Max-Planck-Institut für Mathematik in den Naturwissenschaften Leipzig

Restricted Boltzmann Machines: Introduction and Review

by

Guido Montúfar

Preprint no.: 87 2018

Restricted Boltzmann Machines: Introduction and Review

Guido Montúfar

Abstract The restricted Boltzmann machine is a network of stochastic units with undirected interactions between pairs of visible and hidden units. This model was popularized as a building block of deep learning architectures and has continued to play an important role in applied and theoretical machine learning. Restricted Boltzmann machines carry a rich structure, with connections to geometry, applied algebra, probability, statistics, machine learning, and other areas. The analysis of these models is attractive in its own right and also as a platform to combine and generalize mathematical tools for graphical models with hidden variables. This article gives an introduction to the mathematical analysis of restricted Boltzmann machines, reviews recent results on the geometry of the sets of probability distributions representable by these models, and suggests a few directions for further investigation.

Key words: hierarchical model, latent variable model, exponential family, mixture model, Hadamard product, non-negative tensor rank, expected dimension, universal approximation, Kullback-Leibler divergence, divergence maximization

1 Introduction

This article is intended as an introduction to the mathematical analysis of the restricted Boltzmann machine. Complementary to other existing and excellent introductions, we emphasize mathematical structures in relation to the geometry of the set of distributions that can be represented by this model. There is a large number of works on theory and applications of restricted Boltzmann machines. We review a selection of recent results in a way that, we hope, can serve as a guide to this rich

Guido Montúfar
Department of Mathematics and Department of Statistics, University of California, Los Angeles, USA; Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany.
e-mail: `montufar@math.ucla.edu`

subject, and lets us advertise some of the interesting and challenging problems that still remain to be addressed.

Brief overview

A Boltzmann machine is a model of pairwise interacting units that update their states over time in a probabilistic way depending on the states of the adjacent units. Boltzmann machines have been motivated as models for parallel distributed computing [36, 1, 37]. They can be regarded as stochastic versions of Hopfield networks [38], which serve as associative memories. They are closely related to mathematical models of interacting particles studied in statistical physics, especially the Ising model [39, Chapter 14]. For each fixed choice of interaction strengths and biases in the network, the collective of units assumes different states at relative frequencies that depend on their associated energy, in what is known as a Gibbs-Boltzmann probability distribution [30]. As pair interaction models, Boltzmann machines define special types of hierarchical log-linear models, which are special types of exponential family models [14] closely related to undirected graphical models [42, 40]. In contrast to the standard discussion of exponential families, Boltzmann machines usually involve hidden variables. Hierarchical log-linear models are widely used in statistics. Their geometric properties are studied especially in information geometry [5, 8, 6, 11] and algebraic statistics [21, 72]. The information geometry of the Boltzmann machine was first studied by Amari, Kurata, and Nagaoka [7].

A restricted Boltzmann machine (RBM) is a special type of a Boltzmann machine where the pair interactions are restricted to be between an observed set of units and an unobserved set of units. These models were introduced in the context of harmony theory [70] and unsupervised two layer networks [27]. RBMs played a key role in the development of greedy layer-wise learning algorithms for deep layered architectures [35, 12]. A recommended introduction to RBMs is [24]. RBMs have been studied intensively, with tools from optimization, algebraic geometry, combinatorics, coding theory, polyhedral geometry, and information geometry among others. Some of the advances over the past few years include results in relation to their approximation properties [77, 43, 58, 57], dimension [17, 53, 55], semi-algebraic description [18, 68], efficiency of representation [45, 54], sequential optimization [23, 26], statistical complexity [10], sampling and training [64, 22, 23, 26], information geometry [7, 6, 41].

Organization

This article is organized as follows. In Section 2 we introduce Boltzmann machines, Gibbs sampling, and the associated probability models. In Section 3 we introduce restricted Boltzmann machines and discuss various perspectives, viewing the probability models as marginals of exponential families with Kronecker factoring suffi-

cient statistics, as products of mixtures of product distributions, and as feedforward networks with soft-plus activations. We also discuss a piecewise linear approximation called the tropical RBM model, which corresponds to a feedforward network with rectified linear units. In Section 4 we give a brief introduction to training by maximizing the likelihood of a given data set. We comment on gradient, contrastive divergence, natural gradient, and EM methods. Thereafter, in Section 5 we discuss the Jacobian of the model parametrization and the model dimension. In Section 6 we discuss the representational power, covering two hierarchies of representable distributions, namely mixtures of product distributions and hierarchical log-linear models, depending on the number of hidden units of the RBM. In Section 7 we use the representation results to obtain bounds on the approximation errors of RBMs. In Section 8 we discuss semi-algebraic descriptions and a recent result for a small RBM. Finally, in Section 9 we collect a few open questions and possible research directions.

2 Boltzmann machines

A Boltzmann machine is a network of stochastic units. Each unit, or neuron, can take one of two states. A joint state of all units has an associated energy value which is determined by pair interactions and biases. The states of the units are updated in a stochastic manner at discrete time steps, whereby lower energy states are preferred over higher energy ones. In the limit of infinite time, the relative number of visits of each state, or the relative probability of observing each state, converges to a fixed value that is exponential in the energy differences. The set of stationary limit probability distributions that result from all possible values of the pair interactions and biases, forms a manifold of probability distributions called the Boltzmann machine probability model. The probability distributions for a subset of visible units are obtained via marginalization, adding the probabilities of all joint states that are compatible with the visible states. We make these notions more specific in the following.

Pairwise interacting units

We consider a network defined by a finite set of nodes N and a set of edges $I \subseteq \binom{N}{2}$ connecting pairs of nodes. Each node $i \in N$ corresponds to a random variable, or unit, with states $x_i \in \{0,1\}$. The joint states of all units are vectors $x = (x_i)_{i\in N} \in \{0,1\}^N$. Each unit $i \in N$ has an associated bias $\theta_i \in \mathbb{R}$, and each edge $\{i,j\} \in I$ has an associated interaction weight $\theta_{\{i,j\}} \in \mathbb{R}$. For any given value of the parameter $\theta = ((\theta_i)_{i\in N}, (\theta_{\{i,j\}})_{\{i,j\}\in I})$, the energy of the joint states x is given by

$$E(x;\theta) = -\sum_{i\in N} \theta_i x_i - \sum_{\{i,j\}\in I} \theta_{\{i,j\}} x_i x_j, \quad x \in \{0,1\}^N. \tag{1}$$

In particular, the negative energy function $-E(\cdot;\theta)$ is a linear combination of the functions $x \mapsto x_i$, $i \in N$, $x \mapsto x_i x_j$, $\{i,j\} \in I$, with coefficients θ. It takes lower values when pairs of units with positive interaction take the same states, or also when units with positive bias take state one.

State updates, Gibbs sampling

The Boltzmann machine updates the states of its units at discrete time steps, in a process known as Gibbs sampling. Given a state $x^{(t)} \in \{0,1\}^N$ at time t, the state $x^{(t+1)}$ at the next time step is created by selecting a unit $i \in N$, and then setting $x_i^{(t+1)} = 1$ with probability

$$\Pr\left(x_i^{(t+1)} = 1 | x^{(t)}\right) = \sigma\Big(\sum_{\{i,j\}\in I} \theta_{\{i,j\}} x_j^{(t)} + \theta_i\Big), \tag{2}$$

or $x_i^{(t+1)} = 0$ with complementary probability $\Pr(x_i^{(t+1)} = 0 | x^{(t)}) = 1 - \Pr(x_i^{(t+1)} = 1 | x^{(t)})$. Here $\sigma\colon s \mapsto 1/(1+\exp(-s))$ is the standard logistic function. In particular, the quotient of the probabilities of setting either $x_i = 1$ or $x_i = 0$ is the exponential energy difference $\sum_{\{i,j\}\in I} \theta_{\{i,j\}} x_j + \theta_i$ between the two resulting joint states. The activation probability (2) can be regarded as the output value of a deterministic neuron with inputs x_j weighted by $\theta_{\{i,j\}}$ for all adjacent js, bias θ_i, and activation function σ.

If the unit i to be updated at time t is selected according to a probability distribution r over N, and $T_i(x^{(t+1)}|x^{(t)})$ denotes the Markov transition kernel when choosing unit i, then the total transition kernel is

$$T = \sum_{i\in N} r(i) T_i.$$

In other words, if the state at time t is $x^{(t)}$, then the state $x^{(t+1)}$ at the next time step is drawn from the probability distribution $T(\cdot|x^{(t)})$. More generally, if $p^{(t)}$ is a probability distribution over joint states $x^{(t)} \in \{0,1\}^N$ at time t, then at time $t+1$ we have the probability distribution

$$p^{(t+1)} = p^{(t)} \cdot T.$$

The one step transition kernel T is non zero only between state vectors $x^{(t)}$ and $x^{(t+1)}$ that differ at most in one entry. However, if r is strictly positive, then there is a positive probability of transitioning from any state to any other state in N time steps, so that the N-th power T^N is strictly positive, implying that T is a primitive kernel.

Stationary limit distributions

If T is a primitive kernel, then there is a unique distribution p with $\lim_{t\to\infty} p^0 \cdot T^t = p$, for all start state distributions p^0. This follows from a theorem by Geman and Geman, which also shows that p is the Gibbs-Boltzmann distribution

$$p(x;\theta) = \frac{1}{Z(\theta)} \exp(-E(x;\theta)), \quad x \in \{0,1\}^N, \tag{3}$$

with the energy function $E(\cdot;\theta)$ given in (1) and normalizing partition function $Z(\theta) = \sum_{x'} \exp(-E(x';\theta))$.

The set of stationary distributions (3), for all $\theta \in \mathbb{R}^{|N|+|I|}$, is the Boltzmann machine probability model with interaction structure $G = (N, I)$. This is an exponential family with sufficient statistics $x_i, i \in N$, $x_i x_j$, $\{i, j\} \in I$ and canonical or exponential parameter θ. It is a smooth manifold of dimension $|N| + |I|$, contained in the $2^N - 1$ dimensional simplex of probability distributions on $\{0,1\}^N$,

$$\Delta_{\{0,1\}^N} = \Big\{ p \in \mathbb{R}^{\{0,1\}^N} : \, p(x) \geq 0 \;\text{ for all }\; x \in \{0,1\}^N, \;\text{ and } \sum_{x \in \{0,1\}^N} p(x) = 1 \Big\}.$$

Hidden units, visible marginal distributions

We will be interested in a situation where only a subset $V \subseteq N$ of all units can be observed, while the other units $H = N \setminus V$ are unobserved or hidden. Given the probability distribution $p(x;\theta)$ over the states $x = (x_V, x_H) \in \{0,1\}^V \times \{0,1\}^H$ of all units, the marginal probability distribution over the visible states x_V is given by

$$p(x_V;\theta) = \sum_{x_H \in \{0,1\}^H} p(x;\theta), \quad x_V \in \{0,1\}^V. \tag{4}$$

The set of marginal probability distributions, for all choices of θ, is a subset of the $2^V - 1$ dimensional simplex $\Delta_{\{0,1\}^V}$. It is the image of the fully observable Boltzmann machine probability manifold by the linear map that computes marginal distributions. In general this set is no longer a manifold. It may have a rather complex shape with self intersections and dimension strictly smaller than that of the manifold of distributions of all units. We will be concerned with the properties of this set in the special case where interaction edges are only allowed between visible and hidden units.

3 Restricted Boltzmann machines

The restricted Boltzmann machine (RBM) is a special type of Boltzmann machine where the interactions are restricted to be between visible and hidden units, such that $I = \{\{i,j\} : i \in V, j \in H\}$. This is illustrated in Figure 1. The corresponding probability distributions take the form

$$p(x;\theta) = \frac{1}{Z(\theta)} \sum_{y \in \{0,1\}^H} \exp(y^\top W x + c^\top y + b^\top x), \quad x \in \{0,1\}^V. \tag{5}$$

Here x is the state of the visible units, y is the state of the hidden units, Z is the partition function, and $\theta = (W,b,c)$ denotes the parameters, composed of the interaction weights $W = (w_{j,i})_{j \in H, i \in V}$, the biases of the visible units $b = (b_i)_{i \in V}$, and the biases of the hidden units $c = (c_j)_{j \in H}$. The RBM probability model with n visible and m hidden units is the set of probability distributions of the form (5), for all possible choices of θ. We denote this set by $\mathrm{RBM}_{n,m}$. We will write $[n] = \{1,\dots,n\}$ and $[m] = \{1,\dots,m\}$ to enumerate the visible and hidden units, respectively. We write $\mathscr{X} = \{0,1\}^V$ for the state space of the visible units, and $\mathscr{Y} = \{0,1\}^H$ for that of the hidden units.

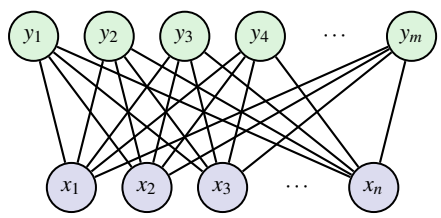

Fig. 1 RBM as a graphical model with visible units $x_1,\dots,x_n$ and hidden units $y_1,\dots,y_m$. Each edge has an associated interaction weight w_{ji}, each visible node has an associated bias weight b_i, and each hidden node an associated bias weight c_j.

An RBM probability model can be interpreted in various interesting and useful ways, as we discuss in the following. These are views of the same object and are equivalent in that sense, but they highlight different aspects.

Product of mixtures

One interpretation the RBM is as a *product of experts* model, meaning that it consists of probability distributions which are normalized entrywise products with factors coming from some fixed models. Factorized descriptions are familiar from graphical models, where one considers probability distributions that factorize into potential functions, which are arbitrary positive valued functions that depend only on certain fixed subsets of all variables. We discuss graphical models in more depth in

Section 6. In the case of RBMs, each factor model is given by mixtures of product distributions. A *product distribution* is a distribution of multiple variables which factorizes as an outer product $q(x_1,\dots,x_n) = \prod_{i\in[n]} q_i(x_i)$ of distributions q_i of the individual variables. A *mixture distribution* is a convex combination $q(x) = \sum_k \lambda_k q_k(x)$, where the λ_k are non-negative weights adding to one, and the q_k are probability distributions from some given set. Indeed, the RBM distributions can be written as

$$\begin{aligned} p(x;\theta) =& \frac{1}{Z(\theta)} \sum_{y\in\{0,1\}^m} \exp(y^\top W x + c^\top y + b^\top x) \\ =& \frac{1}{Z(\theta)} \exp(b^\top x) \prod_{j\in[m]} (1+\exp(W_{j:}x + c_j)) \\ =& \frac{1}{Z(\theta)} \prod_{j\in[m]} (\exp(W'_{j:}x) + \exp(c_j)\exp(W''_{j:}x)). \end{aligned} \tag{6}$$

Here $W'_{j:}$ and $W''_{j:} = W_{j:} + W'_{j:}$ can be chosen arbitrarily in $\mathbb{R}^n$ for all $j \in [m]$, with $b = \sum_{j\in[m]} W'_{j:}$. In turn, for any mixture weights $\lambda_j \in (0,1)$ we can find suitable $c_j \in \mathbb{R}$, and for any distributions $p'_{j,i}$ and $p''_{j,i}$ on $\mathscr{X}_i = \{0,1\}$ suitable $W'_{j,i}$ and $W''_{j,i}$, such that

$$p(x;\theta) = \frac{1}{Z(\theta)} \prod_{j\in[m]} \Big(\lambda_j \prod_{i\in[n]} p'_{j,i}(x_i) + (1-\lambda_j) \prod_{i\in[n]} p''_{j,i}(x_i)\Big). \tag{7}$$

This shows that the RBM model can be regarded as the set distributions that are entrywise products of m terms, with each term being a mixture of two product distributions over the visible states.

Products of experts can be trained in an efficient way, with methods such as contrastive divergence, which we will outline in Section 4. Products of experts also relate to the notion of distributed representations, where each observation is explained by multiple latent causes. This allows RBMs to create exponentially many inference regions, or possible categorizations of input examples, on the basis of only a polynomial number of parameters. This sets RBMs apart from mixture models, and provides one way of breaking the curse of dimensionality, which is one motivation for choosing one network architecture over another in the first place. We discuss more about this further below and in Section 6.

Tensors and polynomial parametrization

A probability distribution on $\{0,1\}^n$ can be regarded as an n-way table or tensor with entries indexed by $x_i \in \{0,1\}$, $i \in [n]$. A tensor p is said to have rank one if it can be factorized as $p = p_1 \otimes \cdots \otimes p_n$, where each p_i is a vector. Thus, non-negative rank one tensors correspond to product distributions. A tensor is said to have non-negative rank k if it can be written as the sum of k non-negative tensors of rank 1, and k is the smallest number for which this is possible. Tensors of non-

negative rank at most k correspond to mixtures of k product distributions. The RBM distributions are, up to normalization, the tensors that can be written as Hadamard (i.e., entrywise) products of m factor tensors of non-negative rank at most two. The representable tensors have the form

$$p = \prod_{j\in[m]} \big(q'_{j,1} \otimes \cdots \otimes q'_{j,n} + q''_{j,1} \otimes \cdots \otimes q''_{j,n}\big), \tag{8}$$

where the $q'_{j,i}$ and $q''_{j,i}$ are non-negative vectors of length two.

In particular, we note that, up to normalization, the RBM distributions have a polynomial parametrization

$$p = \Big(\prod_{i\in[n]} \omega_{0,i}^{x_i}\Big) \prod_{j\in[m]} \Big(1 + \omega_{j,0} \prod_{i\in[n]} \omega_{j,i}^{x_i}\Big), \tag{9}$$

with parameters $\omega_{0,i} = \exp(b_i) \in \mathbb{R}_+$, $\omega_{j,0} = \exp(c_j) \in \mathbb{R}_+$, $j \in [m]$, $\omega_{j,i} = \exp(W_{j,i}) \in \mathbb{R}_+$, $(i,j) \in [n] \times [m]$. The fact that RBMs have a polynomial parametrization makes them, like many other probability models, amenable to be studied with tools from algebra. This is the realm of *algebraic statistics*. Introductions to this area at the intersection of mathematics and statistics are [21, 72]. In algebraic geometry one studies questions such as the dimension and degree of solution sets of polynomial equations. When translated to statistics, these questions relate to parameter identifiability, the number of maximizers of the likelihood function, and other important properties of statistical models.

Kronecker products, harmonium models

As we have seen, the joint distributions of a Boltzmann machine form an exponential family over the states of all units. That is, the joint distributions are given by exponentiating and normalizing vectors from an affine space,

$$p(x,y;\theta) = \frac{1}{Z(\theta)} \exp(\theta^\top F(x,y)), \quad (x,y) \in \mathscr{X} \times \mathscr{Y}. \tag{10}$$

Here the sufficient statistics $F_1,\dots,F_d \colon \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$ span the affine space in question. For an RBM, the sufficient statistics F have a special structure. Recall that the Kronecker product of two matrices is defined by $(a_{i,j})_{i,j} \otimes (b_{k,l})_{k,l} = (a_{i,j}(b_{k,l})_{k,l})_{i,j} = (a_{i,j}b_{k,l})_{(i,k),(j,l)}$. The sufficient statistics for the exponential family of the RBM can be written as a Kronecker product

$$F(x,y) = F^V(x) \otimes F^H(y), \quad (x,y) \in \mathscr{X} \times \mathscr{Y}, \tag{11}$$

where $F^V(x) = (1,x_1,\dots,x_n)^\top$ and $F^H(y) = (1,y_1,\dots,y_m)^\top$ are sufficient statistics of the independence models of the n visible binary units and the m hidden binary units. The independence model is the exponential family of product distributions, $\frac{1}{Z}\exp(\sum_i \theta_i F_i^V(x)) = \frac{1}{Z}\exp(w^\top x + c) = \frac{1}{Z}\prod_{i\in[n]}\exp(w_i x_i)$.

The Kronecker product structure allows us to express the conditional distribution of hidden units given visible units, and vice versa, in the following simple way. Given two vectors a,b, write $\langle a,b\rangle$ for their inner product $a^\top b = \sum_i a_i b_i$. Take any parameter vector $\theta \in \mathbb{R}^{(n+1)(m+1)}$ and arrange its entries into a matrix $\Theta \in \mathbb{R}^{(m+1)\times(n+1)}$, going column by column. Then

$$\begin{aligned}\big\langle \theta, F(x,y)\big\rangle &= \big\langle \theta, F^V(x)\otimes F^H(y)\big\rangle \\ &= \big\langle \Theta^\top F^H(y), F^V(x)\big\rangle \\ &= \big\langle \Theta F^V(x), F^H(y)\big\rangle.\end{aligned}$$

These expression describe following probability distributions:

$$\begin{aligned}p(x,y;\theta) &= \frac{1}{Z(\theta)}\exp\big(\big\langle \theta, F(x,y)\big\rangle\big) \\ p(x|y;\theta) &= \frac{1}{Z\big(\Theta^\top F^H(y)\big)}\exp\big(\big\langle \Theta^\top F^H(y), F^V(x)\big\rangle\big) \\ p(y|x;\theta) &= \frac{1}{Z\big(\Theta F^V(x)\big)}\exp\big(\big\langle \Theta F^V(x), F^H(y)\big\rangle\big).\end{aligned}$$

Geometrically, ΘF^V is a linear projection of F^V into the parameter space of the exponential family with sufficient statistics F^H and, similarly, $\Theta^\top F^H$ is a linear projection of F^H into the parameter space of an exponential family for the visible variables. This is illustrated in Figure 2.

Restricted mixtures of products

The marginal distributions can always be written as

$$p(x;\theta) = \sum_y p(x,y;\theta) = \sum_y p(y;\theta)p(x|y;\theta), \quad x\in\mathscr{X}.$$

In the case of an RBM, the conditional distributions are product distributions $p(x|y;\theta) = \prod_{i\in[n]} p(x_i|y;\theta)$. In turn, the RBM model consists of mixtures of product distributions, with mixture weights $p(y;\theta)$. However, the marginal $p(y;\theta)$ and the tuple of conditionals $p(x|y;\theta)$ have a specific and constrained structure. For instance, as can be seen in Figure 2 for the model $\mathrm{RBM}_{3,2}$, the mixture components have parameter vectors that are affinely dependent. One implication is that $\mathrm{RBM}_{3,2}$ cannot represent any distribution with large values on the even parity strings $000, 011, 101, 110$ and small values on the odd parity strings $001, 010, 100, 111$. This kind of constraint, coming from constraints on the mixture components, have been studied in [54]. An exact description of the constraints that apply to the probability distributions within $\mathrm{RBM}_{3,2}$ was obtained recently in [68]. We comment on this later in Section 8.

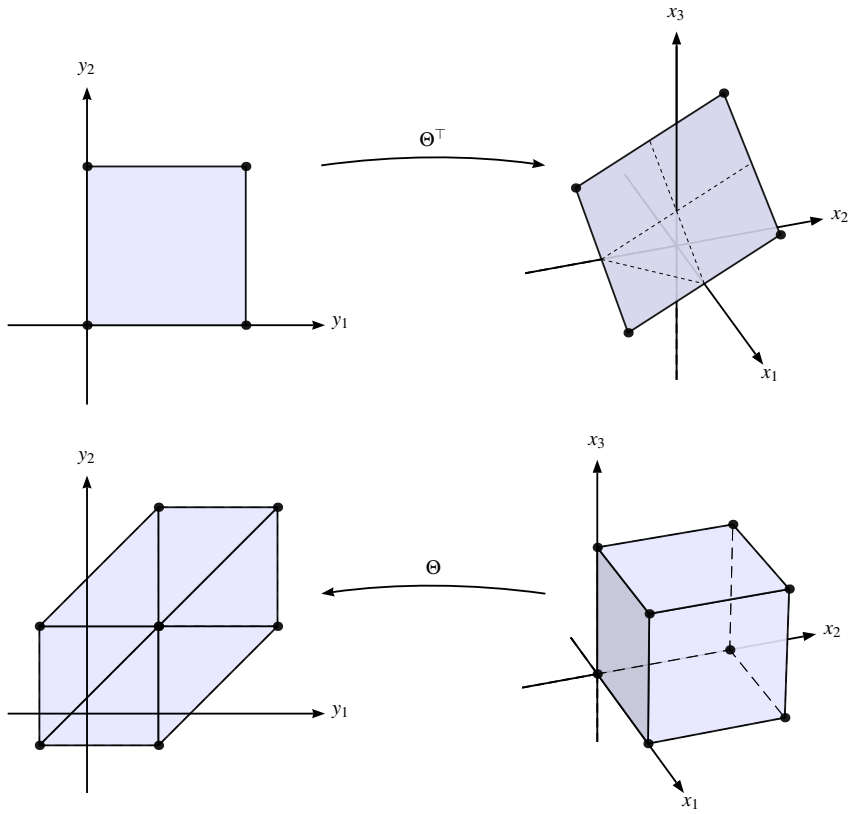

Fig. 2 For an RBM, the conditional distributions $p(X|y;\theta)$ of the visible variables given the hidden variables, are the elements of an exponential family with sufficient statistics F^V and parameters given by projections $\Theta^\top F^H(y)$ of the sufficient statistics F^H of the hidden variables. Similarly, $p(Y|x;\theta)$ are exponential family distributions with sufficient statistics F^H and parameters $\Theta F^V(x)$. The figure illustrates these vectors for $\mathrm{RBM}_{3,2}$ and a choice of θ.

Superposition of soft-plus units

Another useful way of viewing RBMs is as follows. The description as products of mixtures shows that in RBMs the log-probabilities are sums of independent terms. More precisely, they are superpositions of m soft-plus units and one linear unit:

$$\log(p(x;\theta)) = \sum_{j\in[m]} \log(1+\exp(W_{j:}x+c_j)) + b^\top x - \log(Z(\theta)). \qquad (12)$$

A *soft-plus unit* computes a real valued affine function of its arguments, $x \mapsto w^\top x + c$, and then applies the soft-plus non linearity $s \mapsto \log(1+\exp(s))$. A linear unit simply computes $x \mapsto b^\top x + c$.

Log-probabilities correspond uniquely to probability distributions. When studying the space of representable log-probabilities, it is helpful to allow ourselves to add or disregard additive constants, since they correspond to scaling factors that cancel out with the normalization of the probability distributions.

The RBM model can be regarded as the set of negative energy functions (log-probabilities modulo additive constants) that can be computed by a feedforward network with one hidden layer of m soft-plus units and one linear unit, and a single output unit adding the outputs of the hidden units. The situation is illustrated in Figure 3. Feedforward networks are often conceptually easier than stochastic networks or probabilistic graphical models. One point to note is that the output unit of the RBM energy network only computes unweighted sums.

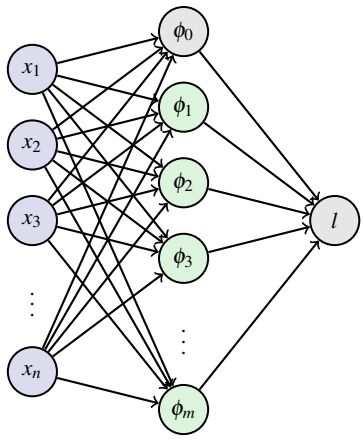

Fig. 3 An RBM model can be regarded as the set of log-probabilities which are computable as the sum of a linear unit ϕ_0 and m soft-plus units ϕ_j, $j = 1, \ldots, m$.

A type of computational unit that is closely related to the soft-plus unit is the *rectified linear unit* (ReLU). A ReLU computes a real valued affine function of its arguments, $x \mapsto w^\top x + c$, followed by rectification $s \mapsto [s]_+ = \max\{0, s\}$. As it turns out, if we replace the soft-plus units by ReLUs in eq. (12), we obtain the so-called tropical RBM model, which is a piecewise linear version of the original model that facilitates a number of computations. We discuss more details of this relationship in the next paragraph.

Tropical RBM, superposition of ReLUs

The tropical RBM model is the set of vectors that we obtain when evaluating log-probabilities of the RBM model using the max-plus algebra and disregarding additive constants. We replace sums by maximum, so that a log-probability vector $l(x;\theta) = \sum_y \exp(y^\top W x + b^\top x + c^\top y)$, $x \in \mathscr{X}$, becomes $\Phi(x;\theta) = \max_y\{y^\top W x +$

$b^\top x + c^\top y\}$, $x \in \mathscr{X}$. We can write this more compactly as

$$\Phi(x;\theta) = \theta^\top F(x, h(x;\theta)), \quad x \in \mathscr{X}, \tag{13}$$

where $F(x,y) = (1, x_1, \ldots, x_n)^\top \otimes (1, y_1, \ldots, y_m)^\top$ is the vector of sufficient statistics, and $h(x;\theta) = \operatorname{argmax}_y \theta^\top F(x,y) = \operatorname{argmax}_y p(y|x;\theta)$ is the *inference function* that returns the most probable y given x. In particular, the tropical RBM model is the image of a piecewise linear map.

We note the following decomposition, which expresses the tropical RBM model as a superposition of one linear unit and m ReLUs. We have

$$\begin{aligned}\Phi(x;\theta) &= \max_y \{y^\top W x + b^\top x + c^\top y\} \\ &= b^\top x + \sum_{j\in[m]} \max_{y_j} \{y_j W_{j:} x + c_j y_j\} \\ &= b^\top x + \sum_{j\in[m]} [W_{j:} x + c_j]_+ .\end{aligned}$$

In turn, the tropical RBM is the set of vectors computable by a sum of one linear unit $x \mapsto b^\top x$ and m ReLUs $x \mapsto [w^\top x + c]_+ = \max\{0, w^\top x + c\}$.

The set of functions that can be represented by a ReLU is closed under multiplication by non-negative scalars. Hence the unweighted sums of m ReLUs, $\sum_{j\in[m]} [w_j^\top x + c_j]_+$, express the same set of functions as the conic combinations of m ReLUs, $\sum_{j\in[m]} \alpha_j [\bar{w}_j^\top x + \bar{c}_j]_+$, where $\alpha_j \geq 0$, $j \in [m]$. For analysis and visualization, we can disregard positive multiplicative factors, and consider convex combinations of m normalized ReLUs. We can normalize each function such that its entry sum equals one. Zero functions cannot be normalized in this way, but they are equivalent to constant functions. The set of normalized functions expressible by a ReLU with two binary inputs is shown in Figure 4. A sum of m ReLUs can realize any convex combinations of m points from this set. Affine functions with positive values correspond to the horizontal square in the middle of the figure, and constant functions to the point at the center of the square. Adding positive / negative constants to a given point corresponds to moving from it towards / away from the center.

Other generalizations

There are numerous generalizations of the regular RBM model.

- A Boltzmann machine can be defined with discrete non-binary states, real valued Gaussian units, or any other type of probability model for each unit. If the hidden variables are defined to take k possible values each, then the RBM defines a Hadamard product of tensors of non-negative rank at most k [53]. In particular, this is a generalization of mixtures of products models. Visible units with more than two states have been used, for example, in collaborative filtering [67].

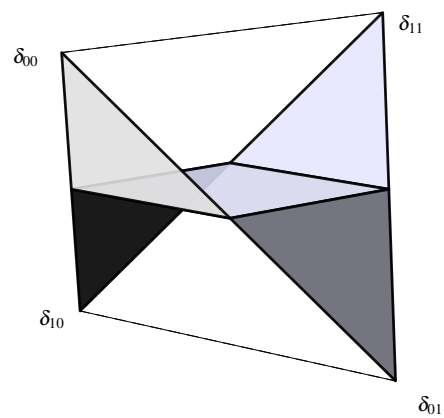

Fig. 4 Illustration of the set of functions $([w^\top x + c]_+)_x$, $x \in \{0,1\}^2$, that can be represented by a ReLU with two binary inputs. This corresponds to the tropical RBM model with zero biases on the visible units. For the visualization of this 3 dimensional set in $\mathbb{R}^4_{\geq 0}$, we scaled the vectors to have entry sum 1 (the zero function is identified with the one function), which results in the shown subset of the simplex with vertices δ_x the indicators of individual inputs $x \in \{0,1\}^2$.

- Viewed as Kronecker product models, with distributions $\frac{1}{Z(\theta)} \sum_y \exp(\theta^\top F^V(x) \otimes F^H(y))$, RBMs can be generalized to have arbitrary factors F^V and F^H, rather than just sufficient statistics of independence models. In this case, the conditional distributions of the visible variables, given the hidden variables, are distributions from the exponential family specified by F^V. This setting has been discussed in [55] and in [76] by the name *exponential family harmonium*.
- We can extend the setting of pair interactions to models with higher order interactions, called higher order Boltzmann machines [69].
- Other generalizations include deep architectures, such as deep belief networks [35] and deep Boltzmann machines [66]. Here one considers a stack of layers of units, with interactions restricted to pairs of units at adjacent layers. The representational power of deep belief networks has been studied in [73, 43, 51, 49] and that of deep Boltzmann machines in [50].
- For some applications, such as discriminative tasks, structured output prediction, stochastic control, one splits the visible units into a set of inputs and a set of outputs. The representational power of *conditional RBMs* has been studied in [52].
- Another line of generalizations are quantum models [9].
- A recent overview on RBM variants for diverse applications was given in [78].

4 Basics of training

We give a short introduction to training. The general idea of training is to adjust the parameters of the Boltzmann machine such that it behaves in a desirable way. To do this, we first decide on a function to measure the desirability of the different possible behaviors, and then maximize that function over the model parameters. The first explicit motivation and derivation of a learning algorithm for Boltzmann machines is by Ackley, Hinton, and Sejnowski [1], based on statistical mechanics. Given a set of examples, the algorithm modifies the interaction weights and biases of the network so as to construct a generative model that produces examples with the same probability distribution of the provided examples.

Maximizing the likelihood of a data set

Based on a set of examples, we aim at generating examples with the same probability distribution. To this end, we can maximize the log-likelihood of the provided examples with respect to the Boltzmann machine model parameters. For a set of examples $x^1,\dots,x^N \in \{0,1\}^n$, the log-likelihood is

$$L(\theta) = \sum_{i=1}^{N} \log p(x^i;\theta) = \sum_x p_{\text{data}}(x) \log p(x;\theta), \tag{14}$$

where p_{data} is the empirical data distribution $p_{\text{data}}(x) = \frac{1}{N}\sum_{i=1}^{N} \delta_{x^i}(x)$, $x \in \mathscr{X}$, and $p(x;\theta)$, $x \in \mathscr{X}$, is the model distribution with parameter $\theta \in \mathbb{R}^d$. Maximizing (14) with respect to θ is equivalent to minimizing the Kullback-Leibler divergence $D(p_{\text{data}}\|p_\theta)$ from p_{data} to the model distribution $p_\theta \equiv p(\cdot;\theta)$, again with respect to θ. The divergence is defined as

$$D(p_{\text{data}}\|p_\theta) = \sum_x p_{\text{data}}(x) \log \frac{p_{\text{data}}(x)}{p(x;\theta)}. \tag{15}$$

In some cases the minimum might not be attained by any value of the parameter θ. However, it is attained as $D(p_{\text{data}}\|p)$ for some distribution p in the closure of $\{p_\theta \colon \theta \in \mathbb{R}^d\} \subseteq \Delta_{\mathscr{X}}$.

Likelihood gradient

In most cases, we do not know how to maximize the log-likelihood in closed form (we discuss a recent exception to this in Section 8). We can search for a maximizer by initializing the parameters at some random value $\theta^{(0)}$ and iteratively adjusting them in the direction of the gradient, as

$$\theta^{(t+1)} = \theta^{(t)} + \alpha_t \nabla L(\theta^{(t)}), \tag{16}$$

until some convergence criterion is met. Here the *learning rate* $\alpha_t > 0$ is a hyper-parameter of the learning criterion that needs to be specified. Typically the user tries a range of values. Often in practice, the parameter updates are computed based only on subsets of the data at the time, in what is known as on-line, mini-batch, or stochastic gradient.

Writing $F\colon \mathscr{X} \times \mathscr{Y} \to \mathbb{R}^d$ for the sufficient statistics of an exponential family of joint distributions of visible and hidden variables, we have

$$\nabla L(\theta) = \langle F \rangle_{\text{data}} - \langle F \rangle_\theta. \tag{17}$$

Here $\nabla = (\frac{\partial}{\partial \theta_1}, \dots, \frac{\partial}{\partial \theta_d})^\top$ is the column vector of partial derivatives with respect to the model parameters, $\langle \cdot \rangle_{\text{data}}$ stands for the expectation value with respect to the joint probability distribution $p_{\text{data}}(x) p_\theta(y|x)$, and $\langle \cdot \rangle_\theta$ stands for the expectation with respect to the joint distribution $p_\theta(x,y)$.

The computation of the gradient can be implemented as follows. We focus on the binary RBM, for which the sufficient statistics take the form

$$F(x,y) = (F_I, F_V, F_H)(x,y) = ((y_j x_i)_{j \in H, i \in V}, (x_i)_{i \in V}, (y_j)_{j \in H}),\ (x,y) \in \{0,1\}^V \times \{0,1\}^H.$$

For the expectation value in (17) involving the data distribution:

- Write a data matrix $\tilde{X} = (x^1, \cdots, x^N)$.
- Collect the activation probabilities of the individual hidden units, in response to each visible data vector, into a matrix $\tilde{Y} = \sigma(c \cdot \mathbb{1}_{1 \times N} + W \cdot \tilde{X})$. Here σ is the logistic function $s \mapsto 1/(1+\exp(-s))$ applied entrywise to the argument, and $\mathbb{1}_{1 \times N}$ is the $1 \times N$ matrix of ones.
- Then

$$\begin{aligned} \langle F_I \rangle_{\text{data}} &= \tilde{Y} \cdot \tilde{X}^\top / N, \\ \langle F_V \rangle_{\text{data}} &= \tilde{X} \cdot \mathbb{1}_{N \times 1} / N, \\ \langle F_H \rangle_{\text{data}} &= \tilde{Y} \cdot \mathbb{1}_{N \times 1} / N. \end{aligned} \tag{18}$$

This calculation is relatively tractable, with order Nnm operations.

For the expectation in (17) with respect to the model distribution:

- Write X for the matrix with columns all vectors in $\{0,1\}^n$ and Y for the matrix with columns all vectors in $\{0,1\}^m$.
- Let $P_{Y \times X}$ be the matrix with entries $p_\theta(x,y)$, with rows and columns indexed by y and x.
- Then

$$\begin{aligned} \langle F_I \rangle_\theta &= Y \cdot P_{Y \times X} \cdot X^\top, \\ \langle F_V \rangle_\theta &= \mathbb{1}_{1 \times 2^m} \cdot P_{Y \times X} \cdot X^\top, \\ \langle F_H \rangle_\theta &= Y \cdot P_{Y \times X} \cdot \mathbb{1}_{2^n \times 1}. \end{aligned} \tag{19}$$

This calculation is possible for small models, but it can quickly become intractable. Since $P_{Y\times X}$ has 2^m rows and 2^n columns, computing its partition function and the expectations requires exponentially many operations in the number of units. In applications n and m may be in the order of hundreds or thousands. In order to overcome the intractability of this computation, a natural approach is to approximate the expectation values by sample averages. We discuss this next.

Contrastive divergence

The expectations $\langle F\rangle_\theta$ with respect to the model distribution can be approximated in terms of sample averages obtained by Gibbs sampling the RBM. One method based on this idea is *contrastive divergence* (CD) [33]. This method has been enormously valuable in practical applications and is the standard learning algorithm for RBMs. The CD algorithm can be implemented as follows.

- As before, write a data matrix $\tilde{X} = (x^1,\dots,x^N)$.
- Then update the state of the hidden units of the RBM by

$$\tilde{Y} = (\sigma(c\cdot \mathbb{1}_{1\times N} + W\cdot \tilde{X}) \geq \mathrm{rand}_{m\times N}).$$

- Update the state of the visible units by

$$\hat{X} = (\sigma(b\cdot \mathbb{1}_{1\times N} + W^\top \tilde{Y}) \geq \mathrm{rand}_{n\times N}).$$

  These updates are the Gibbs sampling state updates described in eq. (2), computed in parallel for all hidden and visible units. Here $\mathrm{rand}_{n\times N}$ is an $n\times N$ array of independent variables uniformly distributed in $[0,1]$, and $\geq$ is evaluated entrywise as a logic gate with binary outputs.
- Now use the reconstructed data $\hat{X}$ to compute $\langle F\rangle_{\mathrm{recon}}$ in the same way as $\tilde{X}$ was used to compute $\langle F\rangle_{\mathrm{data}}$ in eq. (18). The approximate model sample average $\langle F\rangle_{\mathrm{recon}}$ is then used as an approximation of $\langle F\rangle_\theta$.

This calculation involves only order Nnm operations, and remains tractable even for relatively large n and m in the order of thousands.

CD is an approximation to the maximum likelihood gradient. The bias of this method with respect to the actual gradient has been studied theoretically in [23]. There are a number of useful variants of the basic CD method. One can use k Gibbs updates, instead of just one, in what is known as the CD_k method. The larger k, the more one can expect the samples to follow the model distribution. In this spirit, there is also the persistent CD method (PCD) [74], where each sampling chain is initialized at previous samples, rather than at examples form the data set. Another useful technique in this context is parallel tempering [65, 24]. Moreover, basic gradient methods are often combined with other strategies, such as momentum, weight decay, pre-conditioners, second order methods. For more details see the introduction to training RBMs [25] and the useful practical guide [34].

Natural gradient

A natural modification of the standard gradient method is the *natural gradient*, which is based on the notion that the parameter space has an underlying geometric structure. This is the point of view of *information geometry* [3, 8, 6]. A recent mathematical account on this topic is given in the book [11]. The natural gradient method was popularized with Amari's paper [4], which discusses how this method is efficient in learning. In this setting, the ordinary gradient is replaced by a Riemannian gradient, which leads to a parameter update rule of the form

$$\theta^{(t+1)} = \theta^{(t)} + \alpha_t G^{-1}(\theta^{(t)}) \nabla L(\theta^{(t)}), \tag{20}$$

where G is the Fisher information [61]. For a given parametric model $\{p_\theta \colon \theta \in \mathbb{R}^d\}$, the Fisher information is defined as

$$G(\theta) = \mathbb{E}_\theta \left[\nabla \log p(X;\theta) \cdot \nabla^\top \log p(X;\theta) \right].$$

Here $\mathbb{E}_\theta[\cdot]$ denotes expectation with respect to the model distribution $p(X;\theta) \equiv p_\theta$. Amari, Kurata, and Nagaoka [7] discuss the statistical meaning of the Fisher metric. The inverse Fisher matrix divided by the number of observations describes the behavior of the expected square error (covariance matrix) of the maximum likelihood estimator.

For an exponential family model with sufficient statistics $F \colon \mathscr{X} \to \mathbb{R}^d$ and log-partition function $\psi(\theta) = \log Z(\theta)$, the Fisher matrix can be given as the Hessian of the log-partition function, as

$$G(\theta) = \nabla\nabla^\top \psi(\theta) = \mathbb{E}_\theta[F \cdot F^\top] - \mathbb{E}_\theta[F] \cdot \mathbb{E}_\theta[F]^\top = \mathrm{Cov}_\theta[F],$$

which is the covariance of F with respect to the exponential family distribution. This matrix is full rank iff the exponential family parametrization is minimal, meaning that the functions $F_1, \ldots, F_d \colon \mathscr{X} \to \mathbb{R}$ are linearly independent and do not contain the constant function 1 in their linear span.

Consider now the RBM model as the set of visible marginals of an exponential family with sufficient statistics $F \colon \mathscr{X} \times \mathscr{Y} \mapsto \mathbb{R}^d$. The gradient of the visible log-probabilities is

$$\nabla \log p(x;\theta) = \mathbb{E}_\theta[F|x] - \mathbb{E}_\theta[F], \tag{21}$$

where $\mathbb{E}_\theta[F|x] = \sum_y F(x,y) p(y|x;\theta)$ is the conditional expectation of F, given the visible state x, and $\mathbb{E}_\theta[F] = \sum_{x,y} F(x,y) p(x,y;\theta)$ is the expectation with respect to the joint distribution over visible and hidden states. The Fisher matrix takes the form

$$\begin{aligned} G(\theta) &= \mathbb{E}_\theta[\mathbb{E}_\theta[F|X] \cdot \mathbb{E}_\theta[F|X]^\top] - \mathbb{E}_\theta[F] \cdot \mathbb{E}_\theta[F]^\top \\ &= \mathrm{Cov}_\theta[\mathbb{E}_\theta[F|X]]. \end{aligned}$$

The rank of this matrix is equal to the rank of the Jacobian $J(\theta) = [\nabla p(x;\theta)]_x$ of the parametrization of the visible marginal distributions. Verifying whether and when

the Fisher matrix of the RBM has full rank, is a non-trivial problem that we will discuss further in Section 5.

In models with hidden variables, the Fisher matrix is not always full rank. An area that studies the statistical effects of this is *singular learning theory*; see [75, 10]. In practice, for the purpose of parameter optimization, the natural gradient works well even when the model involves singularities, at least so long as the parameter updates don't step into the singular set. The advantages of the natural gradient over the regular gradient have been demonstrated in numerous applications. It tends to be better at handling plateaus, thus reducing the number of required parameter updates, and also to find better local optimizers. On the other hand, computing the Fisher matrix and its inverse is challenging for large systems. Approximations of the relevant expectation values still require a computational overhead over the regular gradient, and in some cases, it is not clear how to balance optimization with other statistical considerations. Approximating the Fisher matrix in an efficient and effective way is an active topic of research. RBMs have been discussed specifically in [60, 31]. Following the notions of the natural gradient, recent works also investigate alternatives and variants of the Fisher metric, for instance based on the Wasserstein metric [48, 44].

Double minimization, EM algorithm

Amari [6, Section 8.1.3] discusses an alternative view on the maximum likelihood estimation problem in probability models with hidden variables. See also [7, 16]. The idea is to regard this as an optimization problem over the model of joint distributions of both visible and hidden variables. Given an empirical data distribution p_V over visible states $x \in \mathscr{X}$, consider the set of joint distributions over $(x,y) \in \mathscr{X} \times \mathscr{Y}$ that are compatible with p_V:

$$E = \Big\{ p(x,y) \colon \sum_{y \in \mathscr{Y}} p(x,y) = p_V(x) \Big\}.$$

This *data manifold* E, being defined by linear equality constraints, is a special type of linear model. Note that it can be written as $E = \{p(x,y) = p_V(x)p(y|x)\}$, where we fix the marginal distribution $p_V(x)$ and are free to choose arbitrary conditional distributions $p(y|x)$ of hidden states given the visible states.

Taking this view, we no longer minimize the divergence from p_V to our model M_V of visible marginal distributions $q_V(x;\theta) = \sum_y q(x,y;\theta)$, but rather we seek for the distributions $q(x,y;\theta)$ in the model M of joint distributions, with the smallest divergence from the data manifold E. The situation is illustrated schematically in Figure 5.

When working with the data manifold E and the joint model M, the maximum likelihood estimation problem becomes a double minimization problem

$$\min_{p \in E, q \in M} D(p \| q). \tag{22}$$

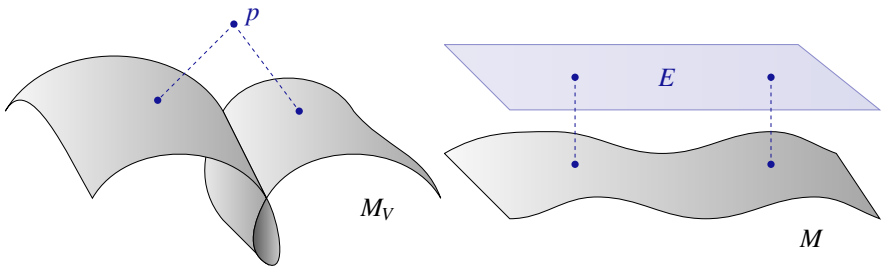

Fig. 5 Schematic illustration of the maximum likelihood estimation problem over the set of visible marginal distributions M_V, and over the set of joint distributions M prior to marginalization.

The minimum of this problem equals the minimum of the original problem

$$\min_{q_V \in M_V} D(p_V \| q_V).$$

To see this, use the chain rule for probability, $P(x,y) = P(x)P(y|x)$, to write

$$\begin{aligned}
\min_{p\in E, q\in M} D(p\|q) &= \min_{p\in E, q\in M} \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{q(x,y)} \\
&= \min_{q\in M} \sum_x p_V(x) \log \frac{p_V(x)}{q_V(x)} + \min_{p(y|x)} \sum_x p_V(x) \sum_y p(y|x) \log \frac{p(y|x)}{q(y|x)} \\
&= \min_{q_V \in M_V} D(p_V \| q_V).
\end{aligned}$$

For simplicity of exposition, we are assuming that the sets E and M are so that the minimum can be attained, e.g., they are closed.

The expression (22) hints at an approach to computing the minimizers. Namely, we can iteratively minimize with respect to each of the two arguments.

- For any fixed value of the second argument, $q \in M$, minimization of the divergence over the first argument $p \in E$ is a convex problem, because E is a linear model. This is solved by the *e-projection* of q onto E, which is given simply by setting $p(y|x) = q(y|x)$.
- For any fixed value of the first argument, $p \in E$, the minimization over the second argument $q \in M$ is also a convex problem, because M is an exponential family. It is solved by the *m-projection* of p onto M, which is given by the unique distribution q in M for which $\sum_{x,y} F(x,y)q(x,y) = \sum_{x,y} F(x,y)p(x,y)$.

This procedure corresponds to the expectation maximization (EM) algorithm [19].

Optimization landscape

In general, for a model with hidden variables, we must assume that the log-likelihood function $L(\theta)$ is non-concave. Gradient methods and other local techniques, such as contrastive divergence and EM, may only allow us to reach critical points or locally optimal solutions. The structure of the optimization landscape and critical points of these methods is the subject of current studies. In Section 8 we discuss results from [68] showing that an RBM model can indeed have several local optimizers with different values of the likelihood function, but also that in some cases, the optimization problem may be solvable in closed form.

5 Dimension

From a geometric standpoint, a basic question we are interested in, is the dimension of the set of distributions that can be represented by our probability model. The dimension is useful when comparing a model against other models, or when testing hypotheses expressed in terms of equality constraints. Under mild conditions, if the dimension is equal to the number of parameters, then the Fisher matrix is regular almost everywhere and the model is generically locally identifiable.

A Boltzmann machine with all units observed is an exponential family, and its dimension can be calculated simply as the dimension of the linear space spanned by the sufficient statistics, disregarding constant functions. This is precisely equal to the number of parameters of the model, since the statistics associated with each of the parameters, bias and interaction weights, are linearly independent.

When some of the units of the Boltzmann machine are hidden, as is usually the case, the set of observable distributions is no longer an exponential family, but rather a linear projection of an exponential family. The marginalization map takes the high dimensional simplex $\Delta_{\mathscr{X}\times\mathscr{Y}}$ to the low dimensional simplex $\Delta_{\mathscr{X}}$. Such a projection can in principle collapse the dimension of the set that is being projected. A simple example where this happens is the set of product distributions. The visible marginals of an independence model are simply the independent distributions of the observed variables, meaning that the hidden variables and their parameters do not contribute to the dimension of the observable model. Another well-known example is the set of mixtures of three product distributions of four binary variables. This model has dimension 13, instead of 14 that one would expect from the number of model parameters. Computing the dimension of probability models with hidden variables often corresponds to challenging problems in algebraic geometry, most prominently the dimension of secant varieties, which correspond to mixture models.

Tropical approach

The first investigation of the dimension of the RBM model was by Cueto, Morton, and Sturmfels [17], using tools from tropical geometry and secant varieties. The tropical approach to the dimension of secant varieties was proposed by Draisma [20]. It can be used in great generality, and it was also used to study non-binary versions of the RBM [53].

As mentioned in Section 3, the tropical RBM consists of piecewise linear approximation of the log-probability vectors of the RBM. The dimension of the tropical RBM is often easy to estimate by combinatorial arguments. A theorem by Bieri and Groves [13, 20] implies that the dimension of the tropical RBM model is a lower bound on the dimension of the original RBM model. Using this method, [17] proved that the RBM model has the expected dimension for most combinations of n and m. However, a number of cases were left open. In fact, for the tropical RBM those cases are still open. A different approach to the dimension of RBMs was proposed in [55], which allowed verifying the conjecture that it always has the expected dimension. In the following we discuss this approach and how it compares to the tropical approach.

Jacobian rank of RBMs and mixtures of products

The dimension of a smoothly parametrized model can be computed as the maximum rank of the Jacobian of the parametrization. For a parametrization $p(x;\theta) = \sum_y p(x,y;\theta)$, with $p(x,y;\theta) = \frac{1}{Z(\theta)}\exp(\sum_i \theta^\top F(x,y))$, the columns of the Jacobian matrix are

$$J_{:x}(\theta) = \sum_y p(x,y;\theta)(F(x,y) - \sum_{x',y'} p(x',y';\theta)F(x',y')), \quad x \in \mathscr{X}. \tag{23}$$

Now we need to consider the specific F and evaluate the maximum rank of the matrix J over the parameter space. In order to simplify this, one possibility is to consider the limit of large parameters θ. The corresponding limit distributions usually have a reduced support and the sum in (23) has fewer nonzero terms. As shown in [55], the dimension bounds from the tropical approach can be obtained in this manner. On the other hand, it is clear that after taking such limits, it is only possible to lower bound the maximum rank. Another problem is that, when the number of parameters is close to the cardinality of $\mathscr{X}$, the rank of the limit matrices is not always easy to compute, with block structure arguments leading to challenging combinatorial problems, such as accurately estimating the maximum cardinality of error correcting codes.

For the analysis it is convenient to work with the denormalized model, which includes all positive scalar multiples of the probability distributions. The dimension of the original model is simply one less. Following (23), and as discussed in [55], the Jacobian for the denormalized RBM is equivalent to the matrix with columns

$$\sum_y p(y|x;\theta)F(x,y) = \sum_y p(y|x;\theta)\hat{x}\otimes\hat{y} = \hat{x}\otimes\hat{\sigma}(Wx+c), \quad x\in\mathscr{X}, \tag{24}$$

where we write $\hat{v} = (1, v^\top)^\top$ for the vector v with an additional 1. Here $\sigma(\cdot) = \exp(\cdot)/(1+\exp(\cdot))$ can be regarded as the derivative of the soft-plus function $\log(1+\exp(\cdot))$. The jth coordinate of $\sigma(Wx+c)$ ranges between 0 and 1, taking larger values the farther x lies in the positive side of the hyperplane $H_j = \{r\in\mathbb{R}^V : W_{j:}r + c_j = 0\}$. In the case of the tropical RBM, the Jacobian is equivalent to the matrix with columns

$$\hat{x}\otimes\hat{\mathbb{1}}_{[Wx+c]_+}, \quad x\in\mathscr{X},$$

where now $\mathbb{1}_{[\cdot]_+}$ corresponds to the derivative of the rectification non-linearity $[\cdot]_+$. The jth coordinate indicates whether the point x lies on the positive side of the hyperplane H_j. The matrices for the RBM and the tropical RBM are illustrated in Figure 6.

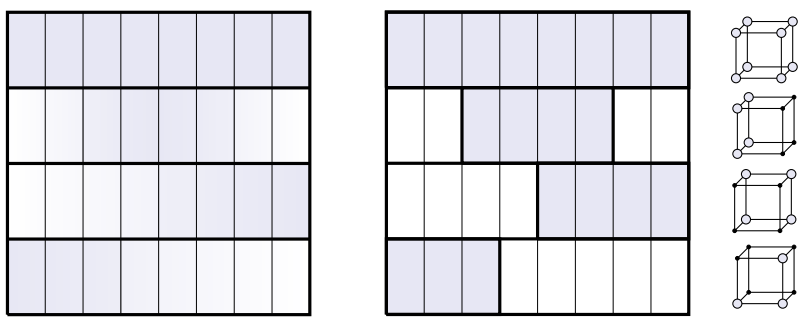

Fig. 6 Illustration of the Jacobian matrix for an RBM with three visible and three hidden units, and its tropical counterpart, together with the corresponding slicings of the visible sufficient statistics. Rows correspond to model parameters and columns to visible states.

In [55] it is shown that (24) can approximate the following matrix, equivalent to the Jacobian of a mixture of $m+1$ product distributions model, arbitrarily well at generic parameters:

$$\hat{x}\otimes\hat{\sigma}'(\tilde{W}x+\tilde{c}), \quad x\in\mathscr{X}.$$

Here $\sigma'(\tilde{W}x+\tilde{c}) = \frac{\exp(\tilde{W}x+\tilde{c})}{\sum_j \exp(\tilde{W}_{j:}x+\tilde{c}_j)}$ is a soft-max unit. In turn, the dimension of the RBM model is bounded below by the dimension of the mixture model. But the results from [15] imply that mixture models of binary product distributions have the expected dimension (except in one case, which for the RBM can be verified by other means). This implies that the RBM model always has the expected dimension:

Theorem 1 ([55, Corollary 26]). *For any $n,m\in\mathbb{N}$ the model* $\mathrm{RBM}_{n,m}$*, with n visible and m hidden binary units, has dimension* $\min\{2^n-1,(n+1)(m+1)-1\}$.

This result implies that, unless the number of parameters exceeds 2^n-1, almost every probability distribution in the RBM model can be represented by only finitely

many different choices of the parameters. One trivial way in which the parameters are not unique, is that we can permute the hidden units without changing the represented distributions, $\sum_{j\in H}\log(1+\exp(w_j x + c_j)) = \sum_{j\in H}\log(1+\exp(w_{\pi(j)}x + c_{\pi(j)}))$ for all $\pi \in H!$. On the other hand, there are also a few probability distributions that can be represented by infinitely many different choices of the parameters. For instance, if $w_j = 0$, then the choice of c_j is immaterial.

The characterization of the parameter fibers $\{\theta \in \mathbb{R}^d \colon p_\theta = p\}$ of the distributions p that can be represented by an RBM model is an important problem, with implications on the parameter optimization problem, which still requires more investigation. We can ask in the first place whether a given distribution p can be represented by an RBM model. We discuss this in the next section.

6 Representational power

The representational power of a probability model can be studied from various angles. An idea is that each parameter allows us to model certain features or properties of the probability distributions. The question then is how to describe and interpret these features. As we have seen, each hidden unit of an RBM can be interpreted as contributing entrywise multiplicative factors which are arbitrary mixtures of two product distributions. Alternatively, each hidden unit can be interpreted as adding a soft-plus unit to the negative energy function of the visible distributions.

Now we want to relate these degrees of freedom with the degrees of freedom of other families of distributions for which we have a good intuition, or for which we can maximize the likelihood function in closed form and compute metrics of the representational power, such as the maximum divergence. The natural approach to this problem is by showing that there exist choices of parameters for which the model realizes a given distribution of interest, or, more generally, a class of distributions of interest. We note that another approach, which we will discuss in Section 8, is by showing that any constraints that apply on the set of distributions from the RBM are less stringent than the constraints that apply on the distributions of interest.

Overview

The representational power of RBMs has been studied in many works. Le Roux and Bengio [43] showed that each hidden unit of an RBM can model the probability of one elementary event. Freund and Haussler [28] used similar arguments to discuss universal approximation. In [51] it was shown that each hidden unit can model the probability of two elementary events of Hamming distance one, which implied improved bounds on the minimal number of hidden units that is sufficient for universal approximation. Generalizing this, [58] showed that each hidden unit can model a block of elementary events with a weighted product distribution, provided certain conditions on the support sets are satisfied. Another line of ideas was

due to [77], showing that each hidden unit can model the coefficient of a monomial in a polynomial representation of the energy function. This analysis was refined in [57], showing that each hidden unit can model the coefficients of as many as n monomials in the energy function.

Mixtures of products and partition models

We discuss a result from [58] showing that an RBM with m hidden units can represent mixtures of $m+1$ product distributions, provided the support sets of m of the mixture components are disjoint. The support of a distribution p on $\mathscr{X}$ is $\operatorname{supp}(p) := \{x \in \mathscr{X} : \, p(x) > 0\}$. The idea is as follows. Consider an entrywise product of the form

$$p_0(x)(1+\lambda p_1(x)) = p_0(x) + \lambda p_0(x) p_1(x), \quad x \in \mathscr{X}. \tag{25}$$

If p_0 and p_1 are product distributions, then so is $p_2 = p_0 p_1$. This is a direct consequence of the fact that the set of product distributions has an affine set of exponential parameters, $\exp(w_0^\top x)\exp(w_1^\top x) = \exp((w_0+w_1)^\top x) = \exp(w_2^\top x)$. In turn, an entrywise product of the form (25) expresses a linear combination of product distributions, provided that p_0 and p_1 are product distributions. The last requirement can be relaxed to hold only over the intersection of the support sets of p_0 and p_1, since the entrywise product will vanish on the other entries either way. When we renormalize, the linear combination becomes a mixture of product distributions, whereby the relative mixture weights are controlled by λ.

Now recall from Section 3 that the RBM distributions can be written as

$$p(x;\theta) = \frac{1}{Z(\theta)} \exp(b^\top x) \prod_{j \in H} (1 + \exp(c_j)\exp(W_{j:}x)). \tag{26}$$

By the previous discussion, we can interpret each factor in (26) as adding a mixture component $p_j(x) = \frac{1}{Z}\exp(W_{j:}x)$, which is a product distribution, so long as the distribution obtained from the preceding factors is a product distribution over the support of p_j. Being an exponential family distribution, p_j has full support, but it can approximate product distributions with restricted support arbitrarily well.

A similar discussion applies to non-binary variables, as shown in [53]. We denote by $\mathrm{RBM}_{\mathscr{X},\mathscr{Y}}$ the RBM with visible states $\mathscr{X} = \mathscr{X}_1 \times \cdots \times \mathscr{X}_n$ and hidden states $\mathscr{Y} = \mathscr{Y}_1 \times \cdots \times \mathscr{Y}_m$. This is the set of marginals of the exponential family with sufficient statistics given by the Kronecker product of the statistics of the independence models on $\mathscr{X}$ and $\mathscr{Y}$, respectively.

Theorem 2 ([49, Theorem 3]). *Let $\mathscr{X} = \mathscr{X}_1 \times \cdots \times \mathscr{X}_n$ and $\mathscr{Y} = \mathscr{Y}_1 \times \cdots \times \mathscr{Y}_m$ be finite sets. The model $\mathrm{RBM}_{\mathscr{X},\mathscr{Y}}$ can approximate any mixture distribution $p(x) = \sum_{i=0}^m \lambda_i p_i(x)$, $x \in \mathscr{X}$, arbitrarily well, where p_0 is any product distribution, and p_i are respectively for all $i \in [m]$, any mixtures of $(|\mathscr{Y}_i| - 1)$ product distributions, with support sets satisfying $\operatorname{supp}(p_i) \cap \operatorname{supp}(p_j) = \emptyset$ for all $1 \le i < j \le m$.*

In particular, the binary $\mathrm{RBM}_{n,m}$ can approximate, to within any desired degree of accuracy, any mixture of $m+1$ product distributions with disjoint supports. Given a collection of disjoint sets $A_1,\dots,A_{m+1} \subseteq \mathscr{X}$, the set of mixtures $p = \sum_j \lambda_j p_j$, where each p_j is a product distribution with support set A_j, is an exponential family on $\cup_j A_j$. More precisely, its topological closure coincides with that of an exponential family with sufficient statistics $\mathbb{1}_{A_j}$, $\mathbb{1}_{A_j} x_i$, $i = 1,\dots,n$, $j = 1,\dots,m+1$. Theorem 2 shows that an RBM can represent all such exponential families, for all choices of disjoint sets $A_1,\dots,A_{m+1}$.

A *partition model* is a special type of mixture model, consisting of all mixtures of a fixed set of uniform distributions on disjoint support sets. Partition models are interesting not only because of their simplicity, but also because they are optimally approximating exponential families of a given dimension. If all support sets of the components, or blocks, have the same size, then the partition model attains the smallest uniform approximation error, measured in terms of the Kullback-Leibler divergence, among all exponential families that have the same dimension [63]. The previous theorem shows that RBMs can approximate certain partition models arbitrarily well. In particular we have:

Corollary 3. *Let $\mathscr{X} = \mathscr{X}_1 \times \cdots \times \mathscr{X}_n$ and $\mathscr{Y} = \mathscr{Y}_1 \times \cdots \times \mathscr{Y}_m$ be finite sets. Let $\mathscr{P}$ be the partition model with partition blocks $\{x_1\} \times \cdots \times \{x_k\} \times \mathscr{X}_{k+1} \times \cdots \times \mathscr{X}_n$ for all $(x_1,\dots,x_k) \in \mathscr{X}_1 \times \cdots \times \mathscr{X}_k$. If $1 + \sum_{j\in[m]} (|\mathscr{Y}_j| - 1) \geq (\prod_{i\in[k]} |\mathscr{X}_i|) / \max_{j\in[k]} |\mathscr{X}_j|$, then each distribution contained in $\mathscr{P}$ can be approximated arbitrarily well by distributions from* $\mathrm{RBM}_{\mathscr{X},\mathscr{Y}}$.

Hierarchical models

Intuitively, each hidden unit of an RBM should be able to mediate certain interactions between the visible units. To make this more concrete, we may ask which distributions from a hierarchical model can be expressed in terms of an RBM, or which parameters of a hierarchical model can be modeled in terms of the hidden units of an RBM. Younes [77] showed that a binary hierarchical model with a total of K pure higher order interactions can be modeled by an RBM with K hidden units. Later, [57] showed that each hidden unit of an RBM can model several parameters of a hierarchical model simultaneously.

Consider a set $S \subseteq 2^V$ of subsets of V. A hierarchical model with interactions S is defined as the set of probability distributions p that can be factorized as

$$p(x) = \prod_{\lambda \in S} \psi_\lambda(x), \quad x \in \mathscr{X}, \tag{27}$$

where each $\psi_\lambda : \mathscr{X} \to \mathbb{R}_+$ is a positive valued function that only depends on the coordinates λ, i.e., satisfies $\psi_\lambda(x) = \psi_\lambda(x')$ whenever $x_i = x'_i$ for all $i \in \lambda$. In practice, we choose a basis to express the potentials as parametrized functions. The set S is conveniently defined as the set of cliques of a graph $G = (V,E)$, and hence these models are also known as hierarchical graphical models. These models are

very intuitive and have been studied in great detail. Each factor ψ_λ is interpreted as allowing us to model arbitrary interactions between the variables x_i, $i \in \lambda$, independently of the variables x_j, $j \in V \setminus \lambda$. Hence, they are a good reference to compare the representational power other models, which is what we want to do for RBMs in the following.

At a high level, the difficulty of comparing RBMs and hierarchical models stems from the fact that their parameters contribute different types of degrees of freedom. While a hidden unit can implement interactions among all visible units it is connected to, certain constraints apply on the values of these interactions. For example, the set of interaction coefficients among two visible variables that can be modeled by one hidden unit is shown in Figure 7.

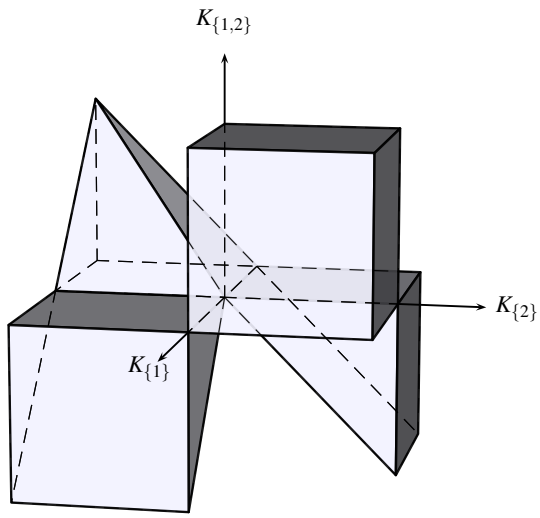

Fig. 7 Interaction coefficients expressible by one RBM hidden unit. Shown is the set of coefficients $(K_{\{1\}}, K_{\{2\}}, K_{\{1,2\}}) \in \mathbb{R}^3$, clipped to a cube centered at the origin, of the polynomials $K_\emptyset + K_{\{1\}}x_1 + K_{\{2\}}x_2 + K_{\{1,2\}}x_1x_2$ expressible in terms of a soft-plus unit on binary inputs. Figure adapted from [57].

To proceed with more details, we first fix a coordinate system. Hierarchical models are conveniently expressed in terms of a basis of orthogonal functions known as characters. For each $\lambda \subseteq V$ we have a function

$$\sigma_\lambda(x) = \prod_{i \in \lambda} (-1)^{x_i}, \quad x \in \{0,1\}^V.$$

The functions σ_λ, $\lambda \subseteq V$, are orthogonal, with $\sum_x \sigma_\lambda(x)\sigma_\mu(x) = 2^n \delta_{\lambda,\mu}$. In turn, we can express any given vector $l \in \mathbb{R}^{\{0,1\}^V}$ as

$$l(x) = \sum_{\lambda \subseteq V} J_\lambda \sigma_\lambda(x), \quad x \in \{0,1\}^V,$$

where the coefficients are given by

$$J_\lambda = \frac{1}{2^n} \sum_{x \in \{0,1\}^V} \sigma_\lambda(x) l(x), \quad \lambda \subseteq V.$$

The change of coordinates from the standard basis δ_x, $x \in \{0,1\}^V$, to the basis of characters σ_λ, $\lambda \subseteq V$, can be interpreted as a Möbius inversion, or also as a Fourier transform.

If we replaced the states $\{0,1\}$ with $\{+1,-1\}$, we could write each σ_λ as a monomial $\prod_{i\in\lambda} x_i$. But we can also use a basis of monomials without changing the states. For each $\lambda \subseteq V$, let

$$\pi_\lambda(x) = \prod_{i\in\lambda} x_i, \quad x \in \{0,1\}^V. \tag{28}$$

Although this is no longer an orthogonal basis, it is conceptually simple and very frequently used in practice. Moreover, for an inclusion closed set $S \subseteq 2^V$, the span of π_λ, $\lambda \in S$, equals that of σ_λ, $\lambda \in S$, such that both bases have the same hierarchical coordinate sub-spaces.

For an inclusion closed set $S \subseteq 2^V$, the binary hierarchical model with interactions S can be parametrized as the exponential family $\mathscr{E}_S$ of distributions of the form

$$p(x) = \frac{1}{Z} \exp\Big(\sum_{\lambda\in S} J_\lambda \prod_{i\in\lambda} x_i \Big), \quad x \in \{0,1\}^V, \tag{29}$$

with parameters $J_\lambda \in \mathbb{R}$, $\lambda \in S$.

Now we proceed with the representation of the parameters of a hierarchical model in terms of an RBM. Recall that the log-probabilities $l = \log(p)$ in the model $\mathrm{RBM}_{n,m}$ are sums of a linear unit and m soft-plus units. For a linear unit $w^\top x + c$, the polynomial coefficients are simply $K_\emptyset = c$, $K_{\{i\}} = w_i$, $i \in V$, and $K_\lambda = 0$ for all $\lambda \subseteq V$ with $|\lambda| \geq 2$. For a soft-plus unit, [57] obtains a partial characterization of the possible polynomial coefficients. In particular, it shows the following.

Lemma 4 ([57, Lemma 5]). *Consider a subset $B \subseteq V$, and let $J_{B\cup\{j\}} \in \mathbb{R}$, $j \in V \setminus B$, and $\varepsilon > 0$. Then there are $w \in \mathbb{R}^V$ and $c \in \mathbb{R}$ such that the soft-plus unit $\log(1 + \exp(w^\top x + c))$ is equal to a polynomial $\sum_\lambda K_\lambda \prod_{i\in\lambda} x_i$ with coefficients satisfying $|K_{B\cup\{j\}} - J_{B\cup\{j\}}| \leq \varepsilon$ for all $j \in V \setminus B$, and $|K_C| \leq \varepsilon$ for all $C \neq B, B\cup\{j\}$, $j \in V \setminus B$.*

This says that each hidden unit of an RBM can model arbitrarily the parameters of a hierarchical model corresponding to the monomials that cover $\prod_{i\in B} x_i$, for any fixed choice of $B \subseteq V$, while at the same time setting all other parameters arbitrarily close to zero, except for the parameter associated with $\prod_{i\in B} x_i$, whose value may be coupled to the values of the other parameters.

We can use this result to describe hierarchical models that can be represented by an RBM. Since each hidden unit of the RBM can model certain subsets of parameters of hierarchical models, we just need to find a sufficiently large number of hidden units which together can model all the required parameters. For example:

- $\mathrm{RBM}_{3,1}$ contains the hierarchical models $\mathscr{E}_S$ with $S = \{\{1\},\{2\},\{3\},\{1,2\},\{1,3\}\}$, $S = \{\{1\},\{2\},\{3\},\{1,2\},\{2,3\}\}$, $S = \{\{1\},\{2\},\{3\},\{1,3\},\{2,3\}\}$. It does not contain the *no-three-way interaction model*, with $S = S_2 = \{\{1\},\{2\},\{3\},\{1,2\},\{1,3\},\{2,3\}\}$.
- The model $\mathrm{RBM}_{3,2}$ contains the no-three-way interaction model $\mathscr{E}_S$ with $S = S_2$. It does not contain the full interaction model, with $S = S_3$. In particular, this model is not a universal approximator.

In general, finding a minimal cover of the relevant set of parameters of hierarchical models in terms of subsets of parameters of the form described in Lemma 4 relates to well-known problems in the theory of combinatorial designs. For S consisting of all sets up to a given cardinality, we can obtain the following bounds.

Theorem 5 ([57, Theorem 11]). *Let $1 \leq k \leq n$ and $\mathscr{X} = \{0,1\}^V$. Every distribution from the hierarchical model $\mathscr{E}_{S_k}$, with $S_k = \{\lambda \subseteq V : |\lambda| \leq k\}$, can be approximated arbitrarily well by distributions from $\mathrm{RBM}_{n,m}$ whenever*

$$m \geq \min\Big\{ \sum_{j=2}^{k} \binom{n-1}{j-1}, \frac{\log(n-1)+1}{n+1} \sum_{j=2}^{k} \binom{n+1}{j} \Big\}.$$

We note that in specific cases there are sharper bounds available, listed in [57].

The hidden units and parameters of an RBM can be employed to model different kinds of hierarchical models. For instance, a limited number of hidden units could model the set of full interactions among a small subset of visible variables, or, alternatively, to model all k-wise interactions among a large set of visible units. Exactly characterizing the largest hierarchical models that can be represented by an RBM is still an open problem for $n \geq 4$.

Universal approximation

The universal approximation question asks for the smallest model within a class of models, which is able to approximate any given probability distribution on its domain to within any desired degree of accuracy. This is a special case of the problems discussed in the previous paragraphs. A direct consequence of Theorem 11 is

Corollary 6. *Let $\mathscr{X} = \mathscr{X}_1 \times \cdots \times \mathscr{X}_n$ and $\mathscr{Y} = \mathscr{Y}_1 \times \cdots \times \mathscr{Y}_m$ be finite sets. The model $\mathrm{RBM}_{\mathscr{X},\mathscr{Y}}$ is a universal approximator whenever*

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

When all units are binary, this implies that an RBM with $2^{n-1}-1$ hidden units is a universal approximator of distributions on $\{0,1\}^n$. Theorem 5 improves this bound as follows:

Corollary 7 ([57, Corollary 12]). *Every distribution on $\{0,1\}^n$ can be approximated arbitrarily well by distributions from $\mathrm{RBM}_{n,m}$ whenever*

$$m \geq \min\left\{2^{n-1}-1, \frac{2(\log(n-1)+1)}{n+1}(2^n-(n+1)-1)+1\right\}.$$

This is the sharpest general upper bound that is available at the moment. A slightly looser but simpler bound is $\frac{2(\log(n)+1)}{n+1}2^n - 1$. Again, in specific cases there are sharper bounds available, listed in [57].

In terms of the necessary number of hidden units for universal approximation, bounds have been harder to obtain. In the general case, we only have lower bounds coming from parameter counting arguments:

Proposition 8. *Let $\mathscr{M}$ be an exponential family over $\mathscr{X} \times \mathscr{Y}$ and $\mathscr{M}_V$ the set of marginals on $\mathscr{X}$. If $\mathscr{M}_V$ is a universal approximator, then $\mathscr{M}_V$ has dimension $|\mathscr{X}|-1$ and $\mathscr{M}$ has dimension at least $|\mathscr{X}|-1$.*

This implies that for $\mathrm{RBM}_{n,m}$ to be a universal approximator, necessarily $m \geq 2^n/(n+1)-1$. There is still a logarithmic gap between the upper and lower bounds. Further closing this gap is an important theoretical problem, which could help us obtain a more complete understanding of the representational power question. In a few small cases we can obtain the precise numbers. For instance, for $n = 2$, the minimal size of a universal approximator is $m = 1$. For $n = 3$ it is $m = 3$. But already for $n = 4$ we can only bound the exact value between 3 and 6.

Relative representational power

As we have seen, RBMs can represent certain mixtures of product distributions. Complementary to this, it is natural to ask how large a mixture of products is needed in order to represent an RBM. Following Section 3, an RBM model consists of tensors which are entrywise products of tensors of with non-negative rank at most two. For many combinations of n and m it turns out that the RBM model represents tensors of the maximum possible rank, 2^m, which implies that the smallest mixture of products that contain the RBM model is as large as one could possibly expect, having 2^m components:

Theorem 9 ([54, Theorem 1.2]). *The smallest k for which the model $\mathscr{M}_{n,k}$, consisting of arbitrary mixtures of k product distributions of n binary variables, contains the model $\mathrm{RBM}_{n,m}$, is bounded by $\frac{3}{4}n \leq \log_2(k) \leq n-1$ when $m \geq n$, by $\frac{3}{4}n \leq \log_2(k) \leq m$ when $\frac{3}{4}n \leq m \leq n$, and satisfies $\log_2(k) = m$ when $m \leq \frac{3}{4}n$.*

As shown in [54] RBMs can express distributions with many more strong modes than mixtures of products with the same number of parameters. A strong mode is a local maximum of the probability distribution, with value larger than the sum of all its neighbors, whereby the vicinity structure is defined by the Hamming distance over the set of elementary events. Distributions with many strong modes have a large non-negative tensor rank. At the same time, [54] shows that an RBM does not always contain a mixture of products model with the same number of parameters. The size of the largest mixture of products that is contained in an RBM is still an open problem.

For hierarchical models, Lemma 4 allows us to formulate an analogous result. The lemma implies that a hidden unit can create non-zero values of any parameter of any arbitrary hierarchical model. In turn, the smallest hierarchical model that contains an RBM must have all possible interactions and hence it is as large as one could possibly expect:

Proposition 10. *Let $n,m \in \mathbb{N}$. The smallest $S \subseteq 2^V$ for which the hierarchical model $\mathscr{E}_S$ on $\{0,1\}^V$ contains* $\mathrm{RBM}_{n,m}$ *is $S = 2^V$.*

7 Divergence bounds

Instead of asking for the sets of distributions that can be approximated arbitrarily well by an RBM, we can take a more refined standpoint and ask for the error in the approximation of a given target distribution. The best possible uniform upper bound on the divergence to a model $\mathscr{M}$ is $D_{\mathscr{M}} = \max_p D(p\|\mathscr{M}) = \max_p \inf_{q\in\mathscr{M}} D(p\|q)$.

Maximizing the divergence to a model, over the set of all possible targets, is an interesting problem in its own right. For instance, the divergence to an independence model is called multi-information and can be regarded as a measure of complexity. The multi-information can be used as an objective function in certain learning problems, as a way to encourage behaviors that are both predictable and diverse. The divergence maximization problem is challenging, even in the case of exponential families with closed formulas for the maximum likelihood estimators. For exponential families models the divergence maximization problem has been studied in particular by Matúš [47], Ay [46], and Rauh [62].

In the case of RBMs, as with most machine learning models used in practice, the situation is further complicated, since we do not have closed formulas for the error minimizers of a given target. The approximation errors of RBMs were studied in [58] by showing that RBMs contain a number of exponential families and providing upper bounds on the divergence to such families. The approach was formulated more generally in [59]. In [56] it was show how to obtain upper bounds on the expected value of the approximation error, when the target distributions are sampled from a given prior. In the following we discuss some of these bounds and also a divergence bound derived from the hierarchical models presented in Section 6.

Upper bounds from unions of mixtures of products and hierarchical models

The Kullback-Leibler divergence from a distribution q to another distribution p is

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

Given some p, we are interested in the best approximation within a given model $\mathscr{M}$. We consider the function that maps each possible target distribution p to

$$D(p\|\mathscr{M}) = \inf_{q\in\mathscr{M}} D(p\|q).$$

The divergence to a partition model $\mathscr{P}_A$ with blocks A_k, $k = 1,\dots,K$, is bounded above by $D(\cdot\|\mathscr{P}_A) \leq \max_k \log|A_k|$. This bound is in fact tight. Corollary 3 shows that RBMs can represent certain partition models. This implies the following bound.

Theorem 11 ([49, Theorem 5]). *Let $\mathscr{X} = \mathscr{X}_1 \times \cdots \times \mathscr{X}_n$ and $\mathscr{Y} = \mathscr{Y}_1 \times \cdots \times \mathscr{Y}_m$ be finite sets. If $1 + \sum_{j\in[m]}(|\mathscr{Y}_j| - 1) \geq |\mathscr{X}_{\Lambda\setminus\{k\}}|$ for some $\Lambda \subseteq [n]$ and $k \in \Lambda$, then*

$$D(\cdot\|\,\mathrm{RBM}_{\mathscr{X},\mathscr{Y}}) \leq \log|\mathscr{X}_{[n]\setminus\Lambda}|.$$

Instead of partition models, we can also consider mixtures of product distributions with disjoint supports, as described in Theorem 2. As discussed in [58] the divergence to a mixture of models with disjoint supports can be bounded tightly from above by the maximum divergence to one of the component models over targets with the same support. Consider a model $\mathscr{M}$ consisting of mixtures $\sum_j \lambda_j p_j$ of distributions $p_j \in \mathscr{M}_j$, where $\mathscr{M}_j$ consists of distributions supported on A_j, and $A_i \cap A_j = \emptyset$ whenever $i \neq j$. Then

$$\max_p D(p\|\mathscr{M}) = \max_j \max_{p:\ \mathrm{supp}(p)\subseteq A_j} D(p\|\mathscr{M}_j).$$

We know that the RBM contains several mixtures of products with disjoint supports. Hence we can further improve the divergence upper bounds by considering the divergence to the union of all the models that are contained in the RBM model. This gives the following bound.

Theorem 12 ([59, Theorem 2]). *If $m \leq 2^{n-1} - 1$,*

$$D(\cdot\|\mathrm{RBM}_{n,m}) \leq \big(n - \lfloor\log_2(m+1)\rfloor - \frac{m+1}{2^{\lfloor\log_2(m+1)\rfloor}}\big)\log(2)\ .$$

A corresponding analysis for RBMs with non-binary units still needs to be worked out.

We can also bound the divergence in terms of the hierarchical models described in Theorem 5, instead of the partition models and mixtures of products mentioned above. Matúš [47] studies the divergence to hierarchical models, and proves, in particular, the following bound.

Lemma 13 ([47, Corollary 3]). *Consider an inclusion closed set $S \subseteq 2^V$ and the hierarchical model $\mathscr{E}_S$ on $\{0,1\}^V$. Then $D(\cdot\|\mathscr{E}_S) \leq \min_{A\in S}\log|\mathscr{X}_{V\setminus A}|$.*

In conjunction with Theorem 7, this directly implies the following bound.

Corollary 14. *Let $n,m \in \mathbb{N}$, and let k be the largest integer with $m \geq \frac{\log(k)+1}{k+1}2^{k+1} - 1$. Then $D(\cdot\|\,\mathrm{RBM}_{n,m}) \leq (n-k)\log(2)$.*

A version of this result for non-binary variables and bounding the divergence to unions of hierarchical models still need to be worked out.

Divergence to polyhedral exponential families

The previous results estimate the divergence to an RBM model by looking at the divergence to exponential families or unions of exponential families that are contained within the RBM model (or within its closure, to be more precise). More generally, we might be interested in estimating the divergence to models whose set of log-probabilities forms a polyhedral shape, as the one shown in Figure 7. Each face of a polyhedron can be extended to an affine space, and hence corresponds to a piece of an exponential family. This allows us to compute the maximum likelihood estimators of a polyhedral family in the following way. A related discussion was conducted recently in [2] in the context of mixtures of products, and in [68] in the context of RBMs.

Given a target distribution p and a model with log-probabilities from a polyhedron $\mathscr{M}$ we proceed as follows.

- For each face $\mathscr{M}_i$ of $\mathscr{M}$, we define a corresponding exponential family $\mathscr{E}_i$. Any basis of the affine hull of $\mathscr{M}_i$ forms a sufficient statistics, and we can take any point in $\mathscr{M}_i$ as a reference measure.
- Then we compute the maximum likelihood estimator $q_i = \operatorname{arginf}_{q\in\mathscr{E}_i} D(p\|q)$ for each individual exponential family $\mathscr{E}_i$. For exponential families the maximum likelihood estimation problem is concave and has a unique solution (possibly on the closure of the exponential family).
- Then we verify which of the projections q_i are feasible, meaning that they satisfy the constraints of the corresponding face $\mathscr{M}_i$.
- Finally, we select among the feasible projections, the one with the smallest divergence to the target distribution p. This is illustrated in Figure 8.

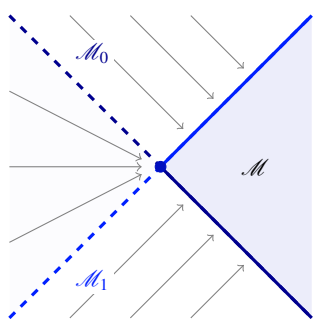

Fig. 8 Illustration of the maximum likelihood projections onto a model whose log-probabilities form a polyhedron. Here the polyhedron $\mathscr{M}$ consists of the points on the positive side of two hyperplanes, $\mathscr{M}_0$ and $\mathscr{M}_1$. Each face of the polyhedron extends to an affine space that corresponds to an exponential family. For each possible target, each exponential family has a unique maximum likelihood projection point. Arrows indicate how targets project to the different faces of $\mathscr{M}$.

Tightness of the bounds

In the previous paragraphs we provided upper bounds on the divergence from arbitrary target distributions to an RBM model. One may wonder about the tightness of these bounds. For the special case of independence models, which are RBMs with no hidden units, the bounds are tight, provided all visible variables have state spaces of equal cardinality. However, already in the case of one single hidden unit, the exact value of the maximum divergence is not known in general.

Experiments on small RBMs [58, 49] seem to indicate that the bounds provided in the previous paragraphs are in good agreement with the actual values. Empirical studies are difficult because of two opposing effects. On the one hand, sequential optimization methods may only lead to sub-optimal approximations of a given target. In fact, part of the motivation for deriving theoretical upper bounds is to monitor the quality of our sequential optimization methods. On the other hand, finding a target distribution with maximum divergence to the model may be a difficult problem itself. It may be that the vast majority of possible targets are not as far to the model as the divergence maximizer. In turn, the theoretical upper bounds could appear pessimistic for most of the targets. In [56] it is shown how to estimate the expected value of the divergence when the target distributions are sampled from a Dirichlet distribution. The average values tend to be indeed much lower than the maximum values.

A recent work [68] shows that the model $\mathrm{RBM}_{3,2}$ has a boundary described in terms of a union of exponential families, and uses this description to obtain the divergence maximizers to the model. It shows that the divergence bounds obtained in Theorem 12 are tight for this particular model.

Theorem 15 ([68, Theorem 3]). *The maximum divergence to* $\mathrm{RBM}_{3,2}$ *is* $\frac{1}{2}\log 2$. *The maximizers are* $\frac{1}{4}(\delta_{000}+\delta_{011}+\delta_{101}+\delta_{110})$ *and* $\frac{1}{4}(\delta_{001}+\delta_{010}+\delta_{100}+\delta_{111})$. *For each of these targets, there is one distinct projection point on each of the six boundary pieces of* $\mathrm{RBM}_{3,2}$.

8 Implicit description

So far we have discussed probability models presented explicitly, as parametric families of distributions. RBMs can also be expressed implicitly, in terms of constraints that apply to the distributions within the model, and only to the distributions within the model. Indeed, since RBMs have a polynomial parametrization, they can be described semi-algebraically as the set of real solutions to a collection of polynomial equations and polynomial inequalities. The *implicitization problem* consists of replacing a parametric description with a description as the solution set of a collection of equations and inequalities. Finding implicit characterizations for graphical models with hidden variables is a significant challenge and a central topic within algebraic statistics [21, 72]. In principle both, explicit and implicit presentations,

can be challenging to interpret in general, for instance when the parametrization is convoluted, or when the constraints correspond to complicated properties of the distributions. However, in some cases the implicit descriptions have a very intuitive statistical interpretation and can allow us to make significant advances over what is possible with a parametric description alone. Implicit descriptions can be extremely useful for hypothesis testing, membership testing, and other related problems. So far there are not many results on the implicit description of RBMs. The following discussion is intended as a motivation.

Markov properties

A fully observable undirected graphical model can be defined in terms of the factorization property (27). Each of the factors can be considered as a parameter, or can be easily parametrized, as shown in (29). Graphical models are usually also motivated and defined in terms of so-called Markov properties, or conditional independence statements. These are constraints that characterize the probability distributions in the model. Undirected graphical models encode conditional independence relations in terms of the structure of the graph. Specifically, a probability distribution is contained in an undirected graphical model with graph G if and only if it satisfies all conditional independence statements encoded by the graph G, namely

$$X_A \perp\!\!\!\perp X_B \mid X_C, \tag{30}$$

whenever A,B,C are disjoint subsets of V for which any path connecting a point in A and a point in B, passes through C. Equation (30) means that p satisfies the equations $p(x_A,x_B|x_C) = p(x_A|x_C)p(x_B|x_C)$, or, equivalently,

$$p(x_A,x_B,x_C) \sum_{x'_A,x'_B} p(x'_A,x'_B,x_C) - \sum_{x'_B} p(x_A,x'_B,x_C) \sum_{x'_A} p(x'_A,x_B,x_C) = 0,$$

for all $x_A \in \mathscr{X}_A$, $x_B \in \mathscr{X}_B$, $x_C \in \mathscr{X}_C$. These are quadratic binomial equations in the indeterminates $p(x) \in \mathbb{R}$, $x \in \mathscr{X}$. A famous theorem by Hammersley and Clifford [32] gives the correspondence between the conditional independence constraints and the factorization property of the joint distributions in a fully observable graphical model. This correspondence is usually limited to strictly positive probability distributions. For distributions that are not strictly positive, which lie at the boundary of the probability simplex, the correspondence is more subtle in general and has been investigated in [29]. The main point here is that we can formulate a parametric set of functions in terms of constraints, or properties of distributions. Moreover, at least in the case of fully observable undirected graphical models, the constraints have an intuitive statistical interpretation.

Constraints in a small RBM

A natural question is what are the constraints that define the visible distributions in a an RBM, and more generally, in a hierarchical model with hidden variables. Aside from RBMs with one single hidden unit, which correspond to mixtures of two product distributions, the RBM with 4 visible 2 hidden and variables has been studied, which turns out to be a hyper-surface defined as the zero set of a polynomial with over a trillion monomials [18].

The constraints that apply to $\mathrm{RBM}_{3,2}$ were studied in [54], obtaining a coarse description of the model. The full semi-algebraic description of this model was then obtained in [68]. The characterization is as follows.

Theorem 16 ([68, Theorem 1]). *The model* $\mathrm{RBM}_{3,2}$ *is the union of six basic semi-algebraic sets, each described by two inequalities, namely:*

$$\begin{aligned}
&\{p_{000}p_{011} \geq p_{001}p_{010}, \quad p_{100}p_{111} \geq p_{101}p_{110}\}\\
&\{p_{000}p_{011} \leq p_{001}p_{010}, \quad p_{100}p_{111} \leq p_{101}p_{110}\}\\
&\{p_{000}p_{101} \geq p_{001}p_{100}, \quad p_{010}p_{111} \geq p_{011}p_{110}\}\\
&\{p_{000}p_{101} \leq p_{001}p_{100}, \quad p_{010}p_{111} \leq p_{011}p_{110}\}\\
&\{p_{000}p_{110} \geq p_{100}p_{010}, \quad p_{001}p_{111} \geq p_{101}p_{011}\}\\
&\{p_{000}p_{110} \leq p_{100}p_{010}, \quad p_{001}p_{111} \leq p_{101}p_{011}\}.
\end{aligned}$$

Each pair of inequalities represents the non-negativity or non-positivity of two determinants. These determinants capture the conditional correlations of two of the variables, given the value of the third variable. The conditional correlation is either non-negative or non-positive for both possible values of the third variable.

This theorem gives a precise description of the geometry of the model. The model is full dimensional in the ambient probability simplex. Hence the description involves only inequalities and no equations (aside from the normalization constraint $\sum_x p_x = 1$). Setting either of the inequalities to an equation gives a piece of the boundary of the model. Each boundary piece is an exponential family which can be interpreted as the set of mixtures of one arbitrary product distribution and one product distribution with support on the states with fixed value of one of the variables, similar to the distributions described in Theorem 2. For these exponential families we can compute the maximum likelihood estimators in closed form, as described in the previous paragraph, and also obtain the exact maximizers of the divergence, given in Theorem 15. With the implicit description at hand [68] also shows that the model $\mathrm{RBM}_{3,2}$ is equal to the mixture model of three product distributions, and that it does not contain any distributions with 4 modes, both statements that had been conjectured in [54].

Coarse necessary constraints

Obtaining the exact constraints that define an RBM model can be difficult in general. In Section 6 we described submodels of the RBM, which can be interpreted

as constraints that are sufficient for probability distributions to be contained in the model, but not necessary. A complementary alternative is to look for constraints that are necessary for distributions to be in the model, but not sufficient. These sometimes are easier to obtain and interpret. An example are strong mode inequalities in mixtures of product distributions [54], and information theoretic inequalities in Bayesian networks [71]. Mode inequality constraints for RBMs have been studied in [54]. Another possible direction was suggested in [68], namely to consider the inequality constraints that apply to mixtures of two product distributions and how they combine when building Hadamard products.

9 Open problems

The theory of RBMs is by no means a finished subject. In the following, I collect a selection of problems, as a sort of work program, addressing which I think is important towards obtaining a more complete picture of RBMs and advancing the theory of graphical models with hidden variables in general.

1. Can we find non-trivial constraints on the sets of representable probability distributions? A related type of questions has been investigated in [54], with focus on the approximation of distributions with many modes, or mixtures of product distributions.
2. Closely related to the previous item, given the number n of visible units, what is the smallest number m of hidden units for which $\mathrm{RBM}_{n,m}$ is a universal approximator? Alternatively, can we obtain lower bounds on the number of hidden units of an RBM that is a universal approximator? Here, of course, we are interested in lower bounds that do not readily follow from parameter counting arguments. The first open case is $n = 4$, for which we have bounds $3 \leq m \leq 6$.
3. What is the smallest tropical RBM that is a universal approximator? Equivalently, what is the smallest m for which a sum of one affine function and m ReLUs can express any function of n binary variables?
4. Characterize the support sets of the distributions in the closure of an RBM. We note that characterizing the support sets of distributions in the closure of an exponential family corresponds to describing the faces its convex support polytope.
5. Also in relation to the first item, obtain an implicit description of the RBM model. The work [68] gives the description of $\mathrm{RBM}_{3,2}$ and ideas for the inequality constraints of larger models. Interesting cases to consider are $\mathrm{RBM}_{4,3}$ (this might be the full probability simplex), $\mathrm{RBM}_{5,2}$, $\mathrm{RBM}_{6,5}$. For the latter [54] obtained some linear inequality constraints.
6. Can we produce explicit descriptions of the maximum likelihood estimators? Here [68] indicates possible avenues.
7. Describe the structure of the likelihood function of an RBM. In particular, what is the number of local and global optimizers? How does this number depend on the empirical data distribution?

8. Describe the critical points of the EM algorithm for an RBM model or for its Zariski closure.
9. Characterize the sets of parameters that give rise to the different distributions expressible by an RBM. When this is finite, are there parameter symmetries other than those coming from relabeling units and states?
10. What is the maximum possible value of the divergence to an RBM model, $D_{n,m} = \max_{p \in \Delta_{\{0,1\}^n}} \inf_{q \in \mathrm{RBM}_{n,m}} D(p\|q)$, and what are the divergence maximizers? We know $D_{3,0} = 2\log 2$ from results for independence models (see, e.g., [59]), and $D_{3,2} = \frac{1}{2}\log 2$ (see Theorem 16 and [68]). The first open case is $D_{3,1}$. Discussions with Johannes Rauh suggest $-\frac{3}{4}\log_2(2\sqrt{3}-3)$.
11. In relation to the previous item, can we provide lower bounds on the maximum divergence from a given union of exponential families?
12. Does the tropical RBM model have the expected dimension? In [17] it was conjectured that it does. The problem remains open, even though [55] gave a proof for the RBM. The description of the tropical RBM as a superposition of ReLUs could be useful here.
13. What is the largest mixture of product distributions that is contained in the RBM model? A result from [54] shows that RBMs do not always contain mixtures of products of the same dimension.
14. What are the largest hierarchical models that are contained in the closure of an RBM model? A partial characterization of the polynomials that are expressible in terms of soft-plus and rectified linear units on binary inputs was obtained in [57]. A full characterization is still missing.
15. Generalize the analysis of hierarchical models contained in RBM models to the case of non-binary variables (both visible and hidden).

Acknowledgements I thank Shun-ichi Amari for inspiring discussions over the years. This review article originated at the IGAIA IV conference in 2016 dedicated to his 80th birthday. I am grateful to Nihat Ay, Johannes Rauh, Jason Morton, and more recently Anna Seigal for our collaborations. I thank Fero Matúš for discussions on the divergence maximization for hierarchical models, lastly at the MFO Algebraic Statistics meeting in 2017. I thank Bernd Sturmfels for many fruitful discussions, and Dave Ackley for insightful discussions at the Santa Fe Institute in 2016.

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 757983).

References

1. D. H. Ackley, G. E. Hinton, and T. J. Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive Science*, pages 147–169, 1985.
2. E. Allman, H. B. Cervantes, R. Evans, S. Hoşten, K. Kubjas, D. Lemke, J. Rhodes, and P. Zwiernik. Maximum likelihood estimation of the latent class model through model boundary decomposition. 2017.
3. S. Amari. *Differential-geometrical methods in statistics*. Lecture notes in statistics. Springer-Verlag, 1985.

4. S. Amari. Natural gradient works efficiently in learning. *Neural Comput.*, 10(2):251–276, Feb. 1998.
5. S. Amari. Information geometry on hierarchical decomposition of stochastic interactions. *IEEE Transaction on Information Theory*, 47:1701–1711, 1999.
6. S. Amari. *Information Geometry and its Applications*, volume 194 of *Applied Mathematical Sciences*. Springer Japan, 2016.
7. S. Amari, K. Kurata, and H. Nagaoka. Information geometry of Boltzmann machines. *IEEE Transactions on Neural Networks*, 3(2):260–271, Mar 1992.
8. S. Amari and H. Nagaoka. *Methods of Information Geometry*. Translations of mathematical monographs. American Mathematical Society, 2007.
9. M. H. Amin, E. Andriyash, J. Rolfe, B. Kulchytskyy, and R. Melko. Quantum Boltzmann machine. *Phys. Rev. X*, 8:021050, May 2018.
10. M. Aoyagi. Stochastic complexity and generalization error of a Restricted Boltzmann Machine in Bayesian estimation. *Journal of Machine Learning Research*, 99:1243–1272, August 2010.
11. N. Ay, J. Jost, H. Lê, and L. Schwachhöfer. *Information Geometry*, volume 64 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer, 2017.
12. Y. Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009. Also published as a book. Now Publishers, 2009.
13. R. Bieri and J. Groves. The geometry of the set of characters iduced by valuations. *Journal für die reine und angewandte Mathematik*, 347:168–195, 1984.
14. L. Brown. *Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory*. Institute of Mathematical Statistics, Hayworth, CA, USA, 1986.
15. M. Catalisano, A. 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.
16. I. Csiszár and G. Tusnády. Information Geometry and Alternating minimization procedures. *Statistics and Decisions*, Supplement Issue 1, 1984.
17. M. A. Cueto, J. Morton, and B. Sturmfels. Geometry of the restricted Boltzmann machine. In M. A. G. Viana and H. P. Wynn, editors, *Algebraic methods in statistics and probability II, AMS Special Session*, volume 2. American Mathematical Society, 2010.
18. M. A. Cueto, E. A. Tobis, and J. Yu. An implicitization challenge for binary factor analysis. *Journal of Symbolic Computation*, 45(12):1296–1315, 2010.
19. 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.
20. J. Draisma. A tropical approach to secant dimensions. *J. Pure Appl. Algebra*, 212(2):349–363, 2008.
21. M. Drton, B. Sturmfels, and S. Sullivant. *Lectures on Algebraic Statistics*. Oberwolfach Seminars. Springer Verlag, 2009.
22. A. Fischer and C. Igel. Contrastive divergence learning may diverge when training restricted Boltzmann machines. In *Frontiers in Computational Neuroscience. Bernstein Conference on Computational Neuroscience (BCCN 2009)*, 2009.
23. A. Fischer and C. Igel. Bounding the bias of contrastive divergence learning. *Neural Computation*, 23(3):664–673, 2010.
24. 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*, pages 14–36, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.
25. A. Fischer and C. Igel. Training restricted Boltzmann machines: an introduction. *Pattern Recognition*, 47(1):25–39, 2014.
26. A. Fischer and C. Igel. A bound for the convergence rate of parallel tempering for sampling restricted Boltzmann machines. *Theoretical Computer Science*, 598:102 – 117, 2015.
27. Y. Freund and D. Haussler. Unsupervised learning of distributions on binary vectors using two layer networks. In J. E. Moody, S. J. Hanson, and R. P. Lippmann, editors, *Advances in Neural Information Processing Systems 4*, pages 912–919. Morgan-Kaufmann, 1992.

28. Y. Freund and D. Haussler. Unsupervised learning of distributions on binary vectors using two layer networks. Technical report, Santa Cruz, CA, USA, 1994.
29. D. Geiger, C. Meek, and B. Sturmfels. On the toric algebra of graphical models. *Ann. Statist.*, 34(3):1463–1492, 06 2006.
30. J. Gibbs. *Elementary Principles in Statistical Mechanics: Developed with Especial Reference to the Rational Foundations of Thermodynamics*. Elementary Principles in Statistical Mechanics: Developed with Especial Reference to the Rational Foundation of Thermodynamics. C. Scribner's sons, 1902.
31. R. Grosse and R. Salakhudinov. Scaling up natural gradient by sparsely factorizing the inverse fisher matrix. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2304–2313, Lille, France, 07–09 Jul 2015. PMLR.
32. J. M. Hammersley and P. E. Clifford. Markov random fields on finite graphs and lattices. Unpublished manuscript, 1971.
33. G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14:1771–1800, 2002.
34. G. E. Hinton. A practical guide to training restricted Boltzmann machines, version 1. Technical report, UTML2010-003, University of Toronto, 2010.
35. G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, July 2006.
36. G. E. Hinton and T. J. Sejnowski. Analyzing cooperative computation. In *Proceedings of the Fifth Annual Conference of the Cognitive Science Society, Rochester NY*, 1983.
37. G. E. Hinton and T. J. Sejnowski. Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1. chapter Learning and Relearning in Boltzmann Machines, pages 282–317. MIT Press, Cambridge, MA, USA, 1986.
38. J. J. Hopfield. Neurocomputing: Foundations of research. chapter Neural Networks and Physical Systems with Emergent Collective Computational Abilities, pages 457–464. MIT Press, Cambridge, MA, USA, 1988.
39. K. Huang. *Statistical Mechanics*. John Wiley and Sons, 2000.
40. M. I. Jordan. Graphical models. *Statist. Sci.*, 19(1):140–155, 02 2004.
41. R. Karakida, M. Okada, and S. Amari. Dynamical analysis of contrastive divergence learning: Restricted Boltzmann machines with Gaussian visible units. *Neural Networks*, 79:78–87, July 2016.
42. S. L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.
43. N. Le Roux and Y. Bengio. Representational power of restricted Boltzmann machines and deep belief networks. *Neural Computation*, 20(6):1631–1649, June 2008.
44. W. Li and G. Montúfar. Natural gradient via optimal transport I. *arXiv preprint arXiv:1803.07033*, 2018.
45. J. Martens, A. Chattopadhya, T. Pitassi, and R. Zemel. On the representational efficiency of restricted Boltzmann machines. In *Advances in Neural Information Processing Systems 26*, pages 2877–2885. Curran Associates, Inc., 2013.
46. F. Matúš and N. Ay. On maximization of the information divergence from an exponential family. In *Proceedings of the WUPES'03*, pages 199–204, 2003.
47. F. Matúš. Divergence from factorizable distributions and matroid representations by partitions. *Information Theory, IEEE Transactions on*, 55(12):5375–5381, Dec 2009.
48. G. Montavon, K.-R. Müller, and M. Cuturi. Wasserstein training of restricted boltzmann machines. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 3718–3726, USA, 2016. Curran Associates Inc.
49. G. Montúfar. Universal approximation depth and errors of narrow belief networks with discrete units. *Neural Computation*, 26(7):1386–1407, 2014.
50. G. Montúfar. Deep narrow Boltzmann machines are universal approximators. In *International Conference on Learning Representations (ICLR 15)*, 2015. Published online at arXiv:1411.3784.
51. 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.

52. G. Montúfar, N. Ay, and K. Ghazi-Zahedi. Geometry and expressive power of conditional restricted Boltzmann machines. *Journal of Machine Learning Research*, 16:2405–2436, 2015.
53. G. Montúfar and J. Morton. Discrete restricted Boltzmann machines. In *Online Proceedings of the 1-st International Conference on Learning Representations (ICLR2013)*, 2013.
54. G. Montúfar and J. Morton. When does a mixture of products contain a product of mixtures? *SIAM Journal on Discrete Mathematics*, 29(1):321–347, 2015.
55. G. Montúfar and J. Morton. Dimension of marginals of Kronecker product models. *SIAM Journal on Applied Algebra and Geometry*, 1(1):126–151, 2017.
56. G. Montúfar and J. Rauh. Scaling of model approximation errors and expected entropy distances. *Kybernetika*, 50(2):234–245, 2014.
57. G. Montúfar and J. Rauh. Hierarchical models as marginals of hierarchical models. *International Journal of Approximate Reasoning*, 88(Supplement C):531–546, 2017.
58. G. Montúfar, J. Rauh, and N. Ay. Expressive power and approximation errors of restricted Boltzmann machines. In *Advances in Neural Information Processing Systems 24*, pages 415–423, 2011.
59. G. Montúfar, J. Rauh, and N. Ay. *Geometric Science of Information: First International Conference, GSI 2013, Paris, France, August 28-30, 2013. Proceedings*, chapter Maximal Information Divergence from Statistical Models Defined by Neural Networks, pages 759–766. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.
60. R. Pascanu and Y. Bengio. Revisiting natural gradient for deep networks. In *International Conference on Learning Representations 2014 (Conference Track)*, Apr. 2014.
61. R. C. Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.*, 37:81–91, 1945.
62. J. Rauh. Finding the maximizers of the information divergence from an exponential family. *IEEE Transactions on Information Theory*, 57(6):3236–3247, 2011.
63. J. Rauh. Optimally approximating exponential families. *Kybernetika*, 49(2):199–215, 2013.
64. R. Salakhutdinov. Learning and evaluating Boltzmann machines. Technical report, 2008.
65. R. Salakhutdinov. Learning in Markov random fields using tempered transitions. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1598–1606. Curran Associates, Inc., 2009.
66. R. Salakhutdinov and G. E. Hinton. Deep Boltzmann machines. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS 09)*, pages 448–455, 2009.
67. 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.
68. A. Seigal and G. Montúfar. Mixtures and products in two graphical models. *To appear in Journal of Algebraic Statistics*, 2018. Preprint available from arXiv:1709.05276.
69. T. J. Sejnowski. Higher-order Boltzmann machines. In *Neural Networks for Computing*, pages 398–403. American Institute of Physics, 1986.
70. 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, Cambridge, MA, USA, 1986.
71. B. Steudel and N. Ay. Information-theoretic inference of common ancestors. *Entropy*, 17(4):2304, 2015.
72. S. Sullivant. *Algebraic Statistics*. 2018.
73. I. Sutskever and G. E. Hinton. Deep, narrow sigmoid belief networks are universal approximators. *Neural Computation*, 20(11):2629–2636, 2008.
74. T. Tieleman. Training restricted Boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 1064–1071, New York, NY, USA, 2008. ACM.
75. S. Watanabe. *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press, New York, NY, USA, 2009.

76. M. Welling, M. Rosen-Zvi, and G. E. Hinton. Exponential family harmoniums with an application to information retrieval. In *Advances in Neural Information Processing Systems 17*, pages 1481–1488, 2005.
77. L. Younes. Synchronous Boltzmann machines can be universal approximators. *Applied Mathematics Letters*, 9(3):109–113, 1996.
78. N. Zhang, S. Ding, J. Zhang, and Y. Xue. An overview on restricted Boltzmann machines. *Neurocomputing*, 275:1186 – 1199, 2018.