{E}_{\mathcal{X},\mathcal{Y}}), |\mathcal{X}| - 1\}. \tag{17}$$

The following Theorem 21 describes the regions of linearity of the map Φ. Each of these regions corresponds to a collection of $\mathcal{Y}_j$-slicings (see Definition 9) of the set $\{A^{(\mathcal{X})}_x \colon x \in \mathcal{X}\}$ for all $j \in [m]$. This result allows us to express the dimension of $\mathrm{RBM}^{\text{tropical}}_{\mathcal{X},\mathcal{Y}}$ as the maximum rank of a class of matrices defined by collections of slicings.

For each $j \in [m]$ let $C_j = \{C_{j,1}, \dots, C_{j,|\mathcal{Y}_j|}\}$ be a $\mathcal{Y}_j$-slicing of $\{A^{(\mathcal{X})}_x \colon x \in \mathcal{X}\}$ and let $A_{C_{j,k}}$ be the $|\mathcal{X}| \times d_{\mathcal{X}}$-matrix with x-th row equal to $(A^{(\mathcal{X})}_x)^\top$ when $x \in C_{j,k}$ and equal to a row of zeros otherwise. Let $A_{C_j} = (A_{C_{j,1}}|\cdots|A_{C_{j,|\mathcal{Y}_j|}}) \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}_j|d_{\mathcal{X}}}$ and $d = \sum_{j\in[m]} |\mathcal{Y}_j|d_{\mathcal{X}}$.

Theorem 21 *On each region of linearity, the tropical morphism* Φ *is the linear map* $\mathbb{R}^d \to \mathrm{RBM}^{\text{tropical}}_{\mathcal{X},\mathcal{Y}}$ *represented by the* $|\mathcal{X}| \times d$*-matrix*

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

modulo constant functions. In particular, $\dim(\mathrm{RBM}^{\text{tropical}}_{\mathcal{X},\mathcal{Y}}) + 1$ *is the maximum rank of* $\mathcal{A}$ *over all possible collections of slicings* $C_1, \dots, C_m$*.*

Proof Again use the homogeneous version of the matrix $A^{(\mathcal{X},\mathcal{Y})}$ as in the proof of Proposition 7; this will not affect the rank of $\mathcal{A}$. Let $\theta_{h_j} = (\theta_{\{j,i\},(h_j,x_i)})_{i\in[n],x_i\in\mathcal{X}_i}$ and let A_{h_j} denote the submatrix of $A^{(\mathcal{X},\mathcal{Y})}$ containing the rows with indices $\{\{j,i\},(h_j,x_i) \colon i \in [n], x_i \in \mathcal{X}_i\}$. For any given $v \in \mathcal{X}$ we have

$$\max\Big\{\langle\theta, A^{(\mathcal{X},\mathcal{Y})}_{(v,h)}\rangle \colon h \in \mathcal{Y}\Big\} = \sum_{j\in[m]} \max\Big\{\langle\theta_{h_j}, A_{h_j}(v, h_j)\rangle \colon h_j \in \mathcal{Y}_j\Big\},$$

from which the claim follows. ■

In the following we evaluate the maximum rank of the matrix $\mathcal{A}$ for various choices of $\mathcal{X}$ and $\mathcal{Y}$ by examining good slicings. We focus on slicings by parallel hyperplanes.

Lemma 22 *For any $x^* \in \mathcal{X}$ and $0 < k < n$ the affine hull of the set $\{A^{(\mathcal{X})}_x \colon d_H(x, x^*) = k\}$ has dimension $\sum_{i\in[n]}(|\mathcal{X}_i| - 1) - 1$.*

Proof Without loss of generality let $x^* = (0, \dots, 0)$. The set $\mathcal{Z}^k := \{A^{(\mathcal{X})}_x \colon d_H(x, x^*) = k\}$ is the intersection of $\{A^{(\mathcal{X})}_x \colon x \in \mathcal{X}\}$ with the hyperplane $H^k := \{z\colon \langle \mathbb{1}, z\rangle = k + 1\}$. Now note that the two vertices of an edge of $Q_\mathcal{X}$ either lie in the same hyperplane H^l, or in two adjacent parallel hyperplanes H^l and H^{l+1}, with $l \in \mathbb{N}$. Hence the hyperplane H^k does not slice any edges of $Q_\mathcal{X}$ and $\mathrm{conv}(\mathcal{Z}^k) = Q_\mathcal{X} \cap H^k$. The set $\mathcal{Z}^k$ is not contained in any proper face of $Q_\mathcal{X}$ and hence $\mathrm{conv}(\mathcal{Z}^k)$ intersects the interior of $Q_\mathcal{X}$. Thus $\dim(\mathrm{conv}(\mathcal{Z}^k)) = \dim(Q_\mathcal{X}) - 1$, as was claimed. ■

Lemma 22 implies the following.

Corollary 23 *Let $x \in \mathcal{X}$, and $2k - 3 \leq n$. There is a slicing $C_1 = \{C_{1,1}, \dots, C_{1,k}\}$ of $\mathcal{X}$ by $k - 1$ parallel hyperplanes such that $\cup_{l=1}^{k-1} C_{1,l} = B_x(2k - 3)$ is the Hamming ball of radius $2k - 3$ centered at x and the matrix $A_{C_1} = (A_{C_{1,1}}|\cdots|A_{C_{1,k-1}})$ has full rank.*

Recall that $\mathfrak{A}_\mathcal{X}(d)$ denotes the maximal cardinality of a subset of $\mathcal{X}$ of minimum Hamming distance at least d. When $\mathcal{X} = \{0, 1, \dots, q-1\}^n$ we write $\mathfrak{A}_q(n, d)$. Let $\mathfrak{K}_\mathcal{X}(d)$ denote the minimal cardinality of a subset of $\mathcal{X}$ with covering radius d.

Proposition 24 (Binary visible units) *Let $\mathcal{X} = \{0, 1\}^n$ and $|\mathcal{Y}_j| = s_j$ for all $j \in [m]$. If $\mathcal{X}$ contains m disjoint Hamming balls of radii $2s_j - 3$, $j \in [m]$ whose complement has full rank, then $\mathrm{RBM}^{\text{tropical}}_{\mathcal{X},\mathcal{Y}}$ has the expected dimension, $\min\{\sum_{j\in[m]}(s_j - 1)(n + 1) + n, 2^n - 1\}$.*

In particular, if $\mathcal{X} = \{0, 1\}^n$ and $\mathcal{Y} = \{0, 1, \dots, s-1\}^m$ with $m < \mathfrak{A}_2(n, d)$ and $d = 4(s-1)-1$, then $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$ has the expected dimension. It is known that $\mathfrak{A}_2(n, d) \geq 2^{n-\lceil\log_2(\sum_{j=0}^{d-2}\binom{n-1}{j})\rceil}$.

Proposition 25 (Binary hidden units) *Let $\mathcal{Y} = \{0, 1\}^m$ and $\mathcal{X}$ be arbitrary.*

- *If $m + 1 \leq \mathfrak{A}_\mathcal{X}(3)$, then $\mathrm{RBM}^{\text{tropical}}_{\mathcal{X},\{0,1\}^m}$ has dimension $(1 + m)(1 + \sum_{i\in[n]}(|\mathcal{X}_i| - 1)) - 1$.*
- *If $m + 1 \geq \mathfrak{K}_\mathcal{X}(1)$, then $\mathrm{RBM}^{\text{tropical}}_{\mathcal{X},\{0,1\}^m}$ has dimension $|\mathcal{X}| - 1$.*

Let $\mathcal{Y} = \{0, 1\}^m$ and $\mathcal{X} = \{0, 1, \dots, q - 1\}^n$, where q is a prime power.

- *If $m + 1 \leq q^{n-\left\lceil\log_q(1+(n-1)(q-1)+1)\right\rceil}$, then $\mathrm{RBM}^{\text{tropical}}_{\mathcal{X},\mathcal{Y}}$ has dimension $(1 + m)(1 + \sum_{i\in[n]}(|\mathcal{X}_i| - 1)) - 1$.*
- *If $n = (q^r - 1)/(q - 1)$ for some $r \geq 2$, then $\mathcal{A}_\mathcal{X}(3) = \mathfrak{K}_\mathcal{X}(1)$, and $\mathrm{RBM}^{\text{tropical}}_{\mathcal{X},\mathcal{Y}}$ has the expected dimension for any m.*

In particular, when all units are binary and $m < 2^{n-\lceil \log_2(n+1)\rceil}$, then $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$ has the expected dimension; this was shown in (Cueto et al., 2010).

Proposition 26 (Arbitrary sized units) *If $\mathcal{X}$ contains m disjoint Hamming balls of radii $2|\mathcal{Y}_1|-3,\dots,2|\mathcal{Y}_m|-3$, and the complement of their union has full rank, then $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}^{\text{tropical}}$ has the expected dimension.*

Proof Propositions 24, 25, and 26 follow from Theorem 21 and Corollary 23 together with the following explicit bounds on $\mathfrak{A}$ by (Gilbert, 1952; Varshamov, 1957):

$$\mathfrak{A}_q(n,d) \geq \frac{q^n}{\sum_{j=0}^{d-1}\binom{n}{j}(q-1)^j}.$$

If q is a prime power, then $\mathfrak{A}_q(n,d) \geq q^k$, where k is the largest integer with $q^k < \frac{q^n}{\sum_{j=0}^{d-2}\binom{n-1}{j}(q-1)^j}$. In particular, $\mathfrak{A}_2(n,3) \geq 2^k$, where k is the largest integer with $2^k < \frac{2^n}{(n-1)+1} = 2^{n-\log_2(n)}$, i.e., $k = n - \lceil \log_2(n+1)\rceil$. ■

Example 27 Many results in coding theory can now be translated directly to statements about the dimension of discrete RBMs. Here is an example. Let $\mathcal{X} = \{1,2,\dots,s\}\times\{1,2,\dots,s\}\times\{1,2,\dots,t\}$, $s \leq t$. The minimum cardinality of a code $C \subseteq \mathcal{X}$ with covering-radius one equals $\mathfrak{K}_{\mathcal{X}}(1) = s^2 - \left\lfloor \frac{(3s-t)^2}{8} \right\rfloor$ if $t \leq 3s$, and $\mathfrak{K}_{\mathcal{X}}(1) = s^2$ otherwise (see Cohen et al., 2005, Theorem 3.7.4). Hence $\mathrm{RBM}_{\mathcal{X},\{0,1\}^m}^{\text{tropical}}$ has dimension $|\mathcal{X}|-1$ when $m+1 \geq s^2 - \left\lfloor \frac{(3s-t)^2}{8} \right\rfloor$ and $t \leq 3s$, and when $m+1 \geq s^2$ and $t > 3s$.

8. Discussion

In this note we study the representational power of RBMs with discrete units. Our results generalize a diversity of previously known results for standard binary RBMs and naïve Bayes models. They help contrasting the geometric-combinatorial properties of distributed products of experts versus non-distributed mixtures of experts.

We estimate the number of hidden units for which discrete RBM models can approximate any distribution to any desired accuracy, depending on the cardinalities of their units' state spaces. This analysis shows that the maximal approximation error increases at most logarithmically with the total number of visible states and decreases at least logarithmically with the sum of the number of states of the hidden units. This observation could be helpful, for example, in designing a penalty term to allow comparison of models with differing numbers of units. It is worth mentioning that the submodels of discrete RBMs described in Theorem 15 can be used not only to estimate the maximal model approximation errors, but also the expected model approximation errors given a prior of target distributions on the probability simplex. See (Montúfar and Rauh, 2014) for an exact analysis of Dirichlet priors. In future work it would be interesting to study the statistical approximation errors of discrete RBMs and to complement our theory by a discussion about parameter estimation and learning algorithms, together with an empirical evaluation.

Our estimation of the maximal approximation errors of discrete RBMs is based on explicit descriptions of classes of probability distributions that can be represented by these models. In future work, it would be interesting to take a closer look at the approximation errors for selected classes of target distributions that could appear in practice.

The combinatorics of tropical discrete RBMs allows us to relate the dimension of discrete RBM models to the solutions of linear optimization problems and slicings of convex support polytopes by normal fans of simplices. We use this to show that the model $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$ has the expected dimension for many choices of $\mathcal{X}$ and $\mathcal{Y}$, but not for all choices. We based our explicit computations of the dimension of RBMs on slicings by collections of parallel hyperplanes, but more general classes of slicings may be considered. The same tools presented in this paper can be used to estimate the dimension of a general class of models involving interactions within layers, defined as Kronecker products of hierarchical models (see Montúfar and Morton, 2013). We think that the geometric-combinatorial picture of discrete RBMs developed in this paper may be helpful in solving various long standing theoretical problems in the future, for example: What is the exact dimension of naïve Bayes models with general discrete variables? What is the smallest number of hidden variables that make an RBM a universal approximator? Do binary RBMs always have the expected dimension?

ACKNOWLEDGMENTS

We are grateful to the ICLR 2013 community for very valuable comments. This work was accomplished in part while G.M. was with the Department of Mathematics at the Pennsylvania State University in 2012 and 2013 and in part during his stays at the Max Planck Institute for Mathematics in the Sciences in September and October 2012. This work was supported in part by DARPA grant FA8650-11-1-7145.

References

M. Aoyagi. Stochastic complexity and generalization error of a restricted Boltzmann machine in Bayesian estimation. *Journal of Machine Learning Research*, 11:1243–1272, April 2010.

N. Ay and A. Knauf. Maximizing multi-information. *Kybernetika*, 42(5):517–538, 2006.

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

Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 153–160. MIT Press, Cambridge, MA, 2007.

M. Á. Carreira-Perpiñán and G. E. Hinton. On contrastive divergence learning. In *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, AISTATS '05, 2005.

M. V. Catalisano, A. V. Geramita, and A. Gimigliano. Secant varieties of $\mathbb{P}^1 \times \cdots \times \mathbb{P}^1$ (n-times) are not defective for $n \geq 5$. *Journal of Algebraic Geometry*, 20:295–327, 2011.

G. Cohen, I. Honkala, S. Litsyn, and A. Lobstein. *Covering Codes*. North-Holland Mathematical Library. Elsevier Science, 2005.

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*, volume 2. American Mathematical Society, 2010.

G. E. Dahl, R. P. Adams, and H. Larochelle. Training restricted Boltzmann machines on word observations. In John Langford and Joelle Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning*, ICML '12, pages 679–686, New York, NY, USA, July 2012. Omnipress.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

J. Draisma. A tropical approach to secant dimensions. *Journal of Pure and Applied Algebra*, 212 (2):349–363, 2008.

Y. Freund and D. Haussler. Unsupervised learning of distributions of binary vectors using 2-layer networks. In J. E. Moody, S. J. Hanson, and R. Lippmann, editors, *Advances in Neural Information Processing Systems 4*, NIPS '91, pages 912–919. Morgan Kaufmann, 1991.

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

G. E. Hinton. Products of experts. In *Proceedings of the 9th International Conference on Artificial Neural Networks*, volume 1 of *ICANN '99*, pages 1–6, 1999.

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, version 1. Technical report, UTML2010-003, University of Toronto, 2010.

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

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

P. M. Long and R. A. Servedio. Restricted Boltzmann machines are hard to approximately evaluate or simulate. In J. Fürnkranz and T. Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning*, ICML '10, pages 703–710. Omnipress, 2010.

D. Lowd and P. Domingos. Naive Bayes models for probability estimation. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML, pages 529–536. ACM Press, 2005.

T. K. Marks and J. R. Movellan. Diffusion networks, products of experts, and factor analysis. In *Proceedings of the 3rd International Conference Independent Component Analysis*, pages 481–485, 2001.

F. Matúš and N. Ay. On maximization of the information divergence from an exponential family. In *Proceedings of the 6th Workshop on Uncertainty Processing*, WUPES '03, pages 199–204. University of Economics, Prague, 2003.

R. Memisevic and G. E. Hinton. Learning to represent spatial transformations with factored higher-order Boltzmann machines. *Neural Computation*, 22(6):1473–1492, June 2010.

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

G. Montúfar. Universal approximation depth and errors of narrow belief networks with discrete units. *Neural Computation*, 26(7):1386–1407, 2014.

G. 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. Montúfar and J. Morton. Geometry of hidden-visible products of statistical models. 2013. In preparation.

G. Montúfar and J. Morton. When does a mixture of products contain a product of mixtures? *Accepted for SIAM Journal on Discrete Mathematics*, 2014. Preprint available at `http://arxiv.org/abs/1206.0387`.

G. Montúfar and J. Rauh. Scaling of model approximation errors and expected entropy distances. *Kybernetika*, 50(2):234–245, 2014. Special issue of the 9th Workshop on Uncertainty Processing (WUPES '12).

G. Montúfar, J. Rauh, and N. Ay. Expressive power and approximation errors of restricted Boltzmann machines. In J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F. C. N. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, NIPS '11, pages 415–423, 2011.

S. Osindero and G. E. Hinton. Modeling image patches with a directed hierarchy of Markov random fields. In J. C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, NIPS '07, pages 1121–1128. MIT Press, Cambridge, MA, 2008.

L. Pachter and B. Sturmfels. Tropical geometry of statistical models. *Proceedings of the National Academy of Sciences of the United States of America*, 101(46):16132–16137, November 2004.

M. Ranzato, A. Krizhevsky, and G. E. Hinton. Factored 3-Way Restricted Boltzmann Machines For Modeling Natural Images. In *Proceedings 13th International Conference on Artificial Intelligence and Statistics*, AISTATS '10, pages 621–628, 2010.

W. E. Roth. On direct product matrices. *Bulletin of the American Mathematical Society*, 40:461–468, 1934.

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 '07, pages 791–798, New York, NY, USA, 2007. ACM.

T. J. Sejnowski. Higher-order Boltzmann machines. In *Neural Networks for Computing*, pages 398–403. American Institute of Physics, 1986.

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

T. Tran, D. Phung, and S. Venkatesh. Mixed-variate restricted Boltzmann machines. In *Proceedings of the 3rd Asian Conference on Machine Learning*, ACML '11, pages 213–229, 2011.

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

M. Welling, M. Rosen-Zvi, and G. E. Hinton. Exponential family harmoniums with an application to information retrieval. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, NIPS '04, pages 1481–1488. MIT Press, Cambridge, MA, 2005.