When Does a Mixture of Products Contain a Product of Mixtures?

Guido F. Montúfar* and Jason Morton†

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

Thursday 1st May, 2014

Abstract

We derive relations between theoretical properties of restricted Boltzmann machines (RBMs), popular machine learning models which form the building blocks of deep learning models, and several natural notions from discrete mathematics and convex geometry. We give implications and equivalences relating RBM-representable probability distributions, perfectly reconstructibe inputs, Hamming modes, zonotopes and zonosets, point configurations in hyperplane arrangements, linear threshold codes, and multi-covering numbers of hypercubes. As a motivating application, we prove results on the relative representational power of mixtures of product distributions and products of mixtures of pairs of product distributions (RBMs) that formally justify widely held intuitions about distributed representations. In particular, we show that an exponentially larger mixture of products, requiring an exponentially larger number of parameters, is required to represent the probability distributions represented as products of mixtures.

Keywords: linear threshold function, Hadamard product, zonotope, tensor rank, hyperplane arrangement
2000 MSC: 51M20, 60C05, 68Q32, 14Q15

*gfm10@psu.edu. Present address: Max Planck Institute for Mathematics in the Sciences.
†morton@math.psu.edu

Contents

1 Introduction **2**

2 Mixtures of products and products of mixtures **5**

3 Geometric Perspectives **6**

3.1 Inference functions, representations, reconstructability 7

3.2 Modes 9

3.3 Modes of mixtures of products 10

3.4 Modes of RBMs 12

3.5 Zonosets and hyperplane arrangements 13

3.6 Linear threshold codes 17

3.7 Multi-covering numbers of hypercubes 19

3.8 Proof of Theorem 5 21

4 Relative representational power **21**

4.1 When does a mixture of products contain an RBM? 21

4.2 When does an RBM contain a mixture of products? 23

5 Discussion **24**

1 Introduction

Two basic ways of combining probability distributions are mixtures, i.e., convex combinations, and Hadamard products, i.e., renormalized entry-wise products. Fixing the number of parameters, are products of small mixtures better than mixtures at approximating interesting or complex probability distributions? The general intuition among practitioners is that using products allows for more modelling power. We compare two canonical representatives of these model classes: mixtures of independent (product) distributions, called naïve Bayes models, and Hadamard products of mixtures of pairs of independent distributions, called restricted Boltzmann machines (RBMs). The mixture of products model $\mathcal{M}_{n,k}$ consists of all convex combinations of k independent distributions of n binary variables (Definition 6); the restricted Boltzmann machine model $\mathrm{RBM}_{n,m}$ consists of all Hadamard products of m mixtures of two independent distributions of n binary variables (Definition 7). Both are graphical probability models with hidden variables, see Figure 1. Besides defining probability distributions on their visible states, these graphical models define conditional distributions between visible and hidden states, which makes them interesting in the context of learning representations. This paper is the result of analyzing following problem.

Problem 1. When does the mixture of product distributions $\mathcal{M}_{n,k}$ contain the product of mixtures of product distributions $\mathrm{RBM}_{n,m}$, and vice versa?

We show that the number of parameters of the smallest mixture of products $\mathcal{M}_{n,k}$ containing the model $\mathrm{RBM}_{n,m}$ grows exponentially in the number of parameters of the latter for any fixed ratio $0 < m/n < \infty$. Theorem 43 gives a solution of order $\log_2(k) = \Theta(\min\{m, n\})$.

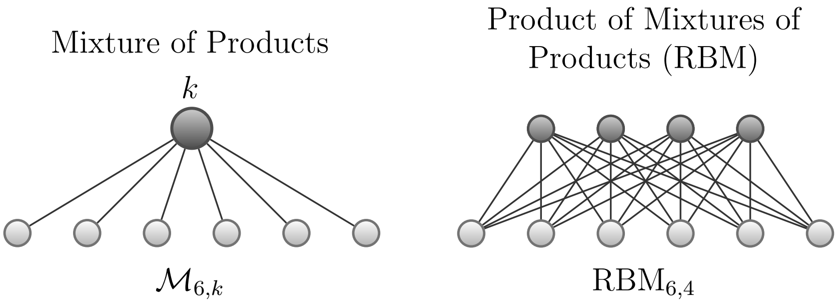

Figure 1: Graphical representation of a mixture of products and a product of mixtures. The dark nodes represent hidden units and the light nodes represent visible units.

See Figure 2 for an illustration of the result. In Theorem 46 we show a complementary result stating that, although RBMs naturally contain small mixture models, in general they do not contain mixture models that match their dimension.

To approach Problem 1, we study the sets of modes (Hamming-local maxima) of probability distributions that can be represented as mixtures of product distributions and as RBMs. We consider the following problems, showing in many cases that they are equivalent or equivalent after adding some necessary conditions.

Problem 2. What sets of length-n binary vectors are

1. the modes or strong modes (Hamming-local maxima) of probability distributions represented by an RBM with m hidden units?

2. perfectly reconstructible by an RBM with m hidden units? An input vector is encoded by the most likely hidden state given the input as visible state, and decoded by the most likely visible state given the encoding as hidden state.

3. the outputs of n linear threshold functions with m input bits?

We find that probability distributions with many strong modes (like the parity functions (probability distributions strictly supported on the binary vectors with even or odd number of ones), can be represented far more compactly by RBMs than by mixtures of products. Modes are described by linear inequalities of the form $p(x) > p(x')$ and can be used to derive polyhedral approximations of probability models. As it turns out, the analysis of modes is closely related to binary classification problems (separation of vertex sets of hypercubes by hyperplane arrangements), and leads to problems such as the following.

Problem 3. What is the smallest arrangement of hyperplanes, if one exists, that slices each edge of a hypercube a given number of times?

We consider the following properties of sets of binary vectors, and derive relations between them, summarized below in Theorem 5.

Definition 4. Let n and m be two non-negative integers and let $\mathcal{C}$ be a subset of $\{0,1\}^n$.

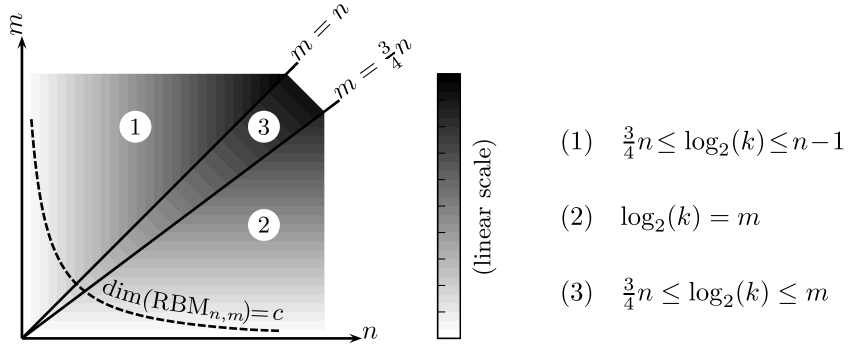

Figure 2: Smallest mixtures of products that can represent an RBM. Shown is a heat map of the logarithm of $k(n,m) = \min\{k' \in \mathbb{N} \colon \mathcal{M}_{n,k'} \supseteq \mathrm{RBM}_{n,m}\}$, depending on $n, m \in \mathbb{N}$. The domain of this function has three regions, each with approximately linear behavior (Theorem 43). An RBM of dimension c has hyperparameters n and m satisfying $nm + n + m = c$ (the dashed hyperbola). Fixing dimension, the RBMs which are hardest to represent as mixtures of product distributions are those with $m/n \approx 1$.

- **LTC**$(n, m, \mathcal{C})$: The set $\mathcal{C}$ is an (n,m)-linear threshold code, i.e., the image of n linear threshold functions with m inputs (Definition 30).
- **HP**$(n, m, \mathcal{C})$: There exists an arrangement $\mathcal{A}$ of n hyperplanes in $\mathbb{R}^m$ such that the vertices of the m-dimensional unit cube intersect exactly the $\mathcal{C}$-cells of $\mathcal{A}$ (Definition 25).
- **ZP**$(n, m, \mathcal{C})$: There is an m-zonoset (i.e., the affine image of an m-cube) in $\mathbb{R}^n$ which intersects exactly the $\mathcal{C}$-orthants of $\mathbb{R}^n$ (Definition 22).
- **SM**$(n, m, \mathcal{C})$: An RBM with n visible and m hidden nodes can represent a distribution with set of strong modes $\mathcal{C}$ (Definition 10).
- **PR**$(n, m, \mathcal{C})$: The set $\mathcal{C}$ is the set of perfectly reconstructible inputs of an RBM with n visible and m hidden nodes (Definition 9).
- **SP**$(n, m, \mathcal{C})$: An RBM with n visible and m hidden nodes can represent a distribution which is strictly positive on $\mathcal{C}$ and zero elsewhere.

We derive implications among the properties **LTC**, **PR**, **HP**, **ZP**, **SM**, and **SP** in two cases: the set $\mathcal{C}$ is arbitrary, and $\mathcal{C}$ consists of vectors which are at least Hamming distance 2 apart.

Theorem 5. *Fix integers n and m. For any $\mathcal{C} \subset \{0,1\}^n$, the following hold.*

1. *The properties* **LTC**, **HP**, *and* **ZP** *are equivalent.*
2. *If $\mathcal{C}$ satisfies* **PR** *or* **SM**, *then it is contained in an* **LTC** *set.*

$$\begin{tikzcd}
& \mathbf{SP} \arrow[dl, dashed, "d_H(\mathcal{C}) \geq 2"'] \arrow[dr, dashed, "d_H(\mathcal{C}) \geq 2"] & \\
\mathbf{SM} \arrow[dr] & & \mathbf{PR} \arrow[dl] \\
\mathbf{HP} \arrow[r, leftrightarrow] & \mathbf{LTC} \arrow[r, leftrightarrow] \arrow[uu, dotted, "\ell_1"'] & \mathbf{ZP}
\end{tikzcd}$$

Figure 3: Illustration of the implications in Theorem 5

3. *If the vectors in $\mathcal{C}$ are at least Hamming distance* 2 *apart, then* **SP** *implies both* **SM** *and* **PR***.*

4. *If the vectors in $\mathcal{C}$ are at least Hamming distance* 2 *apart and $\mathcal{C}$ satisfies an ℓ_1 property (see Theorem 23), then* **LTC** *implies* **SP***.*

This result is proven in Section 3.8 by combining results from Section 3. See Figure 3 for an illustration.

Section 2 contains basic definitions and background on mixtures of product distributions and RBMs. Section 3 discusses geometric perspectives on statistical models and inference, elaborated in various subsections. Section 3.1 discusses inference functions, distributed representations, and reconstructability. Section 3.2 discusses the concept of modes and polyhedral approximations of probability models. Section 3.3 covers the sets of modes of probability distributions realizable as mixtures of product distributions (Theorem 13). In Section 3.5, the modes of probability distributions realizable by RBMs are related to zonosets and hyperplane arrangements (Theorem 23). Section 3.7 discusses multi-covering numbers; the smallest hyperplane arrangements slicing each edge of a hypercube a given number of times. Section 3.8 contains the proof of Theorem 5. Turning to the motivating questions, Section 4 contains our analysis of Problem 1, treating the inclusion of RBMs in mixture models in Section 4.1, and the reverse inclusion in Section 4.2.

2 Mixtures of products and products of mixtures

Let $\mathcal{P}_n$ denote the (2^n-1)-dimensional simplex of probability distributions on $\{0,1\}^n$. An independent (or product) distribution of n binary variables is a probability distribution $p \in \mathcal{P}_n$ that factorizes as $p(x) = p(x_1)\cdots p(x_n)$ for all $x = (x_1, \ldots, x_n) \in \{0,1\}^n$. We denote the set of all product distributions of n binary variables by $\mathcal{M}_{n,1}$. This set is the closure of the n-dimensional exponential family $p_B(x) = \frac{1}{Z(B)}\exp(B^\top x)$ for all $x \in \{0,1\}^n$, with natural parameter $B \in \mathbb{R}^n$ and normalization $Z(B) = \sum_{y\in\{0,1\}^n}\exp(B^\top y)$.

Definition 6. The *k-mixture of product distributions of n binary variables* is the set $\mathcal{M}_{n,k}$ of distributions on $\{0,1\}^n$ expressible as convex combinations $p = \sum_{i\in[k]}\lambda_i q^{(i)}$, where $\sum_{i\in[k]}\lambda_i = 1$, $\lambda_i \geq 0$, and $q^{(i)} \in \mathcal{M}_{n,1}$ for all $i \in [k]$.

The set $\mathcal{M}_{n,k}$ has the same dimension as its Zariski closure in complex projective space, the kth secant variety of the nth Segre product of $\mathbb{P}^1$s. This is the dimension expected from counting parameters, equal to $\min\{nk + (k-1), 2^n - 1\}$, except when $(n,k) = (4,3)$, in which case it has dimension 13 instead of 14. See [4], which answered this century-old question in algebraic geometry. In addition, $\mathcal{M}_{n,k}$ is equal to the probability simplex $\mathcal{P}_n$ if and only if $k \geq 2^{n-1}$, see [16]. In particular, the smallest mixture of products model that can approximate any probability distribution on $\{0,1\}^n$ arbitrarily well has $2^{n-1}(n+1) - 1$ parameters.

Definition 7. The *RBM model* with n visible and m hidden binary units is the set $\mathrm{RBM}_{n,m}$ of distributions on $\{0,1\}^n$ that can be approximated arbitrarily well by distributions of the form

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

where $W \in \mathbb{R}^{m\times n}$ is a matrix of interaction weights between hidden and visible units (with state vectors h and x, respectively), $B \in \mathbb{R}^n$ is a vector of bias weights of the visible units, $C \in \mathbb{R}^m$ is a vector of bias weights of the hidden units, and $Z(W,B,C) = \sum_{x\in\{0,1\}^n} \sum_{h\in\{0,1\}^m} \exp(h^\top W x + B^\top x + C^\top h)$ is a normalization function.

The RBM model is a *product of experts* [12]; each hidden unit corresponds to an expert which is a mixture of two product distributions, see [6]. Each distribution $p \in \mathrm{RBM}_{n,m}$ is also a restricted mixture of 2^m conditional distributions which are product distributions, namely

$$p(x|h) = \frac{\exp((h^\top W + B^\top)x)}{\sum_{x'\in\{0,1\}^n} \exp((h^\top W + B^\top)x')} \qquad \text{for all } h \in \{0,1\}^m.$$

In general, the dimension of the mixture model $\mathcal{M}_{n,2^m}$ is much larger than that of $\mathrm{RBM}_{n,m}$. The set $\mathrm{RBM}_{n,m}$ is known to have dimension $nm+n+m$ when $m < 2^{n-\lceil \log_2(n+1)\rceil}$, and $2^n - 1$ when $m \geq 2^{n-\lfloor \log_2(n+1)\rfloor}$, see [6]. In addition, it is known that $\mathrm{RBM}_{n,m}$ equals $\mathcal{P}_n$ whenever $m \geq 2^{n-1} - 1$, see [17]. It is not known if the latter bound is always tight, but it shows that the smallest RBM model that can approximate any distribution on $\{0,1\}^n$ arbitrarily well has not more than $2^{n-1}(n+1) - 1$ parameters, and hence not more than the smallest mixture of products model.

We will show that the sets of probability distributions representable by RBMs and mixtures of products are quite different. The intersection of both model classes has been studied in [18], where it is shown that $\mathrm{RBM}_{n,m}$ and $\mathcal{M}_{n,m+1}$ intersect at sets of dimension of order at least $mn + m + 2n + 1 - (m-1)\log_2(m+1)$.

3 Geometric Perspectives

In this section we present five points of view on the families of probability distributions defined in the previous section. We consider inference functions and hidden representations defined by these models (Section 3.1), modes and strong modes of their marginal distributions (Sections 3.2, 3.3, and 3.4), zonosets, hyperplane arrangements, and linear threshold codes

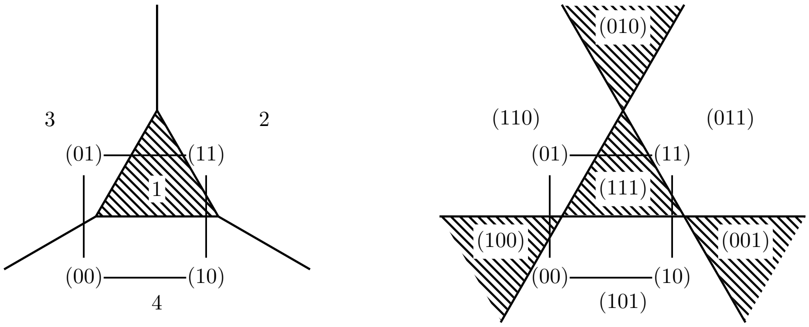

Figure 4: Inference regions of $\mathcal{M}_{2,4}$ (left) and $\mathrm{RBM}_{2,3}$ (right), for a choice of parameters.

that capture their combinatorial structure (Sections 3.5 and 3.6), and the resulting multi-covering numbers of hypercubes (Section 3.7).

Each point of view comes with particular set of related tools and implications for the capabilities of RBMs and competing models. We determine how these five concepts are related, which imply which, and how properties such as the number of strong modes in a marginal distribution translate into each perspective. These observations are summarized in Section 3.8, and in Section 4, where they are applied to distinguish mixtures of products and products of mixtures.

3.1 Inference functions, distributed representations, reconstructability

Hinton [12] discusses advantages of products of experts (Hadamard products of probability models) over mixtures of experts (mixtures of probability models), for modelling "high-dimensional data which simultaneously satisfies many low-dimensional constraints." In products of experts models, each expert can individually ensure that one constraint is satisfied. In the case of RBMs, each hidden unit linearly divides the input space according to its preferred state given the input, which results in a *multi-clustering*, or a partition of the input space into cells where different joint hidden states are most likely. Inference of the most likely hidden state given an input produces a distributed encoding or *distributed representation* of the input vector, as discussed by Bengio in [2, Section 5.3].

Definition 8. The *inference function* of a probability model $p_\theta(v, h)$ with parameter $\theta \in \Omega$ 'explains' each value of v by the most likely value of h according to $\mathrm{up}_\theta \colon v \mapsto \mathrm{argmax}_h\, p_\theta(h|v)$. This defines a partition of the input space into the preimages of all possible outputs, called *inference regions.*

For each choice of parameters $W \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^n$, $C \in \mathbb{R}^m$, the model $\mathrm{RBM}_{n,m}$ defines the inference function

$$\mathrm{up}_{W,B,C} \colon\ \mathbb{R}^n \supset \{0,1\}^n \to \{0,1\}^m;\ v \mapsto \mathrm{argmax}_{h\in\{0,1\}^m}\, h^\top(Wv + C)\ . \qquad (1)$$

The visible state v is explained by the hidden state h which satisfies $\operatorname{sgn}(Wv+C) = \operatorname{sgn}(h^\top - \frac{1}{2}\mathbb{1})$, where $\mathbb{1} := (1,\dots,1)$. There may be several explanations for a given observation, but generically there is only one. Geometrically, the input space $\mathbb{R}^n$ is partitioned into the preimages of the orthants of $\mathbb{R}^m$ by the affine map $\psi\colon \mathbb{R}^n \to \mathbb{R}^m; v \mapsto Wv+C$. This partition corresponds to the intersection of an affine space and the normal fan of an m-cube (the orthants of $\mathbb{R}^m$). The number of inference regions can be as large as $\mathfrak{C}_{\text{aff}}(m,d) = \sum_{i=0}^{d}\binom{m}{i}$, which is the number of orthants of $\mathbb{R}^m$ intersected by a generic d-dimensional affine subspace, where $d \leq \min\{n,m\}$ is the rank of W. When the rank of W is less than m (for example, when $m > n$), then the image of the map ψ does not intersect all orthants of $\mathbb{R}^m$ and there are 'empty' inference regions, i.e., states h which are not the explanation of any input vector v.

The mixture model $\mathcal{M}_{n,k}$, on the other hand, defines, for any choice of the mixture weights λ_i and the natural parameters of each mixture component $B_i \in \mathbb{R}^n$ for $i \in [k]$, an inference function

$$\operatorname{up}_{\lambda,B}\colon \quad v \mapsto \operatorname{argmax}_{i\in[k]}(B_i^\top v - \log(Z(B_i)) + \log(\lambda_i)), \tag{2}$$

where $Z(B_i) = \sum_{v\in\{0,1\}^n}\exp(B_i^\top v)$. In this case, the input space $\mathbb{R}^n$ is partitioned into the at most k regions of linearity of the function $v \mapsto \max\{B_i^\top v - \log(Z(B_i)) + \log(\lambda_i)\colon i \in [k]\}$. This partition corresponds to the intersection of an affine space and the normal fan of a $(k-1)$-simplex.

Figure 4 shows an example of inference regions in $\{0,1\}^2 \subset \mathbb{R}^2$ defined by $\mathcal{M}_{2,4}$ (left panel) and $\mathrm{RBM}_{2,3}$ (right panel), for some specific parameter values. Both models have 7 parameters and are universal approximators of distributions on $\{0,1\}^2$, but they define very different inference regions.

For a fixed input space dimension n, the number of inference regions in $\mathbb{R}^n$ that can be realized by $\mathrm{RBM}_{n,m}$ is of order $\Theta(\binom{m}{\min\{n,m\}})$, which is exponential in the number of parameters of the model, whereas the number of inference regions that can be realized by $\mathcal{M}_{n,k}$ is linear in the number of parameters of the model. Distributed representations can, in principle, learn different explanations to a number of observations that is exponential in the number of model parameters, see [2].

Now we discuss reconstructability. Similarly to the up_θ inference function, a model $p_\theta(v,h)$ defines a $\operatorname{down}_\theta$ inference function, which outputs the most likely visible state $\operatorname{argmax}_v p_\theta(v|h)$ given a hidden state h.

Definition 9. Given a probability model $p_\theta(v,h)$ on $v \in \mathcal{X}$ and $h \in \mathcal{Y}$, a collection of states $\mathcal{C} \subseteq \mathcal{X}$ is *perfectly reconstructible* if there is a choice of the parameter θ for which $\operatorname{down}_\theta(\operatorname{up}_\theta(v)) = v$ for all $v \in \mathcal{C}$.

The ability to reconstruct input vectors is used to evaluate the performance of RBMs in practice, since it can be tested more cheaply than the probability distributions they represent. The reconstructability of input vectors can also be used to define training algorithms, like in the case of auto-encoders.

When writing the joint probabilities $(p_\theta(v,h))_{v,h}$ as a matrix with rows labeled by h and columns by v, a set $\mathcal{C}$ is perfectly reconstructible iff there is a choice of θ for which

$p_\theta(v, \mathrm{up}_\theta(v))$ is the unique maximal entry in the $\mathrm{up}_\theta(v)$-row (and in the v-column) for all $v \in \mathcal{C}$.

For the model $\mathrm{RBM}_{n,m}$, this is the case exactly when for each $v \in \mathcal{C} \subseteq \{0,1\}^n$ there is an $h_v \in \{0,1\}^m$ with $\mathrm{sgn}(Wv + C) = \mathrm{sgn}(h_v - \frac{1}{2}\mathbb{1})$ and $\mathrm{sgn}(h_v^\top W + B^\top) = \mathrm{sgn}(v - \frac{1}{2}\mathbb{1})$. One can see, for example, that all cylinder subsets of $\{0,1\}^n$ of dimension m are perfectly reconstructible, by choosing the corresponding m columns of W equal to $(P - \frac{1}{2}\mathbb{1})$, where P is a permutation matrix and $\mathbb{1}$ is the matrix of ones. Later we will study the ability of RBMs to reconstruct more complicated sets of binary inputs, and how these sets relate to the visible probability distributions represented by the model.

3.2 Modes

We will characterize the ability of RBMs and mixtures of product distributions to represent distributions with many strong modes, in order to draw a distinction between them. As an interesting side remark, note that similar questions, about the number of modes of mixtures of multivariate normal distributions, have been posed in [24], and that the maximal number of modes realizable by mixtures of k normal distributions on $\mathbb{R}^n$ is unknown.

Definition 10. Let p be a probability distribution on a finite set $\mathcal{X}$ of length-n vectors. A vector $x \in \mathcal{X}$ is a *mode* of p if $p(x) > p(y)$ for all $y \in \mathcal{X}$ with $d_H(y,x) = 1$, and a *strong mode* if $p(x) > \sum_{y \in \mathcal{X}: d_H(y,x)=1} p(y)$. Here $d_H(x,y) := |\{i \in [n] \colon x_i \neq y_i\}|$ is the Hamming distance between x and y.

The modes of a distribution are the Hamming-locally most likely events in the space of possible events. Modes are closely related to the support sets and boundaries of statistical models, which have been studied especially for hierarchical and graphical models without hidden variables [11, 14, 23].

We write $\mathcal{G}_{n,m}$ (and $\mathcal{H}_{n,m}$) for the set of distributions in $\mathcal{P}_n$ which have at least m modes (strong modes). For any set $\mathcal{C} \subset \{0,1\}^n$ of vectors with Hamming distance at least 2 from each other, we write $\mathcal{G}_\mathcal{C}$ (and $\mathcal{H}_\mathcal{C}$) for the set of distributions which have modes (strong modes) $\mathcal{C}$. The closures $\overline{\mathcal{G}_\mathcal{C}}$ (and $\overline{\mathcal{H}_\mathcal{C}}$) are convex polytopes inscribed in the probability simplex $\mathcal{P}_n$. The sets of modes that are not realizable by a probability model give a full dimensional polyhedral approximation of the model's complement. See Figure 5 for an example. We will focus most of our consideration on strong modes. These are easier to study than modes, because they are described by fewer inequalities.

The minimum Hamming distance of a set $\mathcal{C} \subseteq \mathcal{X}$ is defined as $d_H(\mathcal{C}) := \min\{d_H(x,y) \colon x \neq y \text{ and } x, y \in \mathcal{C}\}$. Since any two modes have at least Hamming distance two from each other, a distribution on $\{0,1\}^n$ has at most 2^{n-1} modes. There are exactly two subsets of $\{0,1\}^n$ with cardinality 2^{n-1} and minimum distance two. These are the sets of binary strings with an even, respectively odd, number of entries equal to one

$$\begin{aligned} Z_{+,n} &:= \Big\{(x_1,\dots,x_n) \in \{0,1\}^n \colon \sum_{i\in[n]} x_i \text{ is even}\Big\}; \\ Z_{-,n} &:= \Big\{(x_1,\dots,x_n) \in \{0,1\}^n \colon \sum_{i\in[n]} x_i \text{ is odd}\Big\}. \end{aligned}$$

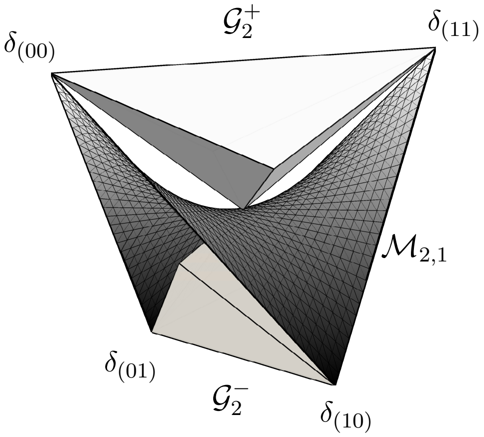

Figure 5: The 3-dimensional probability simplex (a tetrahedron with vertices corresponding to the outcomes $(00), (01), (10), (11)$), with three sets of probability distributions depicted. The dark curved surface is the 2-dimensional manifold $\mathcal{M}_{2,1}$ of product distributions of two binary variables. The angular regions at the top and bottom are the polyhedra $\mathcal{G}_2^+$ and $\mathcal{G}_2^-$ of distributions with two modes. An interactive 3-dimensional graphic object is available at www.personal.psu.edu/gfm10/blogs/gfmc_blog/indepmo.pdf.

We write $\mathcal{G}_{n,2^{n-1}} = \mathcal{G}_{Z_{+,n}} \cup \mathcal{G}_{Z_{-,n}}$, or $\mathcal{G}_n = \mathcal{G}_n^+ \cup \mathcal{G}_n^-$ for short, and similarly $\mathcal{H}_n = \mathcal{H}_n^+ \cup \mathcal{H}_n^-$. Figure 5 illustrates the three-dimensional sets $\mathcal{G}_2^+$ and $\mathcal{G}_2^-$, and the two-dimensional set $\mathcal{M}_{2,1}$ of product distributions on $\{0,1\}^2$. The case of three bits is as follows.

Example 11. The subset $\mathcal{G}_3^+ \subset \mathcal{P}_3$ of distributions on $\{0,1\}^3$ with modes $Z_{+,3}$, is the intersection of $\mathcal{P}_3$ and 12 open half-spaces defined by $p(x) > p(y)$ for all y with $d_H(x,y) = 1$ for all $x \in Z_{+,3}$. The closure of this set is a convex polytope $\overline{\mathcal{G}_3^+}$ with 19 vertices. The Lebesgue volume can be computed (e.g., using the software Polymake [10]): $\operatorname{vol}(\mathcal{G}_3^+)/\operatorname{vol}(\mathcal{P}_3) \approx 0.0179$. The set $\mathcal{H}_3^+$ of distributions with four strong modes $Z_{+,3}$, is the intersection of $\mathcal{P}_3$ and the 4 half-spaces defined by $p(x) > \sum_{y : d_H(x,y)=1} p(y)$ for $x \in Z_{+,3}$.

3.3 Modes of mixtures of products

In this section we characterize the sets of modes and strong modes that can appear in mixtures of product distributions, and show how these can be used to obtain a polyhedral approximation of the set of probability distributions representable in such models.

Problem 12. What is the smallest $k \in \mathbb{N}$ for which $\mathcal{M}_{n,k}$ contains a distribution with l (strong) modes?

A mixture of k uni-modal discrete probability distributions has at most k strong modes; for mixtures of products we have the following.

Theorem 13. *Let $\mathcal{M}$ be the k-mixture of the set of product distributions of n variables $x_i \in \mathcal{X}_i$, $|\mathcal{X}_i| < \infty$ for $i \in [n]$. The sets of strong modes of distributions within this model are exactly the sets of strings in $\mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ of minimum Hamming distance at least two and cardinality at most k. Furthermore, if $p \in \mathcal{M}$ has strong modes $\mathcal{C}$, then every $c \in \mathcal{C}$ is the mode of one mixture component of p.*

Proof. A product distribution q has at most one mode, because the value of the product $q_1(x_1)\cdots q_n(x_n)$ is either maximal, or can be increased by changing only one entry of x. If $q^{(j)}$, $j \in [k]$ are product distributions and x is not a mode of any $q^{(j)}$, then $\sum_{y:d_H(y,x)=1}\sum_{j\in[k]}\alpha_j q^{(j)}(y) \geq \sum_{j\in[k]}\alpha_j q^{(j)}(x)$ for any $\alpha_j \geq 0$, and x is not a strong mode of p. On the other hand, the set of product distributions contains every point measure δ_y, which is just the product distribution $q_1(x_1)\cdots q_n(x_n)$ with $q_i(y_i) = 1$ for all $i \in [n]$. Hence $\mathcal{M}$ contains any distribution $\sum_{y\in\mathcal{C}}\frac{1}{|\mathcal{C}|}\delta_y$ with $|\mathcal{C}| \leq k$. □

Remark 14. By the previous theorem, a mixture of k product distributions can have at most k strong modes, but it can have more than k modes. For instance, there are mixtures of two product distributions on $\{0,1\}^4$ which have more than two modes.

Theorem 13 shows that $(\mathcal{P}_n \setminus \mathcal{M}_{n,m}) \supseteq \mathcal{H}_{n,m+1}$ for all m. We can triangulate $\mathcal{H}_{n,m+1}$ and bound the volume of the complement of $\mathcal{M}_{n,m}$. A rough estimate is:

Proposition 15. *When $m < 2^{n-1}$ the Lebesgue volume of the complement of $\mathcal{M}_{n,m}$ satisfies $\operatorname{vol}(\mathcal{H}_{n,m+1})/\operatorname{vol}(\mathcal{P}_n) \geq 2^{-(m+1)n}K(m+1)$, where $K(m+1) = 2^{m+1}$ if $m+1 \leq 2^k \leq \frac{2^n}{n}$ for some k, and $K(m+1) = 2$ otherwise.*

Proof. Let $\mathcal{P}(\mathcal{Y})$ be the simplex of probability distributions strictly supported on $\mathcal{Y} \subseteq \mathcal{X} := \{0,1\}^n$. This is a regular $(|\mathcal{Y}|-1)$-simplex in $\mathbb{R}^{|\mathcal{X}|}$ with edge-length $\sqrt{2}$. Let $\mathcal{H}(\mathcal{Y})$ denote the set of distributions with strong modes at $\mathcal{Y}$, such that $\mathcal{H}(\mathcal{Y}) = \cap_{y\in\mathcal{Y}}\mathcal{H}(y)$. Let $B_1(y) \subseteq \mathcal{X}$ denote the Hamming ball with center y and radius 1. The set $\mathcal{P}_y(B_1(y)) := \{p \in \mathcal{P}(B_1(y))\colon p(y) \geq p(\mathcal{X}\setminus\{y\})\}$ is a regular n-simplex with edge-length $\frac{\sqrt{2}}{2}$ and vertices $\{\frac{1}{2}(\delta_y + \delta_{\hat{y}})\}_{d_H(\hat{y},y)\leq 1}$. The volume of a regular N-simplex with edge-length l is $\frac{\sqrt{N+1}}{N!\sqrt{2}^N}l^N$. The set $\mathcal{H}(y)$ is the convex hull of $\mathcal{P}_y(B_1(y))$ and $\mathcal{P}(\mathcal{X}\setminus B_1(y))$, and hence $\operatorname{vol}(\mathcal{H}(y)) = 2^{-n}\operatorname{vol}(\mathcal{P})$. If $\mathcal{Y}$ has minimum distance 3 or more, then $\operatorname{vol}(\mathcal{H}(\mathcal{Y})) = 2^{-|\mathcal{Y}|n}\operatorname{vol}(\mathcal{P})$. If the minimum distance is two, then the volume of $\mathcal{H}(\mathcal{Y})$ is larger.

The number $K(m+1)$ is a lower bound on the number of disjoint sets $\mathcal{H}(\mathcal{Y})$ with $|\mathcal{Y}| = m+1$. By the Gilbert-Varshamov bound, if $m+1 \leq 2^k$, where k is the largest integer with $2^k \leq \frac{2^n}{n}$, then there is a set $\mathcal{Y} \subset \mathcal{X}$, $|\mathcal{Y}| = m+1$ of minimum distance 3. Let $\mathcal{Y}' = (\mathcal{Y}\setminus\{y\}) \cup \{y \oplus_2 e_1\}$ (flip one coordinate of one element of $\mathcal{Y}$), such that $\mathcal{H}(\mathcal{Y}) \cap \mathcal{H}(\mathcal{Y}') = \emptyset$. Since $\mathcal{Y}$ has $(m+1)$ elements, there are 2^{m+1} disjoint sets of this form. For any $m+1 \leq 2^{n-1}$, if $\mathcal{Y}$ is a binary code of minimum distance 2, then also $\mathcal{Y}' \oplus_2 e_1$, and $\mathcal{H}(\mathcal{Y}) \cap \mathcal{H}(\mathcal{Y}') = \emptyset$. □

3.3.1 Polyhedral approximation of the full-dimensional model $\mathcal{M}_{3,3}$

Theorem 13 shows that any $p \in \mathcal{M}_{3,3}$ has at most three strong modes. We can show that this is true for modes too.

Proposition 16. *The mixture model of three product distributions on* $\{0,1\}^3$ *cannot realize distributions with four modes:* $\mathcal{M}_{3,3} \cap \mathcal{G}_3 = \emptyset$.

Proof. Assume that $\mathcal{M}_{3,3} \cap \mathcal{G}_3^+ \neq \emptyset$. By the Lemma 17 given below, there are factors $(p^{i,1}, p^{i,2}, p^{i,3}) \in (\mathcal{P}_1)^3, i = 1,2,3$ such that $\operatorname{conv}\{q^i := p^{i,2}p^{i,3}\}_{i=1,2,3}$ intersects $\mathcal{G}_2^+$ and $\mathcal{G}_2^-$. Hence $\operatorname{conv}\{q^1, q^2\}$ intersects $\mathcal{G}_2^+$ and $\operatorname{conv}\{q^2, q^3\}$ intersects $\mathcal{G}_2^-$ (for some enumeration of q^1, q^2, q^3). The mixture of q^2 and q^3 intersects $\mathcal{G}_2^-$ only if (01) and (10) are the unique maxima of q^2 and q^3. Similarly, $\operatorname{conv}\{q^1, q^2\}$ intersects $\mathcal{G}_2^+$ only if (11) and (00) are the unique maxima of q^1 and q^2; a contradiction. □

The proof of Proposition 16 uses the following lemma, which relates the number of modes realizable by $\mathcal{M}_{n,k}$ to the number of modes simultaneously realizable on subsets of variables.

Lemma 17. *Let* $n, k \in \mathbb{N}$ *and* $n \geq 2$. *Let* $p = \sum_{i\in[k]} \lambda_i \prod_{j\in[n]} p^{i,j} \in \mathcal{M}_{n,k}$, *with* $\lambda_i \geq 0$ *and* $p^{i,j} \in \mathcal{P}_1$ *for all* $(i,j) \in [k] \times [n]$. *If* p *has* 2^{n-1} *modes, then for any subset of variables* $I \subsetneq [n]$, $|I| = m$, *the convex hull of the product distributions* $\{\prod_{j\in I} p^{i,j}\}_{i\in[k]} \subset \mathcal{P}_m$ *intersects both* $\mathcal{G}_m^+$ *and* $\mathcal{G}_m^-$.

Proof. We show the special case with $I = \{1, \ldots, n-1\}$. The proof of the general case is a straightforward generalization. Any $q \in \mathcal{M}_{n,k}$ has the following form:

$$q(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{k} \lambda_i p^{i,1}(x_1) p^{i,2}(x_2) \cdots p^{i,n}(x_n), \tag{3}$$

for all $(x_1, x_2, \ldots, x_n) \in \{0,1\}^n$, where $\sum_{i=1}^k \lambda_i = 1$, $\lambda_i \geq 0$ and $p^{i,j} \in \mathcal{P}_1$. For the fixed value $x_1 = 0$ this is a mixture of k products with $(n-1)$ variables, multiplied by a positive constant:

$$q(x_1 = 0, x_2, \ldots, x_n) = c_0 \sum_{i=1}^{k} \lambda_{0,i} p^{i,2}(x_2) \cdots p^{i,n}(x_n), \tag{4}$$

where $\sum_{i=1}^k \lambda_{0,i} = 1$, $\lambda_{0,i} \geq 0$ with $\lambda_{0,i} = \frac{\lambda_i p^{i,1}(x_1=0)}{c_0}$ and $c_0 = \sum_{i=1}^k \lambda_{0,i} p^{i,1}(x_1 = 0)$. A similar observation can be made for the fixed value $x_1 = 1$. If the distribution q is contained in $\mathcal{G}_n^+$, then $q(x_1 = 0, x_2, \ldots, x_n) \in \mathcal{G}_{n-1}^+$ and $q(x_1 = 1, x_2, \ldots, x_n) \in \mathcal{G}_{n-1}^-$, since $q \in \mathcal{G}_n^+$. These two conditional distributions are mixtures of the same k product distributions $\{p^{i,2} \cdots p^{i,n}\}_{i\in[k]}$, even though they may have different mixture weights. □

Remark 18. The sets $\overline{\mathcal{G}_3^+}$ and $\overline{\mathcal{G}_3^-}$ are intersections of half-spaces that contain the uniform distribution. Although $\mathcal{M}_{3,3}$ is full dimensional in $\mathcal{P}_3$, by Proposition 16 the complement of $\mathcal{M}_{3,3}$ contains points arbitrarily close to the uniform distribution!

3.4 Modes of RBMs

In this section we characterize the sets of modes and strong modes that can appear in RBM-distributions.

Problem 19. What is the smallest $m \in \mathbb{N}$ for which $\mathrm{RBM}_{n,m}$ contains a distribution with l (strong) modes?

In particular, what is the smallest m for which the model $\mathrm{RBM}_{n,m}$ can represent the parity function? By Theorem 13 and $\mathrm{RBM}_{n,m} \subseteq \mathcal{M}_{n,2^m}$, any $p \in \mathrm{RBM}_{n,m}$ has at most $\min\{2^m, 2^{n-1}\}$ strong modes. We will see that this bound is not always tight, but often.

In special cases, the analysis of mixtures of products (Section 3.3) is sufficient to make statements about RBMs, for example: The model $\mathrm{RBM}_{4,2}$ is contained in $\mathcal{M}_{4,4}$ and has codimension one in $\mathcal{P}_4$. Its *algebraic implicitization* was studied in [7], i.e., its description as the set of zeros of a collection of polynomials. It was found to be the zero locus of a polynomial of degree 110 in as many as 5.5 trillion monomials. By Theorem 13, $\mathcal{M}_{4,4} \cap \mathcal{H}_4 = \emptyset$ and so $\mathrm{RBM}_{4,2} \cap \mathcal{H}_4 = \emptyset$. Using Proposition 16 and Lemma 17 one can show:

Proposition 20. *The models $\mathcal{M}_{4,4}$ and $\mathrm{RBM}_{4,2}$ cannot realize probability distributions with 8 modes: $\mathcal{M}_{4,4} \cap \mathcal{G}_4 = \emptyset$ and $\mathrm{RBM}_{4,2} \cap \mathcal{G}_4 = \emptyset$.*

Remark 21. The model $\mathrm{RBM}_{n,m}$ contains any distribution with support of cardinality $\min\{m+1, 2^n\}$ [17, Theorem 1]. Therefore, it contains some distributions with $\min\{m+1, 2^{n-1}\}$ strong modes. For example, $\mathrm{RBM}_{n,m}$ contains a uniform distribution on a set of cardinality $\min\{m+1, 2^{n-1}\}$ and minimum Hamming distance 2. In particular, whenever $\mathcal{M}_{n,k+1}$ contains a distribution with strong modes $\mathcal{C}$, then also $\mathrm{RBM}_{n,k}$ contains a distribution with strong modes $\mathcal{C}$. Note also that, since the model $\mathrm{RBM}_{n,m}$ is symmetric under relabeling of any of its variables, there is an RBM distribution with strong modes $\mathcal{C}$ iff there is one with strong modes $\mathcal{C} \oplus_2 x = \{c + x \mod (2) \colon c \in \mathcal{C}\}$ for any $x \in \{0,1\}^n$.

In general, characterizing the sets of modes realizable by RBMs is a more complex problem than it was for mixtures of product distributions, and will necessitate developing characterizations in terms of point configurations called zonosets (Definition 22) and hyperplane arrangements, or in terms of linear threshold functions.

3.5 Zonosets and hyperplane arrangements

Definition 22. Let $m \geq 0$, $n > 0$, $W_i \in \mathbb{R}^n$ for all $i \in [m]$, and $B \in \mathbb{R}^n$. The multiset $\mathcal{Z} = \{\sum_{i \in I} W_i + B\}_{I \subseteq [m]}$ is called an *m-zonoset.*

The convex hull of a zonoset is a zonotope, a well known object in the literature of polytopes. Zonotopes can be identified with hyperplane arrangements and oriented matroids [3, 31].

Given a sign vector $s \in \{-,+\}^n$, the s-orthant of $\mathbb{R}^n$ consists of all vectors $x \in \mathbb{R}^n$ with $\mathrm{sgn}(x) = s$. We say that an orthant has even (odd) parity if its sign vector has an even (odd) number of $+$. The sets of strong modes of RBMs can be described in terms of zonosets as follows.

Theorem 23. *Let $\mathcal{C} \subset \{0,1\}^n$ be a binary code of minimum Hamming distance at least two.*

- *If the model $\mathrm{RBM}_{n,m}$ contains a distribution with strong modes $\mathcal{C}$ (i.e., $\mathrm{RBM}_{n,m} \cap \mathcal{H}_{\mathcal{C}} \neq \emptyset$), or $\mathcal{C}$ has cardinality 2^m and is perfectly reconstructible by $\mathrm{RBM}_{n,m}$, then there is an m-zonoset with a point in each $\mathcal{C}$-orthant of $\mathbb{R}^n$.*

- *If there is an m-zonoset intersecting exactly the $\mathcal{C}$-orthants of $\mathbb{R}^n$ at points of equal ℓ_1-norm, then $\mathrm{RBM}_{n,m} \cap \mathcal{H}_{\mathcal{C}} \neq \emptyset$ and, furthermore, $\mathcal{C}$ is perfectly reconstructible.*

Proof. If there is a p in $\mathrm{RBM}_{n,m} \cap \mathcal{H}_{\mathcal{C}}$, then for each $x \in \mathcal{C}$ there is an $h \in \{0,1\}^m$ for which $p(\cdot|h)$ is uniquely maximized by x (Theorem 13 and $\mathrm{RBM}_{n,m} \subset \mathcal{M}_{n,2^m}$). This is also true if $\mathcal{C}$ is perfectly reconstructible. In this case, $(h^\top W + B^\top)x > (h^\top W + B^\top)v$ for all $v \neq x$, and, equivalently, $\mathrm{sgn}(h^\top W + B^\top) = \mathrm{sgn}(x - \frac{1}{2}(1,\dots,1)^\top)$. The existence of such W and B is equivalent to the existence of a zonoset with a point in each $\mathcal{C}$-orthant of $\mathbb{R}^n$.

Assume now that W, B can be chosen such that all vectors $h^\top W + B^\top$ have the same ℓ_1 norm, equal to K. We have $\frac{1}{2}K = \frac{1}{2}\|h^\top W + B^\top\|_1 = (h^\top W + B^\top)(x_h - \frac{1}{2}(1,\dots,1)^\top) = (h^\top W + B^\top)x_h + h^\top C - \frac{1}{2}B^\top(1,\dots,1)^\top$, where $C = -\frac{1}{2}W(1,\dots,1)^\top$, for some $x_h \in \mathcal{C}$ for all $h \in \{0,1\}^m$. The RBM with parameters $\alpha W, \alpha B$, $C = -\alpha W\frac{1}{2}(1,\dots,1)^\top$, and $\alpha \to \infty$ produces $\frac{1}{2^m}\sum_{h\in\{0,1\}^m} \delta_{x_h} \in \mathcal{H}_{\mathcal{C}}$ as its visible distribution. This also implies that $\mathcal{C}$ is perfectly reconstructible. □

Remark 24. The first part of Theorem 23 remains true if $\mathcal{H}_{\mathcal{C}}$ is extended to the set of distributions for which any $\mathcal{M}_{n,2^m}$-decomposition has a mixture component with mode c, for every $c \in \mathcal{C}$.

Definition 25. A *hyperplane arrangement* $\mathcal{A}$ in $\mathbb{R}^n$ is a finite set of (affine) hyperplanes $\{H_i\}_{i\in[k]}$ in $\mathbb{R}^n$. Choosing an orientation for each hyperplane, each vector $x \in \mathbb{R}^n$ receives a sign vector $\mathrm{sgn}_{\mathcal{A}}(x) \in \{-,0,+\}^k$, where $(\mathrm{sgn}_{\mathcal{A}}(x))_i$ indicates whether x lies on the negative side, inside, or on the positive side of H_i. The set of all vectors in $\mathbb{R}^n$ with the same sign vector is called a *cell* of $\mathcal{A}$.

A necessary condition for the existence of an m-zonoset intersecting all $\mathcal{C}$-orthants of $\mathbb{R}^n$ is that the number of orthants of $\mathbb{R}^n$ that are intersected by an m-dimensional affine space is at least $|\mathcal{C}|$. The maximal number of orthants intersected by a d-dimensional linear subspace of $\mathbb{R}^n$ was derived in [27]. It is not difficult to derive the corresponding number for a d-dimensional affine subspace too:

$$\mathfrak{C}(n,d) = 2\sum_{i=0}^{d-1}\binom{n-1}{i} \qquad \text{and} \qquad \mathfrak{C}_{\mathrm{aff}}(n,d) = \sum_{i=0}^{d}\binom{n}{i}. \tag{5}$$

Cover [5] shows that $\mathfrak{C}(n,d)$ is also the number of partitions of an n-point set in general position in $\mathbb{R}^d$ by central hyperplanes (hyperplanes through the origin). A set of vectors in $\mathbb{R}^d$ is in general position if any d or less are linearly independent. Dually, $\mathfrak{C}_{\mathrm{aff}}(n,d)$ can be seen as the number of cells of a real d-dimensional arrangement of n hyperplanes in general position [22, 26].

In particular, there are affine hyperplanes of $\mathbb{R}^n$ intersecting all but one orthants. Figure 4 (right) is an example showing the intersection of a 2-dimensional affine subspace of $\mathbb{R}^3$ and 7 orthants; four of odd parity and three of even parity. The number of even, or odd, orthants of $\mathbb{R}^n$ intersected by a generic m-dimensional affine space is $\lceil\frac{1}{2}\mathfrak{C}_{\mathrm{aff}}(n,m)\rceil$ or $\lfloor\frac{1}{2}\mathfrak{C}_{\mathrm{aff}}(n,m)\rfloor$. If $m < n$, then $\lfloor\frac{1}{2}\mathfrak{C}_{\mathrm{aff}}(n,m)\rfloor \geq 2^m$. This does not imply, however, that every collection of 2^m even, or odd, orthants can be intersected by an m-zonoset. For example:

Proposition 26. *If n is an odd natural number larger than one, then there is no $(n-1)$-zonoset with a point in every even, or every odd, orthant of $\mathbb{R}^n$.*

Proof. Let $\mathcal{Z}$ be a candidate zonoset; $\mathcal{Z}$ has $(n-1)$ generators, so it lies in an affine hyperplane H of $\mathbb{R}^n$. Let η be a normal vector to H. Assume first that $0 \in H$. All vectors in the orthants $\operatorname{sgn}(\eta)$ and $-\operatorname{sgn}(\eta)$ lie outside H (where we may assign arbitrary sign to zero entries of η). This follows from Stiemke's theorem, see, e.g., [9]. The two orthants have opposite sign vectors and n is odd, so one orthant is even and the other odd. Hence at least one even and one odd orthants do not intersect $\mathcal{Z}$.

Consider now an affine subspace H, and assume it intersects all even orthants. By eq. (5) $\dim(H) \geq n-1$, so H is a hyperplane. Assume without loss of generality that a normal vector to H has only negative entries. Then $H \cap \mathbb{R}^n_{(-\cdots-)}$ is an $(n-1)$-dimensional simplex containing a point of $\mathcal{Z}$. This can be inferred from the number of bounded cells in a d-dimensional arrangement of n hyperplanes in general position, $b(n,d) = \binom{n-1}{d}$ [28, Proposition 2.4]. The orthant $\mathbb{R}^n_{(-\cdots-)}$ is separated by $(n-1)$ coordinate hyperplanes from the orthant $\mathbb{R}^n_{s_i}$ with sign $s_i = (+\cdots+\underset{i}{-}+\cdots+)$ for any $i \in [n]$. Since n is odd and larger than one, $(n-1) > 0$ is even. Since $\mathcal{Z}$ intersects $H \cap \mathbb{R}^n_{s_i}$ for all $i \in [n]$, also $H \cap \mathbb{R}^n_{(-\cdots-)} \subset \operatorname{conv}(\mathcal{Z})$ (details in Lemma 27). On the other hand, the $(n-1)$-generated zonotope of dimension $(n-1)$ is combinatorially equivalent to the $(n-1)$-cube, and no point in its zonoset is contained in the convex hull of any other points. □

We used the following lemma in the proof of Proposition 26.

Lemma 27. *Let P be a polytope with vertex set V. Let $\{H_i\}_{i=1}^r$ the supporting hyperplanes of the facets of P incident to a vertex $v \in V$, and assume P is contained in the intersection of closed half-spaces $\cap_i H_i^+$. If v' is any point in $\cap_i H_i^-$, then the polytope* $\operatorname{conv}(\{v'\} \cup (V \setminus \{v\}))$ *contains P.*

Proof. The case $v' = v$ is trivial, so let $v' \neq v$. It is sufficient to show that v is not a vertex of $Q := \operatorname{conv}(\{v'\} \cup V)$, from which $v \in \operatorname{conv}(\{v'\} \cup (V \setminus \{v\}))$ and $P \subseteq \operatorname{conv}(\{v'\} \cup (V \setminus \{v\}))$ follows. The point v' is a vertex of Q, because $v' \notin P$. Consider first the case where v' is in the interior of $\cap H_i^-$, which is to say that v' is not contained in any H_i^+. If v was a vertex of Q, then one H_i would support a facet of Q (otherwise v' would be incident to all facets incident to v). This would contradict the fact that $v' \notin H_i^+$. The general case $v' \in \cap H_i^-$ results from continuity. □

Proposition 26 allows us to describe some distributions that cannot be represented by RBMs. A code $\mathcal{C} \subset \{0,1\}^n$ extends another code $\mathcal{C}' \subset \{0,1\}^r$, $r \leq n$, if restricting $\mathcal{C}$ to some r indices yields $\mathcal{C}'$.

Corollary 28. *If m is an even non-zero natural number and $m < n$, then* $\mathrm{RBM}_{n,m} \cap \mathcal{H}_{\mathcal{C}} = \emptyset$ *for any code $\mathcal{C} \subset \{0,1\}^n$ extending $Z_{+,m+1}$ or $Z_{-,m+1}$. In particular, when n is an odd natural number larger than one,* $\mathrm{RBM}_{n,n-1}$ *cannot represent any distribution with 2^{n-1} strong modes.*

Proof. If there is an m-zonoset with points in every $\mathcal{C}$-orthant of $\mathbb{R}^n$ and there is a restriction of $\mathcal{C}$ to $Z_{+,m+1}$ or $Z_{-,m+1}$, then there is an m-zonoset contradicting Proposition 26. By Theorem 23, $\mathrm{RBM}_{n,m}$ cannot represent distributions with strong modes $\mathcal{C}$. □

Corollary 28 implies, in particular, that the hierarchical model on the full bipartite graph $K_{n,m}$ does not contain in its closure any distribution supported on a set $\mathcal{Y} \subset \{0,1\}^{n+m}$ with

$$\{(x_{i_1}, \dots, x_{i_{m+1}}) \in \{0,1\}^{m+1} \colon (x_1, \dots, x_n, x_{n+1}, \dots, x_{n+m}) \in \mathcal{Y}\} = Z_{\pm, m+1}$$

for some $1 \leq i_1 < \cdots < i_{m+1} \leq n$.

In Section 3.7 (Corollary 41) we extend the statement of Corollary 28 showing that $\mathrm{RBM}_{6,5}$ cannot represent distributions with 2^{6-1} strong modes.

3.5.1 Polyhedral approximation of the full-dimensional model $\mathrm{RBM}_{3,2}$

The model $\mathrm{RBM}_{3,2}$ is particularly interesting, because it is the smallest candidate of an RBM universal approximator on $\{0,1\}^3$ in terms of the number of mixture components of the mixtures of products that it represents, but it has less than $2^{n-1} - 1$ hidden units, the upper bound for the number of hidden units of the smallest RBM universal approximator given in [17]. Note that the model $\mathrm{RBM}_{3,1} = \mathcal{M}_{3,2}$ is readily full dimensional.

By Corollary 28, the model $\mathrm{RBM}_{3,2}$ does not contain any distribution with four strong modes. We illustrate this explicitly: By Theorem 23, if $\mathrm{RBM}_{3,2} \cap \mathcal{H}_3 \neq \emptyset$, then

$$\operatorname{sgn}\begin{pmatrix} B \\ W_1 + B \\ W_2 + B \\ W_1 + W_2 + B \end{pmatrix} = \begin{pmatrix} + & - & - \\ - & + & - \\ - & - & + \\ + & + & + \end{pmatrix} \tag{6}$$

up to permutations of rows. But it is quickly verified that this equation cannot be satisfied.

The set $\mathcal{H}_3$ is the disjoint union of $\mathcal{H}_3^+ = \mathcal{H}_{Z_{+,3}}$ and $\mathcal{H}_3^- = \mathcal{H}_{Z_{-,3}}$. Its volume satisfies

$$\operatorname{vol}(\mathcal{H}_3) / \operatorname{vol}(\mathcal{P}_3) \approx 0.0078.$$

The set $\overline{\mathcal{H}_3^+}$ is the 7-dimensional simplex defined by the intersection of the 8 half-spaces with inequalities $p(z) \geq \sum_{y : d_H(z,y)=1} p(y)$ for $z \in Z_{+,3}$ and $p(y) \geq 0$ for all $y \in Z_{-,3}$. Its vertices are the uniform distributions on the following sets:

$$\{000, 001, 011, 101\}, \{011, 101, 110, 111\}, \{000, 010, 011, 110\}, \{000, 100, 101, 110\},$$
$$\{000\}, \{011\}, \{101\}, \{110\}.$$

The first four vertices are mixtures of two point measures and one uniform distribution on a pair of Hamming distance one. The last four vertices are the point measures on $Z_{+,3}$. By [18, Theorem 1], all these distributions are contained in $\mathrm{RBM}_{3,2}$ (by symmetry, the vertices of $\overline{\mathcal{H}_3^-}$ are also in $\mathrm{RBM}_{3,2}$). The distributions in the relative interiors of the edges of $\overline{\mathcal{H}_3^+}$ between the first four vertices are not in $\mathrm{RBM}_{3,2}$. The relative interior of edges connecting one of the first four and one of the last four vertices are in $\mathrm{RBM}_{3,2}$ if they have support of cardinality four and are not if they have support of cardinality five. We conjecture that $\mathrm{RBM}_{3,2} \cap \mathcal{G}_3 = \emptyset$.

3.6 Linear threshold codes

Definition 29. A *linear threshold function* (LTF) with m (binary) inputs is a function

$$f\colon \{0,1\}^m \to \{-,+\}; \quad y \mapsto \operatorname{sgn}((\sum_{j\in[m]} w_j y_j) + b);$$

where $w \in \mathbb{R}^m$ is called *weight vector* and $b \in \mathbb{R}$ *bias.* A subset $\mathcal{C} \subset \{0,1\}^m \subset \mathbb{R}^m$ is *linearly separable* iff there exists an LTF with $f(\mathcal{C}) = +$ and $f(\{0,1\}^m \setminus \mathcal{C}) = -$. For convenience we identify $-/+$ and $0/1$ vectors via $- \leftrightarrow 0$ and $+ \leftrightarrow 1$. The opposite $\overline{x}$ of a binary vector x is the vector given by inverting all entries of x.

LTFs are also known as *McCulloch-Pitts neurons* and have been studied intensively in the context of feed-forward artificial networks. The problem of separating subsets of vertices of the m-dimensional hypercube by hyperplane arrangements (multi-label classification) has drawn much attention, see, e.g., [29]. It is known that the logarithm of the number of LTFs with m inputs is asymptotically of order m^2, see [33, 21], but the exact number is only known for $m \leq 9$, see [30, 19, 20]. The study of LTFs simplifies when $f(x_1,\dots,x_m) = \overline{f}(\overline{x_1},\dots,\overline{x_m})$ for all $x \in \{0,1\}^m$, in which case they are called *self-dual.* If an LTF has an equal number of positive and negative points, then it separates every input from its opposite and is self-dual.

Definition 30. A subset $\mathcal{C} \subseteq \{0,1\}^n \cong \{-,+\}^n$ is an (n,m)-*linear threshold code* (LTC) if there exist n linear threshold functions $f_i\colon \{0,1\}^m \to \{0,1\}$, $i \in [n]$ with

$$\{(f_1(y), f_2(y), \dots, f_n(y)) \in \{0,1\}^n \colon y \in \{0,1\}^m\} = \mathcal{C}.$$

Equivalently, $\mathcal{C}$ is an (n,m)-LTC if it is the image of a down inference function of $\mathrm{RBM}_{n,m}$. If all f_i can be chosen self-dual, then $\mathcal{C}$ is called *homogeneous.*

In the following examples, an LTF with m inputs is written as a list of the vertices of the m-cube (the decimal they represent plus one), with a bar on inputs with negative output and no bar on inputs with positive output.

Example 31. Let $n = 3$ and $m = 2$. There are only two ways to linearly separate the vertices of the 2-cube into sets of cardinality two: $12\overline{34}$ and $1\overline{2}3\overline{4}$. These are the only possible columns of a homogeneous LTC with two inputs (up to opposites). The code $Z_{\pm,3}$ is not a $(3,2)$-LTC, because it has three non-equivalent columns. This shows that there does not exist a 2-zonoset with vertices in the four even, or odd, orthants of $\mathbb{R}^3$, and that $\mathrm{RBM}_{3,2}$ does not contain any distributions with four strong modes.

An alternative way of proving this is: The Hamming distance between any two elements of $Z_{\pm,n}$ is even. If the distance of any two vertices of the 2-cube induced by an arrangement of three hyperplanes is even and non-zero, then each edge is sliced at least twice, and in total at least eight edges are sliced (repetitions allowed). On the other hand, each plane slices at most two edges of the 2-cube, and so three planes slice at most 6 edges (repetitions allowed).

Example 32. Let $n = 4$ and $m = 3$. There are 104 ways to linearly separate the vertices of the 3-cube, see [21]. A complete list appears in [3, Section 3.8]. The vertices of the 3-cube

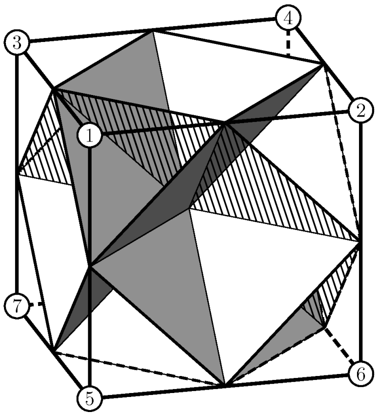

Figure 6: The four slicings of the 3-cube discussed in Example 32.

are in the $Z_{+,4}$-cells of an arrangement of four hyperplanes corresponding to the $(4,3)$-LTC with following LTFs:

$$123\overline{4}5\overline{678},\ 12\overline{3}4\overline{5}6\overline{78},\ 1\overline{2}34\overline{567}\overline{8},\ 1\overline{234}567\overline{8}.$$

This arrangement corresponds to a 3-zonoset with points in the 8 even orthants of $\mathbb{R}^4$ (Theorem 5). The zonoset can be realized as follows:

$$w = \begin{pmatrix} -1 & -1 & -1 & 1 \\ -1 & -1 & 1 & -1 \\ -1 & 1 & -1 & -1 \end{pmatrix}; \qquad b = \frac{1}{2}\begin{pmatrix} 3 & 1 & 1 & 1 \end{pmatrix}; \qquad \mathcal{Z} = \frac{1}{2}\begin{pmatrix} 3 & 1 & 1 & 1 \\ 1 & 3 & -1 & -1 \\ 1 & -1 & 3 & -1 \\ -1 & 1 & 1 & -3 \\ 1 & -1 & -1 & 3 \\ -1 & 1 & -3 & 1 \\ -1 & -3 & 1 & 1 \\ -3 & -1 & -1 & -1 \end{pmatrix}. \tag{7}$$

This choice of w and b corresponds to a central arrangement of four hyperplanes slicing each edge of the 3-cube exactly twice, as shown in Figure 6.

Example 33. Let $n = 5$ and $m = 4$. There are three symmetry types of self-dual LTFs with four input bits, see [19]. The following are representatives of the three types:

$$\begin{aligned} &\overline{1\,2\,3\,4\,5\,6\,7\,8}\,9\,10\,11\,12\,13\,14\,15\,16\ ; \\ &\overline{1\,2\,3\,4\,5\,6\,7}\,8\,\overline{9}\,10\,11\,12\,13\,14\,15\,16\ ; \\ &\overline{1\,2\,3\,4\,5\,6}\,7\,8\,\overline{9\,10}\,11\,12\,13\,14\,15\,16\ . \end{aligned}$$

By Proposition 26, the code $Z_{\pm,5}$ cannot be realized by any 5 LTFs, i.e., as a $(5,4)$-LTC, and $\mathrm{RBM}_{5,4}$ does not contain any distribution with 16 strong modes.

The following example presents a kind of binary code $\mathcal{C}$ of cardinality 2^m with $\mathrm{RBM}_{n,m} \cap \mathcal{H}_{\mathcal{C}} = \emptyset$ which is not covered by Corollary 28.

Example 34. Let $n = 5$ and $m = 3$. Let $x', x'' \in Z_{\pm,4}$ with $d_H(x', x'') = 4$, and

$$\mathcal{C} \quad = \quad \Big\{(x_1, \dots, x_5)\colon\ (x_1, \dots, x_4) \quad \in \quad Z_{\pm,4}, x_5 \quad = \quad \begin{cases} 1 & \text{if } (x_1, \dots, x_4) \in \{x', x''\} \\ 0 & \text{otherwise} \end{cases} \Big\}.$$

If $\mathcal{C}$ is an LTC, some hyperplane separates two vertices of the 3-cube from the other vertices (corresponding to $x_5 = 1$ only for two points). These two vertices must be connected by an edge of the 3-cube. Since $d_H(x', x'') = 4$, four hyperplanes pass through this edge. There are only three different central hyperplanes through an edge of the 3-cube, but four different central hyperplanes are required to produce $Z_{\pm,4}$. Hence $\mathcal{C}$ is not an $(5, 3)$-LTC.

3.7 Multi-covering numbers of hypercubes

The previous section shows that the sets of strong modes realizable by $\mathrm{RBM}_{n,m}$ are related to the solution of the following problem:

Problem 35. Let $m \leq n$. Consider an m-zonoset $\mathscr{Z}$ in $\mathbb{R}^n$ which does not intersect any two orthants separated by a single coordinate hyperplane. How many orthants of $\mathbb{R}^n$ does $\mathscr{Z}$ intersect at most?

As we discuss in the following, this problem is related to the long standing problem of computing the *covering numbers* of hypercubes.

Definition 36. The *covering number* of a hypercube is the smallest number of hyperplanes that slice each edge of the hypercube at least once. An edge is sliced by a hyperplane if the hyperplane intersects the relative interior of the edge, and does not contain any vertices of the hypercube. A *cut* is the collection of all edges sliced by a hyperplane and corresponds to a linear threshold function.

The m hyperplanes with normal vectors equal to the standard basis of $\mathbb{R}^m$ passing through the center of the m-dimensional hypercube slice all its edges. This arrangement is not always optimal. Paterson found 5 hyperplanes slicing all edges of the 6-cube, see [25]. This shows that covering numbers do not behave trivially. The covering numbers are known only for hypercubes of dimension ≤ 6. Computing them in higher dimensions is challenging, even in the cases where all cuts are known. Now:

Proposition 37. *If $Z_{+,n}$ is an $(n, n-1)$-LTC, then there exists an arrangement of n hyperplanes through the center of the $(n-1)$-cube slicing each edge an even non-zero number of times.*

Proof. Given the assumption, there exists an arrangement of n hyperplanes in $\mathbb{R}^{n-1}$ such that each vertex of the $(n-1)$-cube is in one of the $Z_{+,n}$-cells of the arrangement. Each vertex is separated by an even, positive number of hyperplanes from any other vertex, since any two elements of $Z_{+,n}$ differ in an even number of entries. Two vertices cannot be contained in the same cell of the arrangement, since $|Z_{+,n}| = 2^{n-1}$ equals the total number of vertices of the $(n-1)$-cube. The code $Z_{+,n}$ is homogeneous, as each coordinate $i \in [n]$ has the same number of zeros and ones, and hence each hyperplane in the arrangement can be chosen through the center of the cube. □

Proposition 37 motivates the following problem:

Problem 38 (Multi-covering number)**.** What is the smallest arrangement of hyperplanes, if one exists, that slices each edge of a hypercube a given number of times?

Of particular interest is the number of hyperplanes needed to slice each edge of the m-cube an even non-zero number of times. The edges of the m-cube can be sliced exactly twice by $2m$ hyperplanes with normal vectors the standard basis vectors of $\mathbb{R}^m$ counted with multiplicity two. Proposition 26 shows that if m is even and larger than zero, there is no arrangement of $(m+1)$ hyperplanes for which each vertex of the m-cube lies in a different cell and any two vertices are separated by an even number of hyperplanes. This suggests that when m is even, there is no arrangement of $(m+1)$ hyperplanes slicing all edges of the m-cube exactly twice; at least not one for which each vertex lies in a different cell.

There is exactly one way to slice all edges of the 3-cube an even, non-zero number of times by four hyperplanes, namely the way illustrated in Figure 6. To see that this is the only way, note that the 3-cube has twelve edges and that there are only four different cuts that slice six edges.

The 4-cube has 16 vertices, 32 edges, a total of 940 different cuts, 3 symmetry classes of central cuts, and 52 different central cuts. The maximal number of edges sliced by a cut is 12. Hence:

Proposition 39. *There is no arrangement of five hyperplanes, or less, slicing each edge of the four-dimensional cube at least twice.*

The complexity of the next easiest example is considerable. We tested all combinations of six cuts of the 5-cube and found:

Computation 40. *There does not exist an arrangement of six, or less, central hyperplanes slicing all edges of the five-dimensional cube an even non-zero number of times.*

In the following we explain some details of the computation. The 5-cube has 80 edges. There are 47 285 different ways of slicing them with affine hyperplanes, see [8]. A cut is given by the indicator function on the set of edges sliced. A list of the cuts can be found in [32]. An edge of the m-cube corresponds to a pair of binary vectors of length m which differ in exactly one entry. Each edge is parallel to one coordinate vector of $\mathbb{R}^m$. The edges can be organized in m groups, corresponding to their directions. Within each group, the edges are naturally enumerated by the binary vectors of length $(m-1)$ containing the coordinate values that are equal for the two vertices of each edge. The central cuts can be characterized as the cuts which involve only pairs of opposite edges. The 5-cube allows 7 symmetry classes of central cuts and 941 different central cuts. For each choice of 6, or less, central cuts we computed the entry-wise addition of the indicator functions and found that this never produced an even non-zero value in each entry. On the other hand, 5 is the covering number of the 5-cube, see [8], and hence at least 6 hyperplanes are needed to slice each edge twice. As a consequence of Computation 40 we have:

Corollary 41. *The model* $\mathrm{RBM}_{6,5}$ *cannot represent any probability distribution with* 32 *strong modes.*

Indeed, we trained $\mathrm{RBM}_{6,5}$ to approximate the uniform probability distribution on $Z_{+,6}$ and found a Kullback-Leibler divergence minimum of 0.6309 (with base-two logarithm), which is a relatively large value. For this computation we used contrastive divergence [13] and likelihood gradient with numerous parameter initializations.

3.8 Proof of Theorem 5

The equivalence theorem from the introduction (illustrated in Figure 3) is a summary of observations from the previous subsections. For completeness we provide a proof. Recall the definition of **LTC**, **PR**, **HP**, **ZP**, **SM**, and **SP** given in Definition 4. Let n and m be two integers and $\mathcal{C} \subseteq \{0,1\}^n$.

1. The properties **LTC**, **HP,** and **ZP** are equivalent.

   Let W and B be parameters making $\mathcal{C}$ a linear threshold code, so that $\mathcal{C} = \{\mathrm{sgn}(h^\top W + B) \colon h \in \{0,1\}^m\}$. Let W_i, $i = 1, \dots, n$ be the columns of W. The sign of $h^\top W_i + B_i$, $h \in \{0,1\}^m$ indicates which side of hyperplane H_i in the arrangement $\mathcal{A}_{W,B}$ this h lies on. Dually, the sign of $h^\top W_i + B_i$ indicates which side of the i-th coordinate hyperplane in $\mathbb{R}^n$ the point $h^\top W + B$ of the zonoset lies on.

2. If $\mathcal{C}$ satisfies **PR** or **SM**, then it is contained in an **LTC** set.

   For $\mathcal{C}$ to be the perfectly reconstructible, in particular it must be a subset of the image of a down inference function.

   If $\mathcal{C}$ has the **SM** property, its vectors are at least Hamming distance 2 apart. By Theorem 13, each point in $\mathcal{C}$ is the unique maximizer of a conditional distribution $p(\cdot|h)$ and an image point of the down inference function.

3. If the vectors in $\mathcal{C}$ are at least Hamming distance 2 apart, then **SP** implies both **SM** and **PR**.

   SP with Hamming distance two implies that each neighbour of each point in $\mathcal{C}$ has probability zero. Therefore, each point in $\mathcal{C}$ is a strong mode, and **SM**. Writing $p_\theta(v,h)$ as a matrix with rows labeled by h, the Hamming distance two condition implies by Theorem 13 that each row has a single non-zero entry, so $\mathrm{down}_\theta \circ \mathrm{up}_\theta$ is the identity on $\mathcal{C}$, so **PR** holds.

4. If the vectors in $\mathcal{C}$ are at least Hamming distance 2 apart and $\mathcal{C}$ satisfies an ℓ_1 property, then **LTC** implies **SP**.

   This is by Theorem 23.

4 Relative representational power

4.1 When does a mixture of products contain an RBM?

Using the characterizations obtained in Sections 3.3 and 3.4, we now prove the result on relative representational power discussed in the introduction and depicted in Figure 2. To do this, we derive upper bounds for the smallest m such that $\mathrm{RBM}_{n,m}$ contains probability distributions with l strong modes, and show thereby that RBMs can represent many more modes than mixtures of products with the same number of parameters.

Any representability result, as Example 32, combined with the following observation, yields lower bounds on the smallest mixture of products which contains the RBM model.

Observation 42. Assume that for each $i \in [k]$ there is a matrix $W^{(i)} \in \mathbb{R}^{m_i \times n_i}$ and a vector $B^{(i)} \in \mathbb{R}^{n_i}$ which generate a zonoset $\{h^\top W^{(i)} + B^{(i)} \colon h \in \{0,1\}^{m_i}\}$ intersecting K_i even orthants of $\mathbb{R}^{n_i}$. Then

$$W = \begin{pmatrix} W^{(1)} & & \\ & \ddots & \\ & & W^{(k)} \end{pmatrix} \quad \text{and} \quad B = (B^{(1)}, \dots, B^{(k)})$$

generate a zonoset $\{h^\top W + B \colon h \in \{0,1\}^{m_1+\cdots+m_k}\}$ intersecting $\prod_i K_i$ even orthants of $\mathbb{R}^{n_1+\cdots+n_k}$.

The following theorem provides the justification for the statement in the introduction that “the number of parameters of the smallest mixture of products model containing an RBM model grows exponentially in the number of parameters of the RBM for any fixed ratio $0 < m/n < \infty$,” and for Figure 2.

Theorem 43. *Let $n, m \in \mathbb{N}$.*

- *If $4\lceil m/3 \rceil \leq n$, then $\mathrm{RBM}_{n,m} \cap \mathcal{H}_{n,2^m} \neq \emptyset$ and*

$$\mathcal{M}_{n,k} \supseteq \mathrm{RBM}_{n,m} \ \textit{iff}\ k \geq 2^m \ .$$

- *If $4\lceil m/3 \rceil > n$, then $\mathrm{RBM}_{n,m} \cap \mathcal{H}_{n,L} \neq \emptyset$, where $L := \min\{2^l + m - l, 2^{n-1}\}$, $l := \max\{l \in \mathbb{N} \colon 4\lceil l/3 \rceil \leq n\}$, and*

$$\mathcal{M}_{n,k} \supseteq \mathrm{RBM}_{n,m} \ \textit{only if}\ k \geq L \ .$$

Proof. Let $4\lceil m/3 \rceil \leq n$. The *if* direction follows from $\mathrm{RBM}_{n,m} \subseteq \mathcal{M}_{n,2^m}$ for all n and m. For the *only if* direction we show that $\mathrm{RBM}_{n,m}$ contains a probability distribution supported on a set of cardinality 2^m and minimum Hamming distance at least two (a distribution with 2^m strong modes). By Theorem 13 such a distribution is in $\mathcal{M}_{n,k}$ only if $k \geq 2^m$. Consider the following parameters:

$$W = \alpha \left(\begin{array}{ccccc|c} w & & & & & \\ & w & & & & \\ & & \ddots & & & 0 \\ & & & w & & \\ & & & & \tilde{w} & \end{array}\right); \qquad \begin{aligned} B &= \alpha\,(b, b, \dots, b, -1, \dots, -1)\,; \\ b &= \tfrac{1}{2}(3,1,1,1); \\ C &= -W \tfrac{1}{2}(1,\dots,1)^\top = \alpha(1,\dots,1)^\top; \end{aligned}$$

where $\alpha \in \mathbb{R}$ is a constant, w is the 3×4-matrix defined in eq. (7), $\tilde{w}$ consists of the first or the first two rows of w, and B is α times $\lceil m/3 \rceil$ copies of b followed by -1s. Let λ_i be the set of indices $\{1,2,3,4\}+4(i-1)\subset[n]$. For $\alpha\to\infty$ the visible distribution generated by $\mathrm{RBM}_{n,m}$ with parameters W,B,C, is the uniform distribution on following subset of $Z_{+,n}$ of cardinality 2^m: $\{v\in\{0,1\}^n\colon \sum_{j\in\lambda_i}v_j$ is even for all i, and $v_j=0$ for all $j>4\lceil m/3\rceil\}$.

Now let $4\lceil m/3\rceil\geq n$. By the first part, $\mathrm{RBM}_{n,l}$ contains some p with 2^l strong modes in $Z_{+,n}$. Moreover, $\mathrm{RBM}_{n,l+1}$ contains $\mu p+(1-\mu)\delta_x$ for any $p\in\mathrm{RBM}_{n,l}$, $x\in\{0,1\}^n$ and $\mu\in[0,1]$ (see [15]), such that each additional hidden unit can be used to increase the number of strong modes by one, until the set of strong modes is $Z_{+,n}$. □

Remark 44. The statement of the first item of Theorem 43 remains true if $m=1 \mod (3)$ and $4\lfloor m/3\rfloor+2\leq n$. For $n<3$ we have $\mathcal{M}_{n,k}=\mathrm{RBM}_{n,k-1}$ for any $k\in\mathbb{N}$. For $n=3$ we believe that $\mathcal{M}_{3,3}$ and $\mathrm{RBM}_{3,2}$ are very similar, if not equal.

4.2 When does an RBM contain a mixture of products?

Complementary to question of when a mixture of products contains a product of mixtures, in this section we ask what is the smallest m for which $\mathrm{RBM}_{n,m}$ contains $\mathcal{M}_{n,k}$. We focus on an instance which we find particularly interesting:

Problem 45. Does $\mathrm{RBM}_{n,m}$ contain the mixture of products $\mathcal{M}_{n,m+1}$?

Both $\mathrm{RBM}_{n,m}$ and $\mathcal{M}_{n,m+1}$ have $nm+n+m$ parameters and expected dimension $\min\{nm+n+m,2^n-1\}$. The expected dimension is also the true dimension of both models for most choices of n and m [4, 6]. In the following we give a negative answer to Problem 45.

In the previous section we showed that the non-negative rank of probability distributions in the model $\mathrm{RBM}_{n,m}$ is as large as 2^m; there are tables of probabilities (probability distributions) represented by the RBM model, which cannot be represented as non-negative sums of less than 2^m non-negative rank-one tables (product distributions). The rank of a table p is the smallest number k such that p can be written as a sum of k rank-one tables. Here, a multivariate probability distribution $p=p(x_1,\dots,x_n)$ with $x_i\in\mathcal{X}_i$, $|\mathcal{X}_i|=r_i$ for $i=1,\dots,n$ is expressed as an n-way $r_1\times\cdots\times r_n$-table with value $p(x_1,\dots,x_n)$ at the entry $(x_1,\dots,x_n)$. A rank-one table is an outer-product of n vectors of lengths $r_1,\dots,r_n$. A product distribution in $\mathcal{M}_{n,1}$ is the outer-product of the marginal distributions on the variables x_1 through x_n and is a non-negative rank-one table. By definition, the elements of $\mathcal{M}_{n,k}$ have non-negative rank at most k, and therefore also rank at most k. Since $\mathrm{RBM}_{n,m}$ is contained in $\mathcal{M}_{n,2^m}$, any $p\in\mathrm{RBM}_{n,m}$ has rank at most 2^m.

Two models A and B are called *generically distinguishable* if $A\cap B$ has relative measure zero in A and in B. The restriction “generically” is useful, because in most cases of interest the models do intersect (e.g., mixtures of products and RBMs contain the uniform distribution). A *flattening* of a table of probabilities is a way of arranging its entries in a two-way table (i.e., a matrix) by grouping the variables in two groups and considering the joint states of the variables in each of the groups as the states of two variables. The following is an example of a flattening of a table p with four binary variables:

$$p=\begin{pmatrix} p_{00,00} & p_{00,01} & p_{00,10} & p_{00,11}\\ p_{01,00} & p_{01,01} & p_{01,10} & p_{01,11}\\ p_{10,00} & p_{10,01} & p_{10,10} & p_{10,11}\\ p_{11,00} & p_{11,01} & p_{11,10} & p_{11,11}\end{pmatrix}.$$

The matrix rank of any flattening of a table p is upper bounded by the outer-product rank of p. In particular, the vanishing of the $(k+1)\times(k+1)$-minors of flattenings are *algebraic invariants* of the model $\mathcal{M}_{n,k}$.

Theorem 46. *If $m \leq n/2$, then the model $\mathrm{RBM}_{n,m}$ contains points of rank 2^m. If, furthermore, $m+1 \neq 3$ or $n \neq 4$, then the models $\mathrm{RBM}_{n,m}$ and $\mathcal{M}_{n,m+1}$ have dimension $nm+n+m$ and intersect at a set of dimension strictly less than $nm+n+m$.*

Proof. We show that if $m \leq n/2$, then $\mathrm{RBM}_{n,m}$ contains a point p with a flattening of rank 2^m, which implies that p has outer-product rank 2^m. The flattenings of any $q \in \mathcal{M}_{n,k}$ have rank at most k. This gives an algebraic invariant of the mixture of products model $\mathcal{M}_{n,m+1}$ which is not satisfied by elements of $\mathrm{RBM}_{n,m}$. Hence, if both models have the same dimension d, then they intersect at a set of dimension strictly less than d.

Consider the m-cube and the $2m$ hyperplanes through its center consisting of translates of the coordinate hyperplanes with multiplicity two. This hyperplane arrangement slices each edge of the m-cube exactly twice and generates a $(2m, m)$-LTC $\mathcal{C}$ of minimum distance two. The code $\mathcal{C}$ consists of the 2^m binary vectors x in $\{0,1\}^{2m}$ with $x_i = x_{i+1}$ for all odd i, and $\{(x_1, x_3, \dots, x_{2m-1})\colon x \in \mathcal{C}\} = \{0,1\}^m$. In the case $m=3$, for example, the code is

$$\mathcal{C} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}. \tag{8}$$

By Theorem 23, $\mathrm{RBM}_{2m,m}$ contains the uniform distribution on $\mathcal{C}$, $u_{\mathcal{C}}$. View $u_{\mathcal{C}}$ as a linear transformation from the 2^m-dimensional space of real valued functions of $x_1, x_3, \dots, x_{2m-1}$ to the space of functions of $x_2, x_4, \dots, x_{2m}$, then

$$u_{\mathcal{C}} = \begin{pmatrix} 1/2^m & & & \\ & 1/2^m & & \\ & & \ddots & \\ & & & 1/2^m \end{pmatrix}, \tag{9}$$

which has rank 2^m. □

5 Discussion

RBMs create a multi-labeling of their input space by the most likely joint states of their hidden units given the inputs. The number of inference regions that can be generated in this way is of exponential order in the number of RBM parameters. The partitions of $\mathbb{R}^n$ generated by an RBM with n visible and m hidden units can be identified with the intersections of affine spaces of dimension $d \leq \min\{n, m\}$ with the orthants of $\mathbb{R}^m$, whereby each affine space corresponds to a choice of the RBM parameters. We elaborated on the combinatorics of the resulting hyperplane arrangements, and on the combinatorics of point configurations in such hyperplane arrangements, in correspondence with the inference functions on the set of binary input vectors $\{0,1\}^n \subset \mathbb{R}^n$. Although the theory of hyperplane arrangements and linear separation of points is well studied in the literature, it still poses many questions (see examples below).

We analyzed the sets of strong modes of probability distributions represented by RBMs and related them to the hyperplane arrangements and linear threshold codes (multi-labelings). The products of mixtures represented by RBMs are compact representations of probability distributions with many strong modes; of order $\min\{2^m, 2^{n-1}\}$ for the RBM with n visible and m hidden binary units (exponential in the number of parameters). At the same time, Corollaries 28 and 41 show that the hard bound $\min\{2^m, 2^{n-1}\}$ is not always attained. Mixture models of product distributions (naïve Bayes models), on the other hand, generate less restricted input space partitions but into at most as many regions as mixture components, and can only represent probability distributions with a number of strong modes of linear order in the number of model parameters.

These results imply that the smallest mixture model of product distributions that contains an RBM model is, in most cases, as large as one can possibly expect, having one mixture component per state of the RBM hidden units, and thus a number of parameters that is exponential in the number of RBM parameters. RBMs can represent distributions with many strong modes much more compactly than standard mixture models. This gives a concise combinatorial way of differentiating the two models. Fixing dimension, the RBMs which are hardest to represent as mixtures of product distributions are those with about the same number of visible and hidden units. At the same time, we note that there may exist small mixtures of product distributions which cannot be compactly represented by RBMs. For instance, Theorem 46 shows that $\mathcal{M}_{n,m+1} \not\subseteq \mathrm{RBM}_{n,m}$ when $3 \leq m \leq n/2$.

These results aid our understanding of how models complement each other, and why distributed representations in deep learning [2] can be expected to succeed, or when model selection can be based on theory rather than trial-and-error. They confirm the intuition that distributed representations are exponentially more powerful than non-distributed ones, in the case of binary RBMs and taking the number of strong modes, inference functions, and non-trivial perfectly reconstructible input sets as a measure of complexity. Other measures of complexity of probability distributions, such as *multi-information*, which is defined as the Kullback-Leibler divergence to the set of product distributions, are interesting but not necessarily best for differentiating between mixtures of products and RBMs. In terms of multi-information the most complex binary probability distributions have the form $p = \frac{1}{2}(\delta_x + \delta_y)$ with $x_i + y_i = 1$ for all i, see [1], and are contained in any (non-trivial) mixtures of products and RBM models.

A number of problems is covered only partially by our analysis. Some interesting open cases include

- Computing multi-covering numbers for hypercubes of odd dimension larger than five.

- Characterizing the support sets of fully observed RBM models. This problem can be seen to be equivalent to characterizing the face lattices of polytopes defined as Kronecker products of hypercubes.

- Computing the maximal cardinality of linear threshold codes of minimum Hamming distance two. Are there cases where the first item of Theorem 43 holds for $4\lceil m/3 \rceil > n$, $m \leq n-1$, assuming $m \neq n-1$ when n is odd?

- Can $\mathrm{RBM}_{8,7}$ represent probability distributions with 2^7 strong modes?

- Verifying the conjecture that $\mathrm{RBM}_{3,2} \cap \mathcal{G}_3 = \emptyset$, and, in addition, proving or disproving $\mathcal{M}_{3,3} = \mathrm{RBM}_{3,2}$.

Acknowledgments

The authors are grateful to Johannes Rauh for helpful discussions on algebraic invariants, to Nihat Ay for comments on hyperplane arrangements, to Yoshua Bengio for helpful discussions regarding distributed representations, and to Bernd Sturmfels for discussions motivating the analysis of mixtures of products in RBMs. The authors acknowledge use of the RCC-ITS computer cluster at the Pennsylvania State University. This work is supported in part by DARPA grant FA8650-11-1-7145.

References

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

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

[3] A. Björner, M. L. Vergnas, B. Sturmfels, N. White, and G. M. Ziegler. *Oriented Matroids*, volume 46 of *Encyc. of Mathematics and Its Applications*. Cambridge University Press, 1999.

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

[5] T. M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans. on Elect. Computers*, EC-14(3):326–334, 1965.

[6] 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.

[7] M. A. Cueto, E. A. Tobis, and J. Yu. An implicitization challenge for binary factor analysis. *J. Symbolic Computation*, 45:1296–1315, December 2010.

[8] M. R. Emamy-K and M. Ziegler. On the coverings of the d-cube for $d \leq 6$. *Discrete Applied Mathematics*, 156(17):3156–3165, 2008.

[9] L. Flatto. A new proof of the transposition theorem. *Proc. AMS*, 24(1):29–31, 1970.

[10] E. Gawrilow and M. Joswig. Polymake: a framework for analyzing convex polytopes. In *Polytopes – Combinatorics and Computation*, pages 43–74. Birkhäuser, 2000.

[11] D. Geiger, C. Meek, and B. Sturmfels. On the toric algebra of graphical models. *Annals of Statistics*, 34:1463–1492, 2006.

[12] G. E. Hinton. Products of experts. In *Proceedings 9-th ICANN*, volume 1, pages 1–6, 1999.

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

[14] T. Kahle, W. Wenzel, and N. Ay. Hierarchical models, marginal polytopes, and linear codes. *Kybernetika*, 45:189–208, 2009.

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

[16] G. Montúfar. Mixture decompositions of exponential families using a decomposition of their sample spaces. *Kybernetika*, 49(1):23–39, 2013.

[17] 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.

[18] G. Montúfar, J. Rauh, and N. Ay. Expressive power and approximation errors of restricted Boltzmann machines. In *NIPS 24*, pages 415–423, 2011.

[19] S. Muroga, T. Tsuboi, and C. Baugh. Enumeration of threshold functions of eight variables. *Computers, IEEE Transactions on*, C-19(9):818–825, sept. 1970.

[20] OEIS. The on-line encyclopedia of integer sequences, A000609 Number of threshold functions of n or fewer variables, 2010. Published electronically at http://oeis.org, 2010.

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

[22] P. Orlik and H. Terao. *Arrangements of Hyperplanes.* Springer-Verlag, 1992.

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

[24] S. Ray and B. G. Lindsay. The topography of multivariate normal mixtures. *Annals of Statistics*, 33(5):2042–2065, 2005.

[25] M. E. Saks. Slicing the hypercube. In K. Walker, editor, *Surveys in combinatorics*, pages 211–255. Cambridge University Press, New York, NY, USA, 1993.

[26] L. Schläfli. *Theorie der vielfachen Kontinuität.* Cornell University Library historical math monographs. George & Company, 1901.

[27] L. Schläfli. *Gesammelte mathematische Abhandlungen.* Number v. 2 in Gesammelte mathematische Abhandlungen. Birkhäuser, 1953.

[28] R. Stanley. An introduction to hyperplane arrangements. In *Lect. notes, IAS/Park City Math. Inst.*, 2004.

[29] W. Wenzel, N. Ay, and F. Pasemann. Hyperplane arrangements separating arbitrary vertex classes in n-cubes. *Advances in Applied Mathematics*, 25(3):284–306, 2000.

[30] R. O. Winder. Enumeration of seven-argument threshold functions. *Electronic Computers, IEEE Transactions on*, EC-14(3):315–325, 1965. See also correction in EC-16(2):231, 1967.

[31] G. M. Ziegler. *Lectures on polytopes.* Graduate texts in mathematics. Springer-Verlag, 1995.

[32] M. Ziegler. The cut number home page. http://www2.cs.uni-paderborn.de/cs/ag-madh/WWW/CUBECUTS/.

[33] Y. A. Zuev. Asymptotics of the logarithm of the number of threshold functions of the algebra of logic. *Sov. Math. Dok.*, 39(3):512–513, 1989.