JOURNAL OF ALGEBRAIC STATISTICS
Volume 9, No. 1, 2018, p. 1-20
https://jalgstat.library.iit.edu/
ISSN 1309-3452

Mixtures and Products in Two Graphical Models

Anna Seigal[1*], Guido Montúfar[2,3]

[1] *Department of Mathematics, University of California, Berkeley, USA*
[2] *Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany*
[3] *Departments of Mathematics and Statistics, University of California, Los Angeles, USA*

Abstract. We compare two statistical models of three binary random variables. One is a mixture model and the other is a product of mixtures model called a restricted Boltzmann machine. Although the two models we study look different from their parametrizations, we show that they represent the same set of distributions on the interior of the probability simplex, and are equal up to closure. We give a semi-algebraic description of the model in terms of six binomial inequalities and obtain closed form expressions for the maximum likelihood estimates. We briefly discuss extensions to larger models.

2000 Mathematics Subject Classifications: 62E10, 14M07, 14M25, 51M20

Key Words and Phrases: semi-algebraic statistics, restricted Boltzmann machine, mixture model, maximum likelihood estimation

*Corresponding author.

Email addresses: seigal@berkeley.edu (A. Seigal), montufar@math.ucla.edu (G. Montúfar)

1. Introduction

Graphical models are a popular tool for representing multivariate probability distributions in terms of conditional independence relations (see e.g. [6, 11]). Any probability distribution can be modeled by a graphical model, for instance a complete undirected graph imposes no constraints on the distribution. However, certain graphs involve many more parameters than others to represent specific distributions. In the interest of concisely representing data and reducing computational costs, one would like to understand which graph structures best represent which kinds of data. For example, deep architectures, with several layers of hidden variables, have become increasingly important in machine learning (see [5] and references therein). Following [13] (and using their notation) we focus on two important building blocks to such multi-layer architectures:

1. One hidden variable with k states, connected to n observed binary variables. This is the *mixture of products* model $\mathcal{M}_{n,k}$. Up to scaling, it consists of $2 \times \cdots \times 2$ (n times) tensors of non-negative rank at most k,

$$p = \sum_{i=1}^{k} a_i \otimes b_i \otimes \cdots \otimes c_i, \qquad a_i, b_i, \ldots, c_i \in \mathbb{R}^2_{\geq 0}.$$

2. A layer of m hidden binary variables, each connected to n observed binary variables. This is the *restricted Boltzmann machine* (RBM) model $\mathrm{RBM}_{n,m}$, also called the *product of mixtures of products* model. Up to scale, it consists of $2 \times \cdots \times 2$ (n times) tensors that are the Hadamard product of m tensors of non-negative rank at most two,

$$p = \prod_{i=1}^{m} (a_i \otimes b_i \otimes \cdots \otimes c_i + d_i \otimes e_i \otimes \cdots \otimes f_i), \qquad a_i, b_i, \ldots, c_i, d_i, e_i, \ldots, f_i \in \mathbb{R}^2_{\geq 0}. \quad (1)$$

Our main contribution is to find the set of distributions that these models can represent for the first open case $n = 3$. We find the semi-algebraic subset of the simplex that the models occupy. In doing so, we solve questions posed in [13].

The implicit description of a statistical model gives a membership test for distributions, allows the computation of distances to the model (e.g., in terms of the Kullback-Leibler divergence), and suggests model-specific algorithms for parameter estimation [19, 20]. In the above definitions, we consider the polynomial parametrization of the models. They are often defined as marginals of exponential families.[†] The two definitions are equivalent up to closure, see for example [13, Proposition 2.3]. In contrast to the exponential parametrization, we allow zeros in the decomposition, excluding the possibility that p is identically zero.

[†]As marginals of exponential families, RBMs and mixtures of products are given by $p(x) = \frac{1}{Z(W,b,c)} \sum_{y \in \{0,1\}^m} \exp(y^\top W x + c^\top y + b^\top x)$ and $p(x) = \frac{1}{Z(W,b,c)} \sum_{y \in \{e_j \,:\, j=1,\ldots,k\}} \exp(y^\top W x + c^\top y + b^\top x)$, respectively, where $x \in \{0,1\}^n$ and $Z(W,b,c)$ is a normalization function.

We note that $\mathcal{M}_{n,1}$ is the *independence model*, described by the intersection of the Segre variety $\text{Seg}(\mathbb{P}^1 \times \cdots \times \mathbb{P}^1)$ with the probability simplex Δ_{2^n-1} of joint probability distributions of n binary random variables. Also, by definition, $\mathcal{M}_{n,2} = \text{RBM}_{n,1}$. In [2] the description of $\mathcal{M}_{n,2}$ is found. The authors describe the 'formidable obstacles' to extending their results to hidden variables with more than two states.

Three binary variables take joint states in $\{0,1\}^3$. The $2 \times 2 \times 2$ tensor $(p_{ijk})_{0 \leq i,j,k \leq 1}$ stores the probabilities of these elementary events. Such probability distributions lie in the simplex $\Delta_{2^3-1} = \Delta_7$. Strictly positive distributions lie on the interior of the simplex. We obtain the following description of $\text{RBM}_{3,2}$.

Theorem 1.1. *The statistical model* $\text{RBM}_{3,2}$ *is described on the interior of the simplex* Δ_7 *by the union of six basic semi-algebraic sets. One is given by the two inequalities*

$$\{p_{000}p_{011} \geq p_{001}p_{010}, \quad p_{100}p_{111} \geq p_{101}p_{110}\}. \tag{2}$$

The other five are obtained by permuting indices, and/or reversing the inequalities:

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

These binomial inequalities correspond to determinants of slices of the tensor (p_{ijk}). They record conditional correlations in the distribution.

We compare $\text{RBM}_{3,2}$ to the mixture model $\mathcal{M}_{3,3}$ of non-negative rank at most three tensors. Both models are over-parametrized in the seven-dimensional simplex Δ_7, since they have 11 parameters. In [13], it is shown that $\mathcal{M}_{3,3}$ does not fill the simplex. The authors state 'we believe that $\mathcal{M}_{3,3}$ and $\text{RBM}_{3,2}$ are very similar, if not equal.' We resolve this question as follows.

Theorem 1.2. *We have the equality* $\mathcal{M}_{3,3} = \overline{\text{RBM}_{3,2}}$. *Equality* $\mathcal{M}_{3,3} = \text{RBM}_{3,2}$ *holds on the interior of the simplex.*

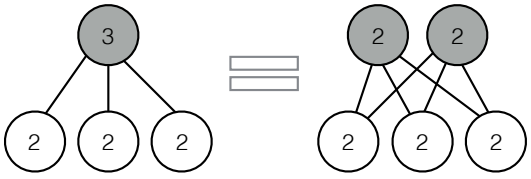


Figure 1: A pictorial representation of Theorem 1.2. The label of a variable is the number of states it has; the shaded nodes are hidden.

The notation $\overline{\text{RBM}_{3,2}}$ refers to the topological closure of $\text{RBM}_{3,2}$. The mixture model $\mathcal{M}_{3,3}$ and the RBM model $\text{RBM}_{3,2}$ look quite different in their parametrization, but this

result shows that they turn out to parametrize the same probability distributions (up to closure). The parametrization of $\mathrm{RBM}_{3,2}$ in (1) does not describe a closed set on the boundary of the simplex. We describe $\mathrm{RBM}_{3,2}$ on the boundary of the simplex in Proposition 2.1. On the other hand, $\mathcal{M}_{3,3}$ is closed (see Proposition 2.3) and we have the following corollary.

Corollary 1.3. *The model $\mathcal{M}_{3,3}$ is described on Δ_7 by the inequalities in Theorem 1.1.*

Previous results showed that $\mathcal{M}_{3,3}$ has relative volume at most 96.4%, and $\mathrm{RBM}_{3,2}$ has relative volume at most 99.2% inside the simplex Δ_7 [13]. Simulations using Theorem 1.1 and Corollary 1.3 estimate the true volume of both of these models to be 75.3%.

We use Theorem 1.1 to prove a conjecture from [13, Section 3.5.1]:

Corollary 1.4. *No distribution in $\mathrm{RBM}_{3,2}$ has four modes.*

For a discrete distribution, a *mode* is a state with larger probability than any of its Hamming neighbour states. Corollary 1.4 is stated as a conjecture $\mathrm{RBM}_{3,2} \cap \mathcal{G}_3 = \emptyset$ in [13], where $\mathcal{G}_3$ denotes distributions on $\{0,1\}^3$ with four modes (the maximum possible number). Note that the models $\mathcal{M}_{3,4}$ and $\mathrm{RBM}_{3,3}$ fill the interior of the simplex Δ_7 [12, 14]. Corollary 1.4 also follows from Theorem 1.2, since no $p \in \mathcal{M}_{3,3}$ has four modes [13, Proposition 3.10].

The remainder of the paper is organized as follows. We derive the implicit description of $\mathrm{RBM}_{3,2}$ in Section 2. We obtain the equality of $\mathrm{RBM}_{3,2}$ and $\mathcal{M}_{3,3}$ in Section 3. We connect this description to triangulations of the three-cube in Section 4, where we also prove Corollary 1.4. We describe the boundary of the model $\mathcal{M} = \mathcal{M}_{3,3} = \overline{\mathrm{RBM}_{3,2}}$ in Section 5, and we study the maximum likelihood problem for the model in Section 6. We explain how to construct a three-dimensional visualization of the model in Section 7. Finally, in Section 8 we study extensions to n binary random variables.

2. The semi-algebraic description of $\mathrm{RBM}_{3,2}$

We first recall the semi-algebraic description of the non-negative rank at most two model $\mathcal{M}_{3,2}$ given in [2]. The model is described in Δ_7 by the union of four basic semi-algebraic sets. On the interior of the simplex, one of the sets is given by the inequalities

$$\begin{array}{lll} p_{000}p_{011} \geq p_{010}p_{001}, & p_{000}p_{101} \geq p_{100}p_{001}, & p_{000}p_{110} \geq p_{100}p_{010}, \\ p_{100}p_{111} \geq p_{110}p_{101}, & p_{010}p_{111} \geq p_{110}p_{011}, & p_{001}p_{111} \geq p_{101}p_{011}. \end{array} \tag{3}$$

The other three sets are obtained by reversing the signs of the inequalities in *two out of the three* columns of (3). For example:

$$\begin{array}{lll} p_{000}p_{011} \geq p_{010}p_{001}, & p_{000}p_{101} \leq p_{100}p_{001}, & p_{000}p_{110} \leq p_{100}p_{010}, \\ p_{100}p_{111} \geq p_{110}p_{101}, & p_{010}p_{111} \leq p_{110}p_{011}, & p_{001}p_{111} \leq p_{101}p_{011}. \end{array} \tag{4}$$

One way to get a distribution in $\mathrm{RBM}_{3,2}$ is to take the Hadamard product of a distribution satisfying (3) with one satisfying (4). We find the semi-algebraic description for

all distributions expressible as such a Hadamard product. It is defined by the polynomial inequalities in (2). From this, swapping indices gives the full semi-algebraic description of the restricted Boltzmann machine $\mathrm{RBM}_{3,2}$ on the interior of the simplex. Note that the independence model $\mathcal{M}_{3,1}$ is obtained on the interior of Δ_7 by setting the inequalities in (3) or (4) to equalities.

2.1. On the interior of the simplex

The binomial inequalities above translate to linear inequalities in the log-probabilities. The inequalities are independent of scaling and we can work with unnormalized distributions. For a strictly positive distribution $p = (p_{ijk})$, we take the log distribution $l_{ijk} = \log(p_{ijk})$. Taking the logarithm of the inequalities in (3) gives the polyhedron

$$X = \begin{Bmatrix} l_{000} + l_{011} - l_{001} - l_{010} \geq 0, & l_{100} + l_{111} - l_{101} - l_{110} \geq 0 \\ l_{000} + l_{101} - l_{001} - l_{100} \geq 0, & l_{010} + l_{111} - l_{011} - l_{110} \geq 0 \\ l_{000} + l_{110} - l_{010} - l_{100} \geq 0, & l_{001} + l_{111} - l_{011} - l_{101} \geq 0 \end{Bmatrix}.$$

Similarly, we define Y to be the log-probabilities satisfying the logarithms of (4),

$$Y = \begin{Bmatrix} l_{000} + l_{011} - l_{001} - l_{010} \geq 0, & l_{100} + l_{111} - l_{101} - l_{110} \geq 0 \\ l_{000} + l_{101} - l_{001} - l_{100} \leq 0, & l_{010} + l_{111} - l_{011} - l_{110} \leq 0 \\ l_{000} + l_{110} - l_{010} - l_{100} \leq 0, & l_{001} + l_{111} - l_{011} - l_{101} \leq 0 \end{Bmatrix}.$$

Taking the Hadamard product in probability space is the same as taking the sum in log-probability space. Therefore, showing Theorem 1.1 is equivalent to proving that the Minkowski sum $X + Y = \{x + y : x \in X, y \in Y\}$ is

$$W = \{l_{000} + l_{011} - l_{001} - l_{010} \geq 0, \quad l_{100} + l_{111} - l_{101} - l_{110} \geq 0\}.$$

The two polyhedra X and Y are eight-dimensional in $\mathbb{R}^8$. The lineality spaces of a polyhedron is the space obtained by setting all the inequalities in their descriptions to equalities. For both X and Y, the lineality space is the set of tensors (l_{ijk}) for which the tensor $(\exp(l_{ijk}))$ is rank one. It is spanned by the rows of the matrix

$$\begin{array}{c} \begin{matrix} l_{000} & l_{100} & l_{010} & l_{001} & l_{110} & l_{101} & l_{011} & l_{111} \end{matrix} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{pmatrix}. \end{array}$$

The polyhedron W is also eight-dimensional. It has a six-dimensional lineality space that

is spanned degenerately by the rows of the matrix

$$\begin{array}{c} \begin{array}{cccccccc} l_{000} & l_{100} & l_{010} & l_{001} & l_{110} & l_{101} & l_{011} & l_{111} \end{array} \\ \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}. \tag{5}$$

Using the software `polymake` [9], we can find a description for the quotient of X or Y by its lineality space. They are both triangular bipyramids.

Proof of Theorem 1.1.

We aim to show that $W = X + Y$. We begin with the containment $X + Y \subseteq W$. Summing the first equations in X and Y yields

$$x_{000} + y_{000} + x_{011} + y_{011} - x_{001} - y_{001} - x_{010} - y_{010} \geq 0$$

while summing the second equations from X and Y gives

$$x_{100} + y_{100} + x_{111} + y_{111} - x_{101} - y_{101} - x_{110} - y_{110} \geq 0.$$

Translating back to the l-coordinates, we get $l_{000} + l_{011} - l_{001} - l_{010} \geq 0$ and $l_{100} + l_{111} - l_{101} - l_{110} \geq 0$. Hence $X + Y \subseteq W$.

For the reverse containment $W \subseteq X + Y$ we require a spanning set for W in which every basis vector lies either in X or in Y. The first four rows of the lineality space of W in (5) lie in X, while the last four rows lie in Y. Hence any non-negative combination of the lineality space lies in W. To extend to negative linear combinations we multiply the spanning set by -1. The first four rows of the negation of (5) lie in Y, and the last four are in X.

It remains to find a basis for the two-dimensional polytope obtained by taking the quotient of W by its lineality space. The quotient is spanned by non-negative combinations of any two linearly independent vectors in W not in its lineality space. For example $l_{000} \in X$ and $l_{100} \in Y$. All non-negative combinations of these lie in $X + Y$. This concludes the proof. □

2.2. On the boundary of the simplex

Theorem 1.1 gives a semi-algebraic description for the restricted Boltzmann machine $\mathrm{RBM}_{3,2}$ on the interior of the simplex Δ_7. However, for p in the boundary of the simplex $\partial\Delta_7$, the inequalities in Theorem 1.1 are not sufficient for membership in $\mathrm{RBM}_{3,2}$.

Proposition 2.1. *The intersection* $\mathrm{RBM}_{3,2} \cap \partial\Delta_7$ *is given by distributions which satisfy*

$$\begin{array}{c} \textit{If the probability of a state vanishes, so does the} \\ \textit{probability of one of its Hamming neighbour states.} \end{array} \tag{6}$$

Proof. First we show that $p \in \mathrm{RBM}_{3,2} \cap \partial\Delta_7$ satisfies condition (6). Since p lies on the boundary of Δ_7, one of its entries vanishes. Assume without loss of generality $p_{000} = 0$. Then condition (6) means that $p_{100}p_{010}p_{001} = 0$. Since $p \in \mathrm{RBM}_{3,2}$, it is the product of two distributions in $\mathcal{M}_{3,2}$. That is,

$$p_{ijk} = (q_{ijk} + r_{ijk})(s_{ijk} + t_{ijk}),$$

where q, r, s, t are rank one non-negative $2 \times 2 \times 2$ tensors. Up to swapping factors the $(0,0,0)$ entry of the tensor $q + r$ must vanish. Hence $q_{000} = r_{000} = 0$. Since q and r are rank one, they must vanish on a slice. Both q and r vanish in at least one of the locations $(0,0,1)$, $(0,1,0)$ and $(1,0,0)$, hence so does p.

For the converse, we consider some $p \in \partial\Delta_7$ satisfying (6) and we aim to show that $p \in \mathrm{RBM}_{3,2}$. As before, we can assume $p_{000} = 0$. Condition (6) implies that one of $p_{001}, p_{010}, p_{100}$ must also vanish. We reorder indices such that p_{010} vanishes. The distribution admits the Hadamard factorization

$$p = \left[\begin{array}{cc||cc} 0 & 0 & p_{001} & p_{011} \\ p_{100} & p_{110} & p_{101} & p_{111} \end{array}\right] = \left[\begin{array}{cc||cc} 0 & 0 & p_{001} & p_{011} \\ p_{101} & p_{111} & p_{101} & p_{111} \end{array}\right] * \left[\begin{array}{cc||cc} 0 & 0 & 1 & 1 \\ \frac{p_{100}}{p_{101}} & \frac{p_{110}}{p_{111}} & 1 & 1 \end{array}\right].$$

If $p_{101}, p_{111} \neq 0$, both factors are non-negative rank two and the distribution lies in $\mathrm{RBM}_{3,2}$. If $p_{101} = 0$, then $p_{111}p_{100}p_{001} = 0$ and if $p_{111} = 0$ then $p_{110}p_{101}p_{011} = 0$. In both of these cases the distribution consists of two pairs of non-zero adjacent entries, hence lies in $\mathcal{M}_{3,2}$, which is a subset of $\mathrm{RBM}_{3,2}$. Hence in all cases the distribution lies in $\mathrm{RBM}_{3,2}$. □

Condition (6) is stricter than the restriction of the inequalities in Theorem 1.1 to the boundary of the simplex. The model $\mathrm{RBM}_{3,2}$ is a semi-algebraic subset of the simplex that is not closed. We give an example of a distribution that lies in the closure of the model, but not in the model.

Example 2.2. Consider the distribution

$$p_{ijk} = \begin{cases} \frac{1}{3}, & (i,j,k) \in \{(0,0,1),(0,1,0),(1,0,0)\} \\ 0, & \text{otherwise.} \end{cases}$$

Observe that $p \in \mathcal{M}_{3,3}$, since $p = \frac{1}{3}(e_0 \otimes e_0 \otimes e_1 + e_0 \otimes e_1 \otimes e_0 + e_1 \otimes e_0 \otimes e_0)$ has non-negative rank three and entries summing to one. Since p does not satisfy the conditions in Proposition 2.1, $p \notin \mathrm{RBM}_{3,2}$. We give a sequence of distributions $(p_n) \subset \mathrm{RBM}_{3,2}$, such that $p_n \to p$. Consider

$$p_n \propto \left[\begin{array}{cc||cc} \epsilon & 1 & 1 & \epsilon \\ 1 & \epsilon & \epsilon & \epsilon^4 \end{array}\right],$$

where $||$ divides the two slices of the tensor, and $\epsilon = \frac{1}{n}$. As $n \to \infty$, $p_n \to p$. The scaling factor can be subsumed to either factor in the following decomposition.

$$\begin{aligned} p_n &\propto \left[\begin{array}{cc||cc} \epsilon & 1 & \epsilon^2 & \epsilon \\ 1 & \epsilon & \epsilon & \epsilon^2 \end{array}\right] * \left[\begin{array}{cc||cc} 1 & 1 & \epsilon^{-2} & 1 \\ 1 & 1 & 1 & \epsilon^2 \end{array}\right] \\ &= \left(\begin{bmatrix}\epsilon\\1\end{bmatrix} \otimes \begin{bmatrix}1\\0\end{bmatrix} \otimes \begin{bmatrix}1\\\epsilon\end{bmatrix} + \begin{bmatrix}1\\\epsilon\end{bmatrix} \otimes \begin{bmatrix}0\\1\end{bmatrix} \otimes \begin{bmatrix}1\\\epsilon\end{bmatrix}\right) * \left(\begin{bmatrix}1\\1\end{bmatrix} \otimes \begin{bmatrix}1\\1\end{bmatrix} \otimes \begin{bmatrix}1\\0\end{bmatrix} + \begin{bmatrix}\epsilon^{-1}\\\epsilon\end{bmatrix} \otimes \begin{bmatrix}\epsilon^{-1}\\\epsilon\end{bmatrix} \otimes \begin{bmatrix}0\\1\end{bmatrix}\right) \end{aligned}$$

This decomposition shows that $p_n \in \mathrm{RBM}_{3,2}$ for each n. Hence $\mathrm{RBM}_{3,2}$ is not closed.

In the example above, the entries of one of the tensors in the decomposition are unbounded as $n \to \infty$. They are multiplied by very small entries in the other term so that the limiting tensor p is bounded. Such situations can be avoided on the interior of the simplex, where the model $\mathrm{RBM}_{3,2}$ is closed, and can also be avoided for the mixture model $\mathcal{M}_{3,3}$.

Proposition 2.3. *The model $\mathcal{M}_{n,k}$ is closed for all n and k.*

Proof. Consider a convergent sequence of tensors $p_n \to p$, where each $p_n \in \mathcal{M}_{n,k}$. We show that the limiting tensor p also lies in $\mathcal{M}_{n,k}$. By definition, each p_n can be written as the sum of k non-negative rank one tensors $p_n = a_n + b_n + \cdots + c_n$. Since the entries of p_n are bounded above by 1, and the entries of $a_n, b_n, \ldots, c_n$ are non-negative, the entries of $a_n, b_n, \ldots, c_n$ are also bounded above by 1. By the Bolzano Weierstrass Theorem, there exists a subsequence of the a_n, call it a_{n_j}, that converges. Its limit, a, is a non-negative rank one tensor. Taking $p_{n_j} \to p$ as our new convergent sequence, we repeat the argument to find a convergent subsequence of the b_{n_j} which converges to a non-negative rank one tensor b. Repeating k times we obtain a subsequence of the p_n whose limit is $a + b + \cdots + c$. Hence $p = a + b + \cdots + c \in \mathcal{M}_{n,k}$. The result also follows directly from topological considerations, since $\mathcal{M}_{n,k}$ is the image of the closed set $(\Delta_1)^{nk} \times \Delta_{k-1}$ under a polynomial map. □

3. Equality of $\mathrm{RBM}_{3,2}$ and $\mathcal{M}_{3,3}$

We prove Theorem 1.2 by proving the two directions of the containment in two lemmas. The second sentence of the theorem (equality on the interior of the simplex) follows from the first (equality of the model closures) by the fact that $\mathrm{RBM}_{3,2}$ is closed on the interior of the simplex.

Lemma 3.1. *We have the containment of statistical models* $\mathrm{RBM}_{3,2} \subseteq \mathcal{M}_{3,3}$.

Proof. Consider a distribution $p \in \mathrm{RBM}_{3,2}$. If $p \in \partial\Delta_7$ then it satisfies (6) and we can assume without loss of generality $p_{000} = p_{001} = 0$. Then

$$p = \left[\begin{array}{cc||cc} 0 & 0 & 0 & 0 \\ p_{100} & 0 & p_{101} & 0 \end{array}\right] + \left[\begin{array}{cc||cc} 0 & p_{010} & 0 & p_{011} \\ 0 & 0 & 0 & 0 \end{array}\right] + \left[\begin{array}{cc||cc} 0 & 0 & 0 & 0 \\ 0 & p_{110} & 0 & p_{111} \end{array}\right]$$

is an expression for p as the sum of three non-negative rank one terms, hence $p \in \mathcal{M}_{3,3}$.

It remains to consider distributions p with no entries vanishing. We name the six determinants by $d_{i,j}$ where $i \in \{1,2,3\}$ denotes which index is fixed in the determinant, and $j \in \{0,1\}$ gives the value of the fixed index:

$$\begin{aligned} d_{1,0} &= p_{000}p_{011} - p_{001}p_{010}, \quad d_{1,1} = p_{100}p_{111} - p_{101}p_{110}, \\ d_{2,0} &= p_{000}p_{101} - p_{001}p_{100}, \quad d_{2,1} = p_{010}p_{111} - p_{011}p_{110}, \\ d_{3,0} &= p_{000}p_{110} - p_{010}p_{100}, \quad d_{3,1} = p_{001}p_{111} - p_{011}p_{101}. \end{aligned} \tag{7}$$

As we will see in Section 4 and Figure 3b, we can relabel indices such that determinants $d_{2,1}$ and $d_{1,1}$ have opposite signs. We can write p as

$$p = \left[\begin{array}{cc||cc} p_{000} & 0 & p_{001} & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] + \left[\begin{array}{cc||cc} 0 & 0 & 0 & 0 \\ p_{100} & x & p_{101} & \frac{p_{101}}{p_{100}}x \end{array}\right] + \left[\begin{array}{cc||cc} 0 & p_{010} & 0 & p_{011} \\ 0 & y & 0 & \frac{p_{011}}{p_{010}}y \end{array}\right],$$

where $x = \frac{p_{100}p_{111}\cdot d_{2,1}}{p_{101}d_{2,1} - p_{011}d_{1,1}}$ and $y = \frac{p_{010}p_{111}\cdot d_{1,1}}{p_{011}d_{1,1} - p_{101}d_{2,1}}$. Since the signs of $d_{2,1}$ and $d_{1,1}$ are different this expression for p is non-negative rank three, hence $p \in \mathcal{M}_{3,3}$. The denominator of x and y is non-zero, provided that $d_{2,1}$ or $d_{1,1}$ is non-zero. If some determinant vanishes, a non-negative rank three decomposition is obtained from the rank one tensor of that face plus the non-negative rank two representation of the opposite face.

Note that x and y are not both non-negative for $p \notin \mathrm{RBM}_{3,2}$ by Figure 4: there is no way to rotate or reflect the cube such that determinants $d_{2,1}$ and $d_{2,2}$ have opposite sign. □

Lemma 3.2. *We have the containment of statistical models* $\mathcal{M}_{3,3} \subseteq \overline{\mathrm{RBM}_{3,2}}$.

Proof. Consider a distribution $p + q \in \mathcal{M}_{3,3}$ where p is non-negative rank two, q is non-negative rank one, and no entries of p or q vanish. Up to swapping values 0 and 1 in one index, p being non-negative rank two means it satisfies the six binomial inequalities in (3). Equivalently, its determinants $d_{i,j}$ from (7) have sign pattern $(+,+,+,+,+,+)$, meaning that $d_{i,j} \geq 0$ for all i, j. We assume for contradiction that $p + q \notin \mathrm{RBM}_{3,2}$. This means $p + q$ has three "$-$" in its sign pattern, $d_{i,j} < 0$ for these pairs i, j. After adding tensor q, *three* determinants have swapped sign: $d_{1,h}$, $d_{2,h}$, $d_{3,h}$ for $h = 0$ or 1.

Take non-negative vectors $a, b, c \in \mathbb{R}^2_{\geq 0}$ such that $q_{ijk} = a_i b_j c_k$. Assume determinant $d_{3,h}$ of $p+q$ is negative: $(p_{00h} + a_0b_0c_h)(p_{11h} + a_1b_1c_h) - (p_{01h} + a_0b_1c_h)(p_{10h} + a_1b_0c_h) < 0$. Multiplying this expression out, and using $p_{00h}p_{11h} \geq p_{01h}p_{10h}$, gives

$$p_{00h}a_1b_1 + p_{11h}a_0b_0 < p_{10h}a_0b_1 + p_{01h}a_1b_0. \tag{8}$$

Hence either $p_{00h}b_1 < p_{01h}b_0$ or $p_{11h}b_0 < p_{10h}b_1$ must hold, and likewise either $p_{00h}a_1 < p_{10h}a_0$ or $p_{11h}a_0 < p_{01h}a_1$ must hold. Furthermore, rearranging (8) yields

$$\frac{1}{p_{00h}}(p_{00h}a_1 - p_{10h}a_0)(p_{00h}b_1 - p_{01h}b_0) + \left(p_{11h} - \frac{p_{10h}p_{01h}}{p_{00h}}\right)a_0b_0 < 0.$$

Since the last term is non-negative, this implies that $\frac{1}{p_{00h}}(p_{00h}a_1 - p_{10h}a_0)(p_{00h}b_1 - p_{01h}b_0) < 0$, hence exactly one of $p_{00h}a_1 < p_{10h}a_0$ and $p_{00h}b_1 < p_{01h}b_0$ holds. Similarly, (8) yields

$$\frac{1}{p_{11h}}(p_{11h}a_0 - p_{01h}a_1)(p_{11h}b_0 - p_{10h}b_1) + \left(p_{00h} - \frac{p_{01h}p_{10h}}{p_{11h}}\right)a_1b_1 < 0,$$

implying exactly one of $p_{11h}a_0 < p_{01h}a_1$ and $p_{11h}b_0 < p_{10h}b_1$ holds. Repeating the above for determinants $d_{2,h}$ and $d_{1,h}$ gives the following $2^3 = 8$ options:

$$\begin{aligned}
I^{(1)}_{ab} &= \{p_{00h}b_1 < p_{01h}b_0, \quad p_{11h}a_0 < p_{01h}a_1\}, \quad I^{(2)}_{ab} = \{p_{11h}b_0 < p_{10h}b_1, \quad p_{00h}a_1 < p_{10h}a_0\},\\
I^{(1)}_{ac} &= \{p_{0h0}a_1 < p_{1h0}a_0, \quad p_{1h1}c_0 < p_{1h0}c_1\}, \quad I^{(2)}_{ac} = \{p_{1h1}a_0 < p_{0h1}a_1, \quad p_{0h0}c_1 < p_{0h1}c_0\},\\
I^{(1)}_{bc} &= \{p_{h00}c_1 < p_{h01}c_0, \quad p_{h11}b_0 < p_{h01}b_1\}, \quad I^{(2)}_{bc} = \{p_{h11}c_0 < p_{h10}c_1, \quad p_{h00}b_1 < p_{h10}b_0\}.
\end{aligned}$$

If either inequality from $I^{(1)}_{ab}$ is satisfied, the inequalities of $I^{(2)}_{ab}$ cannot be satisfied, and likewise for I_{ac} and I_{bc}. To conclude the proof, we derive a contradiction from these options.

Let $h = 0$. Assume the inequalities in $I^{(1)}_{ab}$ hold. Then one of the inequalities from $I^{(2)}_{bc}$ is satisfied, hence $I^{(1)}_{bc}$ cannot hold. If $I^{(1)}_{ac}$ also holds, combining $p_{110}a_0 < p_{010}a_1$ from $I^{(1)}_{ab}$ with $p_{000}a_1 < p_{100}a_0$ from $I^{(1)}_{ac}$ gives $p_{110}p_{000} < p_{010}p_{100}$, contradicting the hypothesis that p satisfies the inequalities in (3). If $I^{(2)}_{ac}$ holds, combining inequalities involving c gives $p_{000}p_{011} < p_{001}p_{010}$, also a contradiction. Likewise, if $I^{(2)}_{ab}$ holds then $I^{(1)}_{ac}$ must hold. If $I^{(1)}_{bc}$ also holds, combining the inequalities involving c implies $p_{101}p_{000} < p_{100}p_{001}$, a contradiction. If $I^{(2)}_{bc}$ holds, combining inequalities involving b gives $p_{110}p_{000} < p_{100}p_{010}$, also a contradiction. The case $h = 1$ follows by analogous reasoning.

This shows that an open dense subset of $\mathcal{M}_{3,3}$ is contained in $\mathrm{RBM}_{3,2}$. It remains to consider when p or q has some vanishing entry. Such cases are in the closure of the above, hence they lie in the closure of $\mathrm{RBM}_{3,2}$. □

Proof of Theorem 1.2. Lemma 3.1, and the closedness of the model $\mathcal{M}_{3,3}$, imply the inclusion of closures $\overline{\mathrm{RBM}_{3,2}} \subseteq \mathcal{M}_{3,3}$. Combining with the inclusion in Lemma 3.2 gives $\mathcal{M}_{3,3} \subseteq \overline{\mathrm{RBM}_{3,2}} \subseteq \mathcal{M}_{3,3}$, hence the two models are equal up to closure. Theorem 1.1 implies that $\mathrm{RBM}_{3,2}$ is closed on the interior of the simplex, hence we have $\mathcal{M}_{3,3} = \mathrm{RBM}_{3,2}$ on the interior of the simplex. □

4. Connection to triangulations of the three-cube

Let $\mathcal{M}$ be the statistical model $\mathcal{M}_{3,3} = \overline{\mathrm{RBM}_{3,2}}$. We characterize $\mathcal{M}$ on the interior of Δ_7 in terms of triangulations. This allows us to prove Corollary 1.4. We describe below how to triangulate the three-cube using a positive distribution $p \in \Delta_7$. Membership in $\mathcal{M}$ is determined by how this triangulation restricts to the faces of the cube.

Consider a generic, strictly positive distribution $p \in \Delta_7$. Its tensor of log-probabilities $(l_{ijk}) = \log(p_{ijk})$ induces a triangulation of the three-cube. For two observed variables,

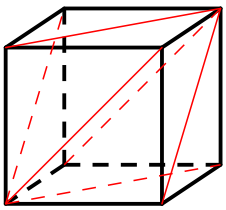
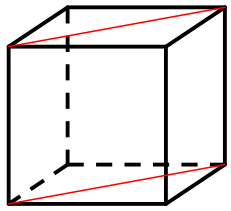
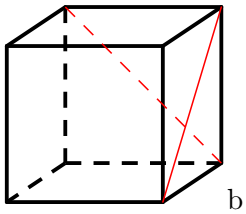


Figure 2: Distributions in $\mathcal{M}_{3,2}$ give (rotations of) this triangulation.



a

Figure 3: Two characterizations of the triangulations from $\mathcal{M}$ (up to rotation). Empty faces can be triangulated in either direction.



b

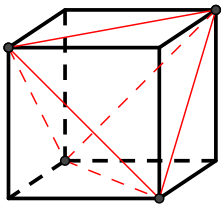


Figure 4: Distributions not in $\mathcal{M}$ give this triangulation (up to rotation).

the set-up is shown in [4, Figure 1]. In three dimensions, we do the higher-dimensional analogue: we assign the height l_{ijk} to each vertex $(i,j,k) \in \{0,1\}^3$ of the three-cube. Then we take the upper part of the convex hull of the points (i,j,k,l_{ijk}) in four-dimensional space, which we refer to as the upper hull, and project it back to the three-dimensional cube. The facets in the upper hull project to tetrahedra that triangulate the cube. Figures 2, 3, and 4 illustrate different types of triangulations of the cube by showing how the triangulations restrict to the faces of the cube. Rotating or reflecting the cubes is allowed within each type; this corresponds to relabeling indices of the distributions. The empty faces in Figure 3 can be triangulated in either of the two possible directions, subject to the condition that the triangulation of the faces is a restriction of a triangulation of the whole cube. Some of the 2^6 possible triangulations of the six faces do not come from a triangulation of the whole cube.

The triangulation of a face of the cube is an equivalent description of the sign of one of the determinants in (7). Hence, we can use the triangulation types to capture the inequalities that describe the model $\mathcal{M}$.

Proposition 4.1. *The model $\mathcal{M}$ contains distributions with triangulations of the form shown in Figure 3. Distributions with triangulations of the form shown in Figure 4 lie outside of $\mathcal{M}$. Triangulations in Figure 2 are special cases of those in Figure 3 and come from distributions in $\mathcal{M}_{3,2}$.*

Proof. There are 20 linear expressions in the coordinates l_{ijk} whose signs determine the triangulation, see [4, page 1325]. Six of these equations determine how the triangulation restricts to the faces of the cube. These are the logarithms of the binomial equations that define $\mathcal{M}$. Hence we can see whether $\exp(l_{ijk})$ lies in $\mathcal{M}$ by looking at how the triangulation induced by (l_{ijk}) restricts to the faces of the cube. The equations that define $\mathcal{M}_{3,2}$ and $\mathcal{M}_{3,1}$ are also of this form.

In the language of triangulations, being in $\mathcal{M}$ means we triangulate *at least one pair of opposite faces in the same direction*, as in Figure 3a. The condition for being in $\mathcal{M}_{3,2}$ is that

every pair of opposite faces is triangulated in the same direction, with sign compatibility as in Figure 2. Triangulations of distributions not in $\mathrm{RBM}_{3,2}$ *triangulate every pair of opposite faces in opposing directions*, as in Figure 4. An alternate characterization of such triangulations is that every pair of adjacent faces is triangulated in a continuous way. If, conversely, a pair of adjacent faces is triangulated in a discontinuous way, as in Figure 3b, the distribution lies in $\mathcal{M}$. □

The three-cube has 74 possible triangulations which fall into six triangulation types, see [10, Figure 1]. In [4] the authors study these triangulations in the context of epistasis in evolutionary biology. We can re-phrase Proposition 4.1 in terms of the numbering of the triangulation types from [10, Page 1657]. The model $\mathrm{RBM}_{3,2}$ only contains distributions with triangulation types 3, 4, 5 and 6. Triangulation types 1 and 2 come from distributions that lie outside of the model. Triangulation type 6 is from distributions in $\mathcal{M}_{3,2}$. Triangulation type 4 can be distinguished from 3 and 5 using the empty faces of the cube in Figure 3a. The empty faces can be triangulated in either direction. If a pair of opposite empty faces are triangulated in different directions from one another we have type 3 or 5. Type 4 occurs if both empty pairs of opposite faces are triangulated in the same direction, but not with the right sign-compatibility for $\mathcal{M}_{3,2}$ membership.

Proof of Corollary 1.4. The idea of the proof is to show that distributions with four modes restrict to the faces of the cube as shown in Figure 4. Assume we have a distribution with four modes. Without loss of generality, the four numbers l_{000}, l_{011}, l_{101}, and l_{110} exceed the values of their neighbours. Consider a face of the cube, for example the face $\langle l_{000}, l_{001}, l_{010}, l_{011}\rangle$. Since $l_{000} \geq l_{001}$ and $l_{011} \geq l_{010}$, we have

$$l_{000} + l_{011} - l_{010} - l_{001} \geq 0,$$

which determines how the triangulation of (l_{ijk}) restricts to the face. Repeating for the other faces gives the triangulation of the faces shown in Figure 4.

Distributions on $\partial\Delta_7 \cap \mathrm{RBM}_{3,2}$ have at least two adjacent entries vanishing, by (6). This excludes the possibility of having four modes. □

5. The boundary of the model

We saw that the statistical model $\mathcal{M} = \mathcal{M}_{3,3} = \overline{\mathrm{RBM}_{3,2}}$ is defined by the binomial inequalities in Theorem 1.1. Setting the inequalities in Theorem 1.1 to equalities gives the Zariski closure of the boundary of the model.

Proposition 5.1. *Distributions on the boundary of* $\mathcal{M}$ *are given by* $2\times 2\times 2$ *tensors with a* 2×2 *slice of rank* ≤ 1.

That is, the Zariski closure of the boundary of the model is a union of hypersurfaces $\{d_{i,j} = 0\}$, for $1 \leq i \leq 3$, $0 \leq j \leq 1$. This is also the Zariski closure of the boundary of the model $\mathcal{M}_{3,2}$ from [2]. Proposition 5.1 says the boundary of $\mathcal{M}$ consists of mixtures of

three product distributions with disjoint supports in $\{0,1\}^3$. Mixtures of products with disjoint supports were used in [15] to study the representational power of RBMs.

The following is a converse result. It implies that $\mathrm{RBM}_{3,2}$ is closed on the interior of the simplex. Furthermore, within the simplex of probability distributions, the Zariski closure of the boundary is contained in the closure of the model. This result (which fails for $\mathcal{M}_{3,2}$) is useful in Section 6 when we study maximum likelihood estimation.

Lemma 5.2. *Every distribution of three binary random variables with a rank one* 2×2 *slice, and strictly positive entries, lies in the models* $\mathrm{RBM}_{3,2}$ *and* $\mathcal{M}_{3,3}$.

Proof. As in the proof of Lemma 3.1, if the determinant of a distribution p vanishes, a non-negative rank three decomposition is obtained from the rank one tensor of that slice plus the non-negative rank two representation of the opposite slice. This proves the result for $\mathcal{M}_{3,3}$.

It remains to build a decomposition of p as $(q+r)(s+t)$ where q, r, s, t are rank one non-negative $2 \times 2 \times 2$ tensors, and multiplication is entry-wise, as in (1). Assume without loss of generality that $d_{3,1} = 0$. Let q be the rank one tensor with slices q_{**1} and p_{**1} equal, where q_{**0} is set to be the smallest scalar multiple of p_{**1} that zeros out an entry of p_{**0}. The notation p_{**0} refers to the slice p_{ij0} for $i, j \in \{0,1\}$. Then $p - q$ consists of at most three non-zero entries. Let r be the tensor which satisfies $r_{ijk} = p_{ijk} - q_{ijk}$ for two of the three entries at which $p \neq q$. Since these two entries can be chosen to be Hamming neighbours, r is rank one. And since $p - q$ is non-negative, r is non-negative. There remains at most one entry where equality $p = q + r$ does not hold: let i, j, k be such that $p_{ijk} > q_{ijk} + r_{ijk}$. Let s be the all ones tensor, and let t be the tensor with just one non-zero entry, $t_{ijk} = \frac{p_{ijk}}{q_{ijk}+r_{ijk}} - 1$. Then t is also non-negative and rank one, and $p = (q+r)(s+t)$ as required. □

In the log-probability coordinates, the boundary of $\mathcal{M}$ is the union of hyperplanes:

$$\begin{array}{ll} \mathcal{L}_{1,0} = \{l_{000} + l_{011} - l_{001} - l_{010} = 0\}, & \mathcal{L}_{1,1} = \{l_{100} + l_{111} - l_{101} - l_{110} = 0\}, \\ \mathcal{L}_{2,0} = \{l_{000} + l_{101} - l_{001} - l_{100} = 0\}, & \mathcal{L}_{2,1} = \{l_{010} + l_{111} - l_{011} - l_{110} = 0\}, \\ \mathcal{L}_{3,0} = \{l_{000} + l_{110} - l_{010} - l_{100} = 0\}, & \mathcal{L}_{3,1} = \{l_{001} + l_{111} - l_{011} - l_{101} = 0\}. \end{array} \tag{9}$$

The intersection poset of a hyperplane arrangement is the set of all intersections of hyperplanes, ordered by reverse inclusion [18]. In Figure 5 we give the intersection poset of the pieces of the boundary of $\mathcal{M}$. As an example of its non-generic structure, in Figure 5 we highlight three codimension three flats that are intersections of four hyperplanes.

We can study the combinatorics of the arrangement using its characteristic polynomial $\chi(t) = \sum_f \mu(f) t^{\dim(f)}$. The summation is taken over all flats in the arrangement, and μ is the Möbius function (indicated in Figure 5 next to each node). Evaluating the characteristic polynomial at $t = -1$ gives the number of full dimensional regions of the ambient space defined by the arrangement (see [18])

$$|\chi(-1)| = 46.$$

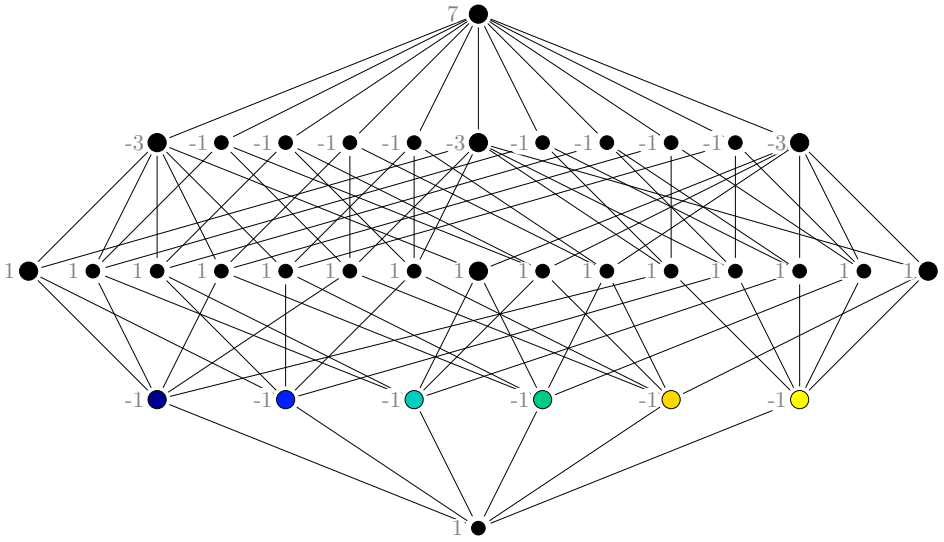


Figure 5: Intersection poset of the boundary pieces of $\mathcal{M}$. The lowest node is the ambient space $\mathbb{R}^8$. At the first level are the six boundary pieces. At the second level are the 15 pairwise intersections. The enlarged nodes are $\mathcal{L}_{i,0} \cap \mathcal{L}_{i,1}$. The third level contains the 11 distinct codimension three intersections. The top intersection corresponds to the independence model. The nodes are labeled with their Möbius function value.

For comparison, a generic four dimensional central arrangement of six hyperplanes defines 52 regions. Ours is a central arrangement (the origin is in all hyperplanes) hence all 46 regions are unbounded cones. Of the 46 regions the model $\mathcal{M}$ occupies 44. The model $\mathcal{M}_{3,2}$ occupies four of the regions.

Since the six boundary pieces (9) are linear equations in log probability space, they define exponential families. For instance, the exponential family $\mathcal{L}_{1,0}$ consists of all distributions whose log-probabilities have a vanishing inner product with $[1,-1,-1,1,0,0,0,0]^\top$. A sufficient statistic is any set of vectors spanning the kernel of this vector. Since intersections of exponential families are exponential families, each element in the intersection poset in Figure 5 is also an exponential family.

6. Maximum likelihood

In this section we give a closed-form formula for maximum likelihood estimation to the model $\mathcal{M}$. We also find the distributions whose divergence to the model is greatest.

Consider an empirical probability distribution coming from some data. The maximum likelihood estimation problem asks for the distribution in a statistical model with smallest Kullback-Leibler (KL) divergence to the data distribution. The KL divergence from p to q is defined as $D(p\|q) := \sum_x p_x \log \frac{p_x}{q_x}$, where x ranges over the possible states of p and q. This is zero if and only if $p = q$ and it is set to $+\infty$ when $\text{supp}(p) \not\subseteq \text{supp}(q)$. The distributions in the closure of a model that minimize the KL divergence are called *reverse information projections* (rI-projections) [7]. In general they are not unique, but for exponential families they are.

6.1. Reversed information projections

To study the maximum likelihood estimation problem for the model $\mathcal{M}$, we first find the rI-projections to each boundary piece of the model. We use the description of the boundary pieces as exponential families from Section 5. Proposition 5.1 means we only need to consider projections onto the six boundary pieces, not onto the entire intersection poset (as we would have to for $\mathcal{M}_{3,2}$, see [1]). For a distribution $p \in \Delta_7 \backslash \mathcal{M}$, each rI-projection will lie on one of the boundary pieces, and there is at most one projection point in each boundary piece. Taking the projection that minimizes divergence, over the six boundary pieces, gives the rI-projection to the whole model.

Let $\mathcal{P}_{i,j}$ be the toric hypersurface in the simplex obtained by exponentiating the hyperplane $\mathcal{L}_{i,j}$ in log-probability space and normalizing. The following proposition concerns maximum likelihood estimation for that toric model.

Proposition 6.1. *The unique rI-projection of $p \in \Delta_7$ onto $\mathcal{P}_{1,0}$, denoted $p_{\mathcal{P}_{1,0}}$, is found by taking the best rank one approximation in the slice p_{0jk}, $j,k \in \{0,1\}$, and leaving the other slice unchanged. In symbols,*

$$p_{\mathcal{P}_{1,0}}(X) = \begin{cases} p(X_2|X_1)p(X_3|X_1)p(X_1), & X_1 = 0 \\ p(X), & X_1 = 1 \end{cases},$$

where X is the random variable on state space $\{0,1\}^3$ and X_i is its ith coordinate. The divergence from p to $\mathcal{P}_{1,0}$ is

$$D(p\|\mathcal{P}_{1,0}) = p(X_1 = 0) \cdot I_p(X_2; X_3|X_1 = 0),$$

where $I_p(X_2; X_3|X_1 = 0) = D(p(X_2X_3|X_1 = 0)\|p(X_2|X_1 = 0)p(X_3|X_1 = 0))$ is the conditional mutual information of the two variables X_2 and X_3, given $X_1 = 0$. The rI-projections to the five other pieces follow analogously.

Proof. This follows applying [15, Lemma 3.2] to the exponential family described in Proposition 5.1 and using the fact that the rI-projection of a distribution to an independence model is given by the product of its marginals. □

The distributions whose rI-projections to $\mathcal{P}_{1,0}$ coincide are those with the same values p_{1jk}, $j,k \in \{0,1\}$ and fixed marginals on p_{0jk}, $j,k \in \{0,1\}$. The rI-projection to the entire model is the boundary projection with smallest divergence value. It has divergence

$$D(p\|\mathcal{M}) = \min_{i=1,2,3,\ j=0,1} D(p\|\mathcal{P}_{i,j}).$$

The rI-projection of any p to an exponential family is unique, so there are at most six rI-projections to $\mathcal{M}$.

Remark 6.2. For the $\mathcal{M}_{3,3}$ and $\mathrm{RBM}_{3,2}$ parametrizations of $\mathcal{M}$, each rI-projection may be realized by several distinct choices of the parameters. This implies that there are several choices of parameters associated with each local maximizer of the likelihood function.

6.2. Divergence maximizers

The maximum divergence to a statistical model is a measure of the representational power of that model. The uniform distribution on the sets of vectors with even or odd parity need the maximum number of components to be arbitrarily well approximated by a mixture of products distribution (see [12]). Here, we show that these parity distributions have the largest divergence to the model $\mathcal{M}$.

Proposition 6.3. *The maximum divergence to $\mathcal{M}$ is $\frac{1}{2}\log 2$. The maximizers are $u^+ := \frac{1}{4}(\delta_{000}+\delta_{011}+\delta_{101}+\delta_{110})$ and $u^- := \frac{1}{4}(\delta_{001}+\delta_{010}+\delta_{100}+\delta_{111})$. There are six rI-projections of u^+, one in each boundary piece:*

$$
\begin{aligned}
u^+_{\mathcal{P}_{1,0}} &= \frac{1}{8}(\delta_{000}+\delta_{001}+\delta_{010}+\delta_{011}) + \frac{1}{4}(\delta_{101}+\delta_{110})\\
u^+_{\mathcal{P}_{1,1}} &= \frac{1}{8}(\delta_{100}+\delta_{101}+\delta_{110}+\delta_{111}) + \frac{1}{4}(\delta_{011}+\delta_{000})\\
u^+_{\mathcal{P}_{2,0}} &= \frac{1}{8}(\delta_{000}+\delta_{001}+\delta_{100}+\delta_{101}) + \frac{1}{4}(\delta_{011}+\delta_{110})\\
u^+_{\mathcal{P}_{2,1}} &= \frac{1}{8}(\delta_{010}+\delta_{011}+\delta_{110}+\delta_{111}) + \frac{1}{4}(\delta_{000}+\delta_{101})\\
u^+_{\mathcal{P}_{3,0}} &= \frac{1}{8}(\delta_{000}+\delta_{010}+\delta_{100}+\delta_{110}) + \frac{1}{4}(\delta_{011}+\delta_{101})\\
u^+_{\mathcal{P}_{3,1}} &= \frac{1}{8}(\delta_{001}+\delta_{011}+\delta_{101}+\delta_{111}) + \frac{1}{4}(\delta_{000}+\delta_{110}).
\end{aligned}
$$

The projection points of u^- are given in a similar way.

Proof. Proposition 6.1 shows that the indicated distributions are the rI-projections of u^+ onto the individual boundary pieces of $\mathcal{M}$. There can be no more than six projection points and hence we have a complete list. The fact that $\frac{1}{2}\log 2$ is the maximum possible divergence to $\mathcal{M}$ follows from upper bounds for mixtures of products and RBMs given in [16]. Both u^+ and u^- attain this upper bound.

Now we show that u^+ and u^- are the only divergence maximizers. Assume without loss of generality that some maximizer p has an rI-projection onto $\mathcal{M}$ in $\mathcal{P}_{1,0}$. Then $D(p\|\mathcal{P}_{1,0}) = p(X_1=0)I_p(X_2;X_3|X_1=0) \leq D(p\|\mathcal{P}_{1,1}) = p(X_1=1)I_p(X_2;X_3|X_1=1) \leq (1-p(X_1=0))\log 2$. The last inequality follows since, for two binary variables, the mutual information is maximized by a uniform distribution on strings of Hamming distance 2 (see [3]). The maximum value $\frac{1}{2}\log 2$ is attained only if $p(X_1=0)=p(X_1=1)=\frac{1}{2}$ and both $p(X_2X_3|X_1=0)$ and $p(X_2X_3|X_1=1)$ are uniform on pairs of Hamming distance 2. If these two conditional distributions were equal, then $p\in\mathcal{M}$, and p is not a divergence maximizer. Hence the pairs are different. This shows that p is a uniform distribution on 4 strings of equal parity. □

Remark 6.4. Proposition 6.3 shows that the upper bound on the maximum divergence to mixtures of products and RBMs from [16, Theorems 1 and 2] is tight in the case of $\mathcal{M}_{3,3}$ and $\mathrm{RBM}_{3,2}$. Moreover it shows that for a given data point, $\mathrm{RBM}_{3,2}$ can have up

to 6 global maximizers of the likelihood, and that generically this will be the number of local maximizers.

An interesting question is whether we can characterize the points in the probability simplex that project to the different boundary pieces of the model. That is, to provide a *decision boundary* separating the regions of the simplex that are closer to each part of the model, with respect to the KL divergence. In our case, these decision boundaries are neither linear families nor exponential families.

7. Visualization in three dimensions

In [17, Figure 3], a first attempt was made to visualize the model $\mathcal{M}$. In this section, we explain how to draw the seven-dimensional model $\mathcal{M}$ using a three dimensional figure. We make use of the following change of basis (corresponding to the basis of characters) in the log-probability coordinates:

$$\begin{pmatrix} m_{\emptyset} \\ m_{\{3\}} \\ m_{\{2\}} \\ m_{\{2,3\}} \\ m_{\{1\}} \\ m_{\{1,3\}} \\ m_{\{1,2\}} \\ m_{\{1,2,3\}} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{pmatrix} \begin{pmatrix} l_{000} \\ l_{001} \\ l_{010} \\ l_{011} \\ l_{100} \\ l_{101} \\ l_{110} \\ l_{111} \end{pmatrix}.$$

The boundary pieces of the model can be written in terms of just four of these coordinates:

$$\begin{aligned} \mathcal{L}_{1,0} &= \{m_{\{2,3\}} + m_{\{1,2,3\}} = 0\}, \quad \mathcal{L}_{1,1} = \{m_{\{2,3\}} - m_{\{1,2,3\}} = 0\}, \\ \mathcal{L}_{2,0} &= \{m_{\{1,3\}} + m_{\{1,2,3\}} = 0\}, \quad \mathcal{L}_{2,1} = \{m_{\{1,3\}} - m_{\{1,2,3\}} = 0\}, \\ \mathcal{L}_{3,0} &= \{m_{\{1,2\}} + m_{\{1,2,3\}} = 0\}, \quad \mathcal{L}_{3,1} = \{m_{\{1,2\}} - m_{\{1,2,3\}} = 0\}. \end{aligned}$$

Hence it suffices to visualize the combinations of coordinates $(m_{\{1,2\}}, m_{\{1,3\}}, m_{\{2,3\}}, m_{\{1,2,3\}})$ that lie in the model. Furthermore, if a vector satisfies the inequalities above, then so does any scalar multiple of it. This means we need consider only those $(m_{\{1,2\}}, m_{\{1,3\}}, m_{\{2,3\}}, m_{\{1,2,3\}})$ lying on the three-dimensional sphere. The value of $m_{\{1,2,3\}}$ can be found up to sign from the other three coordinates. We draw the model in coordinates

$$(\overline{m}_{\{1,2\}}, \overline{m}_{\{1,3\}}, \overline{m}_{\{2,3\}}) = \frac{(m_{\{1,2\}}, m_{\{1,3\}}, m_{\{2,3\}})}{\|(m_{\{1,2\}}, m_{\{1,3\}}, m_{\{2,3\}}, m_{\{1,2,3\}})\|_2}, \tag{10}$$

with separate panels for the different signs of $m_{\{1,2,3\}}$. Figure 6 shows pieces $\mathcal{L}_{1,0}$ and $\mathcal{L}_{1,1}$. The whole model is shown in Figure 7.

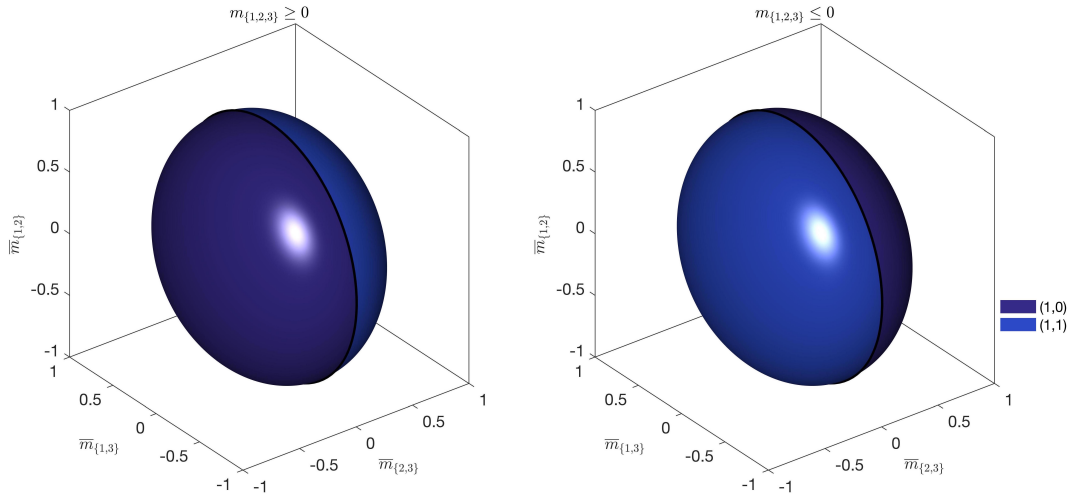


Figure 6: Illustration of two boundary pieces of the model $\mathcal{M}$. The set $\mathcal{L}_{1,0}$ is in dark blue, and $\mathcal{L}_{1,1}$ is in light blue. The points enclosed by the surface correspond to distributions in the complement of the two basic semi-algebraic sets of $\mathrm{RBM}_{3,2}$ enclosed by $\mathcal{L}_{1,0}$ and $\mathcal{L}_{1,1}$. The black line is $\{m_{\{2,3\}} = m_{\{1,2,3\}} = 0\}$, along which $\mathcal{L}_{1,0}$ and $\mathcal{L}_{1,1}$ meet. The non-linearity of the surfaces is due to normalizing with respect to the $\|\cdot\|_2$ norm.

8. Outlook

We proved the rather surprising fact that a mixture of products and a product of mixtures represent the same set of probability distributions. Although for larger models this is known not to be true in general [13], it points at a close similarity of both models.

In most previous work on the representational power of RBMs, membership in the model is determined by constructing parameters that realize certain probability distributions. In contrast, the implicit descriptions discussed here fully characterize distributions that are in the model. As we have shown, the semi-algebraic description also allows the computation of maximum likelihood estimates and divergence maximizers, both of which appear quite difficult to obtain via other methods.

The natural next step is to extend the analysis to larger models. However, the description for larger models involves complicated equality constraints. For example, in [8] the Zariski closure of the model $\mathrm{RBM}_{4,2}$ is found. It is the zero set of a single degree 110 polynomial with at least 17,214,912 terms. The binomial inequalities we obtain here are more tractable.

In light of this, it appears natural to consider approximate descriptions of larger RBM models in terms of inequality constraints only. A relaxation of larger statistical models, given in terms of inequalities only, would provide lower bounds on the maximal divergence and the minimal size of universal approximators.

In [2] the authors show that the model $\mathcal{M}_{n,2}$ consists of supermodular distributions with flattening rank at most two. Distributions in larger RBM models are Hadamard products of non-negative tensors of rank at most two (products of tensors proportional to distributions in $\mathcal{M}_{n,2}$). Ignoring the equations, we have the set of supermodular tensors,

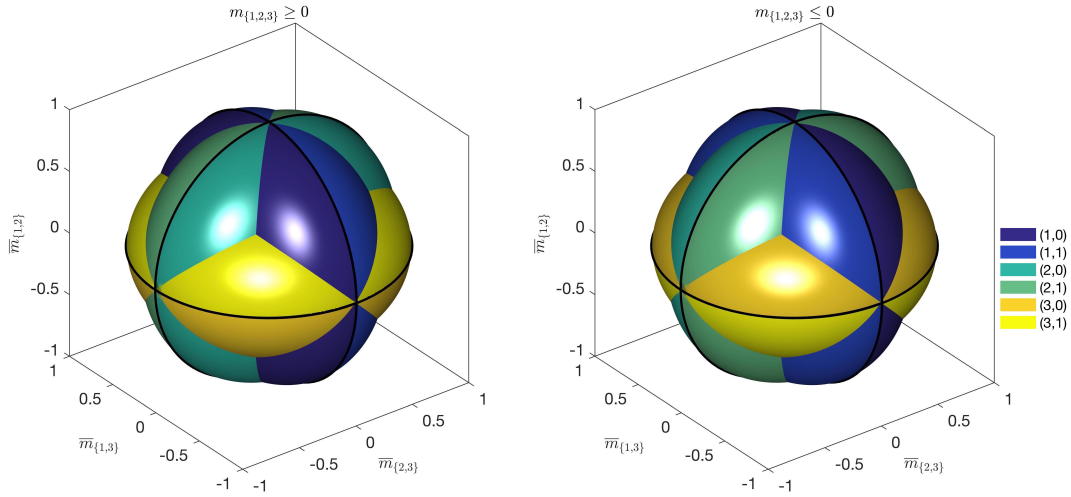


Figure 7: Illustration of $\mathcal{M}$ in the (10) coordinates. The model occupies the space inside the three-sphere that is outside any of the blue, green, or yellow surfaces. The colours correspond to the six boundary pieces of the model. Within each orthant, the part of the sphere outside all three surfaces is a triangular bipyramid. Four bipyramids make up the model $\mathcal{M}_{3,2}$.

which consists of basic semi-algebraic sets satisfying binomial quadratic inequalities as in (3). Hence the algebraic boundary of Hadamard products of supermodular tensors is again a union of exponential families, for which we may hope to obtain maximum likelihood estimates in closed form.

ACKNOWLEDGEMENTS: We are grateful to Bernd Sturmfels for fruitful discussions. GM acknowledges support from the Erwin Schrödinger Institute. This project received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 757983).

References

[1] 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. *arXiv:1710.01696*, 2017.

[2] E. S. Allman, J. A. Rhodes, B. Sturmfels, and P. Zwiernik. Tensors of nonnegative rank two. *Linear Algebra and its Applications*, 473:37 – 53, 2015. Special issue on Statistics.

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

[4] N. Beerenwinkel, L. Pachter, and B. Sturmfels. Epistasis and shapes of fitness landscapes. *Statistica Sinica*, 17(4):1317–1342, 2007.

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

[6] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[7] I. Csiszár and P. C. Shields. Information theory and statistics: A tutorial. *Foundations and Trends in Communications and Information Theory*, 1(4):417–528, 2004.

[8] 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.

[9] E. Gawrilow and M. Joswig. Polymake: a framework for analyzing convex polytopes. *Polytopes—combinatorics and computation*, pages 43–73, 1997.

[10] P. Huggins, B. Sturmfels, J. Yu, and D. Yuster. The hyperdeterminant and triangulations of the 4-cube. *Mathematics of Computation*, 77:1653–1679, 2008.

[11] S. Lauritzen. *Graphical Models*. Oxford Statistical Sci. Ser. Clarendon Press, Oxford, 1996.

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

[13] 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.

[14] G. Montúfar and J. Rauh. Hierarchical models as marginals of hierarchical models. *International Journal of Approximate Reasoning*, 88:531–546, 2017.

[15] 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. Curran Associates, Inc., 2011.

[16] G. Montúfar, J. Rauh, and N. Ay. *Maximal Information Divergence from Statistical Models Defined by Neural Networks*, pages 759–766. Springer, Berlin, Heidelberg, 2013.

[17] A. Seigal. The algebraic statistics of an Oberwolfach workshop. Snapshots of modern mathematics from Oberwolfach, 2017.

[18] R. P. Stanley. *Geometric Combinatorics*, chapter An Introduction to Hyperplane Arrangements, pages 389–496. Number 13 in IAS/Park City Math. Ser. AMS, Providence, RI, 2007.

[19] S. Sullivant. *Algebraic Statistics*. Book, to appear, draft copy, 2017.

[20] P. Zwiernik. *Semialgebraic statistics and latent tree models*, volume 146. Monographs on Statistics and Applied Probability, Chapman & Hall/CRC, Boca Raton, FL, 2016.