Contemporary Mathematics
Volume **516**, 2010

Geometry of the Restricted Boltzmann Machine

María Angélica Cueto, Jason Morton, and Bernd Sturmfels

Abstract. The restricted Boltzmann machine is a graphical model for binary random variables. Based on a complete bipartite graph separating hidden and observed variables, it is the binary analog to the factor analysis model. We study this graphical model from the perspectives of algebraic statistics and tropical geometry, starting with the observation that its Zariski closure is a Hadamard power of the first secant variety of the Segre variety of projective lines. We derive a dimension formula for the tropicalized model, and we use it to show that the restricted Boltzmann machine is identifiable in many cases. Our methods include coding theory and geometry of linear threshold functions.

1. Introduction

A primary focus in algebraic statistics is the study of statistical models that can be represented by polynomials in the model parameters. This class of algebraic statistical models includes graphical models for both Gaussian and discrete random variables [**13, 15**]. In this article we study a family of binary graphical models with hidden variables. The underlying graph is the complete bipartite graph $K_{k,n}$.

The k white nodes in the top row of Figure 1.1 represent hidden random variables. The n black nodes in the bottom row represent observed random variables. The restricted Boltzmann machine (RBM) is the undirected graphical model for binary random variables specified by this bipartite graph. We identify the model with the set M_n^k of its joint distributions inside the probability simplex Δ_{2^n-1}.

The graphical model for Gaussian random variables represented by Figure 1.1 is the *factor analysis* model, whose algebraic properties were studied in [**3, 10, 14**]. Thus, the restricted Boltzmann machine is the binary undirected analog of factor analysis. Our aim here is to study this model from the perspectives of algebra and geometry. Unlike in the factor analysis study [**14**], an important role will now be played by *tropical geometry* [**28**]. This was already seen for $n = 4$ and $k = 2$ in the solution by Cueto and Yu [**8**] of the implicitization challenge in [**15**, Problem 7.7].

2010 *Mathematics Subject Classification.* 62E10, 68T05, 14Q15, 51M20.

Key words and phrases. Algebraic statistics, tropical geometry, deep belief network, Hadamard product, secant variety, Segre variety, inference function, linear threshold function.

María Angélica Cueto was supported by a UC Berkeley Chancellor's Fellowship. Jason Morton was supported in part by DARPA grant HR0011-05-1-0007 and NSF grant DMS-0354543. Bernd Sturmfels was supported in part by NSF grants DMS-0456960 and DMS-0757236.

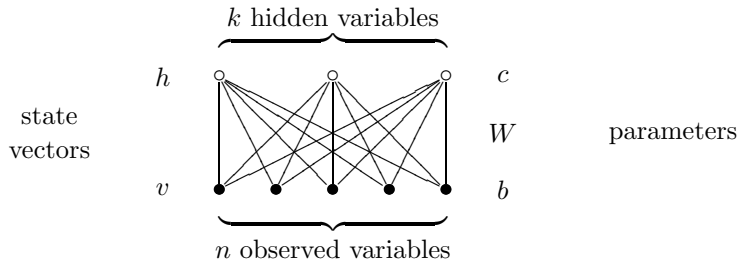


Figure 1.1: Graphical representation of the restricted Boltzmann machine.

The restricted Boltzmann machine has been the subject of a recent resurgence of interest due to its role as the building block of the deep belief network. Deep belief networks are designed to learn feature hierarchies to automatically find high-level representations for high-dimensional data. A deep belief network comprises a stack of restricted Boltzmann machines. Given a piece of data (state of the lowest visible variables), each layer's most likely hidden states are treated as data for the next layer. A new effective training methodology for deep belief networks, which begins by training each layer in turn as an RBM using contrastive divergence, was introduced by Hinton et al. [**18**]. This method led to many new applications in general machine learning problems including object recognition and dimensionality reduction [**19**]. While promising for practical applications, the scope and basic properties of these statistical models have only begun to be studied. For example, Le Roux and Bengio [**23**] showed that any distribution with support on r visible states may be arbitrarily well approximated provided there are at least $r+1$ hidden nodes. Therefore, any distribution can be approximated with 2^n+1 hidden nodes.

The question which started this project is whether the restricted Boltzmann machine model is *identifiable*, i.e. whether the parametrization of the model is locally one-to-one. The dimension of the fully observed binary graphical model on $K_{k,n}$ is equal to $nk+n+k$, the number of nodes plus the number of edges. We conjecture that this dimension is preserved under the projection corresponding to the algebraic elimination of the k hidden variables. Here is the precise statement:

CONJECTURE 1.1. The restricted Boltzmann machine has the expected dimension, i.e. M_n^k is a semialgebraic set of dimension $\min\{nk+n+k, 2^n-1\}$ in Δ_{2^n-1}.

This conjecture is shown to be true in many special cases. In particular, it holds for all k when $n+1$ is a power of 2. This is a consequence of the following:

THEOREM 1.1. *The restricted Boltzmann machine has the expected dimension* $\min\{nk+n+k, 2^n-1\}$ *when* $k \leq 2^{n-\lceil \log_2(n+1)\rceil}$ *and when* $k \geq 2^{n-\lfloor \log_2(n+1)\rfloor}$.

We note that Theorem 1.1 covers most cases of restricted Boltzmann machines as used in practice, as those generally satisfy $k \leq 2^{n-\lceil \log_2(n+1)\rceil}$. In particular, we conclude that the model is identifiable in these cases. The case of large k is primarily of theoretical interest and has been studied recently in [**23**].

This paper is organized as follows. In Section 2 we introduce four geometric objects, namely, the RBM model, the RBM variety, the tropical RBM model, and

the tropical RBM variety, and we formulate a strengthening of Conjecture 1.1. Section 3 is concerned with the case $k = 1$. Here the RBM variety is the variety of secant lines of the Segre variety $(\mathbb{P}^1)^n \subset \mathbb{P}^{2^n-1}$. The general case $k > 1$ arises from that secant variety by way of a construction we call the *Hadamard product of projective varieties*, as shown in Proposition 2.1. In Section 4 we analyze the tropical RBM model, we establish a formula for its dimension (Theorem 4.1), and we draw on results from coding theory to derive Theorem 1.1 and Table 4.1. In Section 5 we study the piecewise-linear map that parameterizes the tropical RBM model. The *inference functions* of the model (in the sense of [**16, 28**]) are k-tuples of *linear threshold functions*. We discuss the number of these functions. Figure 5.2 shows the combinatorial structure of the tropical RBM model for n=3 and k=1.

2. Algebraic varieties, Hadamard product and tropicalization

We begin with an alternative definition of the restricted Boltzmann machine. This "machine" is a statistical model for binary random variables where n of the variables are visible and k of the variables are hidden. The states of the hidden and visible variables are written as binary vectors $h \in \{0,1\}^k$ and $v \in \{0,1\}^n$ respectively. We introduce $nk + n + k$ model parameters, namely, the entries of a real $k \times n$ matrix W and the entries of two vectors $b \in \mathbb{R}^n$ and $c \in \mathbb{R}^k$, and we set

$$\psi(v,h) \;=\; \exp(h^\top W v + b^\top v + c^\top h). \tag{2.1}$$

The probability distribution on the visible random variables in our model equals

$$p(v) \;=\; \frac{1}{Z} \cdot \sum_{h \in \{0,1\}^k} \psi(v,h), \tag{2.2}$$

where $Z = \sum_{v,h} \psi(v,h)$ is the *partition function*. We denote by M_n^k the subset of the open probability simplex Δ_{2^n-1} consisting of all such distributions $(p(v) : v \in \{0,1\}^n)$ as the parameters W, b and c run over $\mathbb{R}^{k\times n}$, $\mathbb{R}^n$ and $\mathbb{R}^k$ respectively.

In what follows we refer to M_n^k as the *RBM model* with n visible nodes and k hidden nodes. It coincides with the binary graphical model associated with the complete bipartite graph $K_{k,n}$ as described in the Introduction. This is indicated in Figure 1.1 by the labeling with the states v, h and the model parameters c, W, b.

The parameterization in (2.1) is not polynomial because it involves the exponential function. However, it is equivalent to the polynomial map obtained by replacing each model parameter by its value under the exponential function:

$$\gamma_i = \exp(c_i)\,, \;\; \omega_{ij} = \exp(W_{ij})\,, \;\; \beta_j = \exp(b_j).$$

This coordinate change translates (2.1) into the squarefree monomial

$$\psi(v,h) \;=\; \prod_{i=1}^{k} \gamma_i^{h_i} \cdot \prod_{i=1}^{k} \prod_{j=1}^{n} \omega_{ij}^{h_i v_j} \cdot \prod_{j=1}^{n} \beta_j^{v_j},$$

and we see that the probabilities in (2.2) can be factored as follows:

$$p(v) \;=\; \frac{1}{Z}\, \beta_1^{v_1} \beta_2^{v_2} \cdots \beta_n^{v_n} \prod_{i=1}^{k} \bigl(1 + \gamma_i\, \omega_{i1}^{v_1} \omega_{i2}^{v_2} \cdots \omega_{in}^{v_n}\bigr) \qquad \text{for} \quad v \in \{0,1\}^n. \tag{2.3}$$

The RBM model M_n^k is the image of the polynomial map $\mathbb{R}_{>0}^{nk+k+n} \to \Delta_{2^n-1}$ whose vth coordinate equals (2.3). The Tarski-Seidenberg Theorem from real algebraic geometry implies that M_n^k is a semialgebraic subset of Δ_{2^n-1}.

When faced with a high-dimensional semialgebraic set in statistics, it is often useful to simplify the situation by disregarding all inequalities and by replacing the real numbers $\mathbb{R}$ by the complex numbers $\mathbb{C}$. This leads us to considering the Zariski closure V_n^k of the RBM model M_n^k. This is the algebraic variety in the complex projective space $\mathbb{P}^{2^n-1}$ parameterized by (2.3). We call V_n^k the *RBM variety*.

Given any two subvarieties X and Y of a projective space $\mathbb{P}^m$, we define their *Hadamard product* $X * Y$ to be the closure of the image of the rational map

$$X \times Y \dashrightarrow \mathbb{P}^m\,,\ (x,y) \mapsto (x_0y_0 : x_1y_1 : \ldots : x_my_m).$$

For any projective variety X, we may consider its Hadamard square $X^{[2]} = X * X$ and its higher Hadamard powers $X^{[k]} = X * X^{[k-1]}$. If M is a subset of the open simplex Δ_{m-1} then its Hadamard powers $M^{[k]}$ are also defined by componentwise multiplication followed by rescaling so that the coordinates sum to one. This construction is compatible with taking Zariski closures, i.e. we have $\overline{M^{[k]}} = \overline{M}^{[k]}$.

In the next section we shall take a closer look at the case $k = 1$, and we shall recognize V_n^1 as a secant variety and M_n^1 as a phylogenetic model. Here, we prove that the case of $k > 1$ hidden nodes reduces to $k = 1$ using Hadamard powers.

PROPOSITION 2.1. *The RBM variety and model factor as Hadamard powers:*

$$V_n^k = (V_n^1)^{[k]} \quad \text{and} \quad M_n^k = (M_n^1)^{[k]}.$$

PROOF. A strictly positive vector p with coordinates $p(v)$ as in (2.3) admits a componentwise factorization into similar vectors for $k = 1$, and, conversely, the componentwise product of k probability distributions in M_n^1 becomes a distribution in M_n^k after division by the partition function. Hence $M_n^k = (M_n^1)^{[k]}$ in Δ_{2^n-1}. The equation $V_n^k = (V_n^1)^{[k]}$ follows by passing to the Zariski closure in $\mathbb{P}^{2^n-1}$. □

The emerging field of *tropical mathematics* is predicated on the idea that $\log(\exp(x) + \exp(y))$ is approximately equal to $\max(x, y)$ when x and y are quantities of different scale. For a first introduction see [**32**], and for further reading see [**5, 9, 12, 25**] and references therein. The process of passing from ordinary arithmetic to the max-plus algebra is known as *tropicalization*. The same approximation motivates the definition of the *softmax* function in the neural networks literature. A statistical perspective is offered in work by Pachter and the third author [**29, 28**].

If $q(v)$ approximates $\log(p(v))$ in the sense of tropical mathematics, and if we disregard the global additive constant $-\log Z$, then (2.2) translates into the formula

$$q(v) \;=\; \max\bigl\{\, h^\top W v + b^\top v + c^\top h \;:\; h \in \{0,1\}^k \,\bigr\}. \tag{2.4}$$

This expression is a piecewise-linear concave function $\mathbb{R}^{nk+n+k} \to \mathbb{R}$ on the space of model parameters (W, b, c). As v ranges over $\{0,1\}^n$, there are 2^n such concave functions, and these form the coordinates of a piecewise-linear map

$$\Phi : \mathbb{R}^{nk+n+k} \to \mathbb{TP}^{2^n-1}. \tag{2.5}$$

Here $\mathbb{TP}^{2^n-1}$ denotes the *tropical projective space* $\mathbb{R}^{2^n}/\mathbb{R}(1,1,\ldots,1)$, as in [**5, 12**]. The image of the map Φ is denoted TM_n^k and is called the *tropical RBM model*. The map Φ is the *tropicalization* of the given parameterization of the RBM model. It is our objective to investigate its geometric properties.

This situation fits precisely into the general scheme of parametric maximum a posterior (MAP) inference introduced in [**28**] and studied in more detail by Elizalde and Woods [**16**]. In Section 5 below, we discuss the statistical relevance of the map

Φ and we examine its geometric properties. Of particular interest are the domains of linearity of Φ, and how these are mapped onto the cones of the model TM_n^k.

Finally, we define the *tropical RBM variety* TV_n^k to be the tropicalization of the RBM variety V_n^k. As explained in [**29**, §3.4] and [**28**, §3], the tropical variety TV_n^k is the intersection in $\mathbb{TP}^{2^n-1}$ of all the tropical hypersurfaces $\mathcal{T}(f)$ where f runs over *all* polynomials that vanish on V_n^k (or on M_n^k). By definition, $\mathcal{T}(f)$ is the union of all codimension one cones in the normal fan of the Newton polytope of f. If the homogeneous prime ideal of the variety V_n^k were known then the tropical variety TV_n^k could in theory be computed using the algorithms in [**5**] which are implemented in the software `Gfan` ([**21**]). However, this prime ideal is not known in general. In fact, even for small instances, its computation is very hard and relies primarily on tropical geometry techniques such as the ones developed in [**8**]. For instance, the main result in [**8**] states that the RBM variety V_4^2 is a hypersurface of degree 110 in $\mathbb{P}^{15}$, and it remains a challenge to determine a formula for the defining irreducible polynomial of this hypersurface. To appreciate this challenge, note that the number of monomials in the relevant multidegree equals 5 529 528 561 944.

Here is a brief summary of the four geometric objects we have introduced:

- The semialgebraic set $M_n^k \subset \Delta_{2^n-1}$ of probability distributions represented by the restricted Boltzmann machine. We call M_n^k the *RBM model*.
- The Zariski closure V_n^k of the RBM model M_n^k. This is an algebraic variety in the complex projective space $\mathbb{P}^{2^n-1}$. We call V_n^k the *RBM variety*.
- The image TM_n^k of the tropicalized parameterization Φ. This is the subset of $\mathbb{TP}^{2^n-1}$ consisting of all optimal score value vectors in the MAP inference problem for the RBM. We call TM_n^k the *tropical RBM model*.
- The tropicalization TV_n^k of the variety V_n^k. This is a tropical variety in the tropical projective space $\mathbb{TP}^{2^n-1}$. We call TV_n^k the *tropical RBM variety*.

We have inclusions $M_n^k \subset V_n^k$ and $TM_n^k \subset TV_n^k$. The latter inclusion is the content of the second statement in [**28**, Theorem 2]. We shall see that both inclusions are strict even for $k = 1$. For example, M_3^1 is a proper subset of $V_3^1 \cap \Delta_7 = \Delta_7$ since points in this set must satisfy the inequality $\sigma_{12}\sigma_{13}\sigma_{23} \geq 0$ as indicated in Theorem 3.2 below. Likewise, TM_3^1 is a proper subfan of $\mathbb{TP}^7 = TV_3^1$. This subfan will be determined in our discussion of the secondary fan structure in Example 5.1.

The dimensions of our four geometric objects satisfy the following chain of equations and inequalities:

$$\begin{aligned} \dim(TM_n^k) \,\leq\, \dim(TV_n^k) \,=\, \dim(V_n^k) \\ = \dim(M_n^k) \,\leq\, \min\{nk+n+k, 2^n-1\}. \end{aligned} \tag{2.6}$$

Here, the tropical objects TM_n^k and TV_n^k are polyhedral fans, and by their dimension we mean the dimension of any cone of maximal dimension in the fan. When speaking of the dimension of V_n^k we mean the Krull dimension of the projective variety, and for the model M_n^k we mean its dimension as a semialgebraic set.

The leftmost inequality in (2.6) holds because $TM_n^k \subset TV_n^k$. The left equality holds by the Bieri-Groves Theorem (cf. [**12**, Theorem 4.5]) which ensures that every irreducible variety has the same dimension as its tropicalization.

Every polynomial function that vanishes on the image of the map p in (2.3) also vanishes on V_n^k. This means that the model M_n^k is *Zariski dense* in the variety V_n^k. From this we conclude the validity of the second equality in (2.6). Finally, the

rightmost inequality in (2.6) is seen by counting parameters in the definition (2.1)–(2.2) of the RBM model M_n^k, and by bounding its dimension by the dimension of the ambient space Δ_{2^n-1}.

We conjecture that both of the inequalities in (2.6) are actually equalities:

CONJECTURE 2.1. The tropical RBM model has the expected dimension, i.e. TM_n^k is a polyhedral fan of dimension $\min\{nk+n+k, 2^n-1\}$ in $\mathbb{TP}^{2^n-1}$.

In light of the inequalities (2.6), Conjecture 2.1 implies Conjecture 1.1. In Section 4 we shall prove some special cases of these conjectures, including Theorem 1.1.

3. The first secant variety of the n-cube

We saw in Proposition 2.1 that the RBM for $k \geq 2$ can be expressed as the Hadamard power of the RBM for $k = 1$. Therefore, it is crucial to understand the model with one hidden node. In this section we fix $k = 1$ and we present an analysis of that case. In particular, we shall give a combinatorial description of the fan TM_n^1 which shows that it has dimension $2n+1$, as stated in Conjecture 2.1.

We begin with a reparameterization of our model that describes it as a secant variety. Let λ, $\delta_1, \ldots, \delta_n$, $\epsilon_1, \ldots, \epsilon_n$ be real parameters which range over the open interval $(0,1)$, and consider the polynomial map $p : (0,1)^{2n+1} \to \Delta_{2^n-1}$ whose coordinates are given by

$$p(v) \;=\; \lambda \prod_{i=1}^{n} \delta_i^{1-v_i}(1-\delta_i)^{v_i} \;+\; (1-\lambda) \prod_{i=1}^{n} \epsilon_i^{1-v_i}(1-\epsilon_i)^{v_i} \quad \text{for } v \in \{0,1\}^n. \tag{3.1}$$

PROPOSITION 3.1. *The image of p coincides with the RBM model M_n^1.*

PROOF. Recall the parameterization (2.3) of the RBM model M_n^1 from Section 2:

$$p(v) \;=\; \frac{1}{Z}\,\beta_1^{v_1}\beta_2^{v_2}\cdots\beta_n^{v_n}\bigl(1+\gamma\,\omega_1^{v_1}\omega_2^{v_2}\cdots\omega_n^{v_n}\bigr) \quad \text{for} \quad v \in \{0,1\}^n. \tag{3.2}$$

We define a bijection between the parameter spaces $\mathbb{R}_{>0}^{2n+1}$ and $(0,1)^{2n+1}$ as follows:

$$\beta_i = \frac{1-\delta_i}{\delta_i} \quad \text{and} \quad \omega_i = \frac{\delta_i}{1-\delta_i}\frac{1-\epsilon_i}{\epsilon_i} \quad \text{for } i = 1,2,\ldots,n,$$

$$\gamma \;=\; Z(1-\lambda)\epsilon_1\epsilon_2\cdots\epsilon_n \quad \text{where} \quad Z = (\lambda\delta_1\delta_2\cdots\delta_n)^{-1}.$$

This substitution is invertible and it transforms (3.2) into (3.1). □

Proposition 3.1 shows that M_n^1 is the first mixture of the independence model for n binary random variables. In phylogenetics, it coincides with the *general Markov model on the star tree* with n leaves. A semi-algebraic characterization of that model follows as a special case from recent results of Zwiernik and Smith [**34**]. We shall present and discuss their characterization in Theorem 3.2 below.

First, however, we remark that the Zariski closure of a mixture of an independence model is a secant variety of the corresponding Segre variety. This fact is well-known (see e.g. [**15**, §4.1]) and is here easily seen from (3.1). We conclude:

COROLLARY 3.1. *The first RBM variety V_n^1 coincides with the first secant variety of the Segre embedding of the product of projective lines $(\mathbb{P}^1)^n$ into $\mathbb{P}^{2^n-1}$, and the first tropical RBM variety TV_n^1 is the tropicalization of that secant variety.*

We next describe the equations defining the first secant variety V_n^1. The coordinate functions $p(v)$ are the entries of an n-dimensional table of format $2\times 2\times\cdots\times 2$. For each set partition $\{1,2,\ldots,n\} = A \sqcup B$ we can write this table as an ordinary two-dimensional matrix of format $2^{|A|}\times 2^{|B|}$, with rows indexed by $\{0,1\}^A$ and columns indexed by $\{0,1\}^B$. These matrices are the *flattenings* of the $2\times 2\times\cdots\times 2$-table. Pachter and Sturmfels [**28**, Conjecture 13] conjectured that the homogeneous prime ideal of the projective variety $V_n^1 \subset \mathbb{P}^{2^n-1}$ is generated by the 3×3-minors of all the flattenings of the table $(p(v))_{v\in\{0,1\}^n}$. This conjecture has been verified computationally for $n \leq 5$. A more general form of this conjecture was stated in [**17**, §7]. The set-theoretic version of that general conjecture was proved by Landsberg and Manivel in [**22**, Theorem 5.1]. Their results imply:

THEOREM 3.1 (Landsberg-Manivel). *The projective variety $V_n^1 \subset \mathbb{P}^{2^n-1}$ is the common zero set of the 3×3-minors of all the flattenings of the table $(p(v))_{v\in\{0,1\}^n}$.*

We now come to the inequalities that determine M_n^1 among the real points of V_n^1. For any pair of indices $i,j \in \{1,2,\ldots,n\}$ we write σ_{ij} for the covariance of the two random variables X_i and X_j obtained by marginalizing the distribution, and we write $\Sigma = (\sigma_{ij})$ for the $n\times n$-covariance matrix. We regard Σ as a polynomial map from the simplex Δ_{2^n-1} to the space $\mathbb{R}^{\binom{n+1}{2}}$ of symmetric $n\times n$-matrices. The off-diagonal entries of the covariance matrix Σ are the 2×2-minors obtained by marginalization from the table $(p(v))$. For example, for $n = 4$ the covariances are

$$\sigma_{12} \;=\; \det\begin{pmatrix} p_{0000}+p_{0001}+p_{0010}+p_{0011} & p_{0100}+p_{0101}+p_{0110}+p_{0111} \\ p_{1000}+p_{1001}+p_{1010}+p_{1011} & p_{1100}+p_{1101}+p_{1110}+p_{1111} \end{pmatrix},$$

$$\sigma_{13} \;=\; \det\begin{pmatrix} p_{0000}+p_{0001}+p_{0100}+p_{0101} & p_{0010}+p_{0011}+p_{0110}+p_{0111} \\ p_{1000}+p_{1001}+p_{1100}+p_{1101} & p_{1010}+p_{1011}+p_{1110}+p_{1111} \end{pmatrix}, \qquad \text{etc.}$$

Zwiernik and Smith [**34**] gave a semi-algebraic characterization of the general Markov model on a trivalent phylogenetic tree in terms of covariances and moments. The statement of their characterization is somewhat complicated, so we only state a weaker necessary condition rather than the full characterization. Specifically, applying [**34**, Theorem 4.2] to the star tree on n leaves implies the following result.

COROLLARY 3.2. *If a probability distribution $p \in \Delta_{2^n-1}$ lies in the first RBM model M_n^1 then all its matrix flattenings (as in Theorem 3.1) have rank ≤ 2 and*

$$\sigma_{ij}\sigma_{ik}\sigma_{jk} \;\geq\; 0 \qquad \text{for all distinct triples } i,j,k \in \{1,2,\ldots,n\}.$$

These inequalities follow easily from the parameterization (3.2), which yields

$$\sigma_{ij} \;=\; \lambda(1-\lambda)(\delta_i-\epsilon_i)(\delta_j-\epsilon_j)\,\frac{\delta_i\delta_j}{\prod_{s=1}^n \delta_s}\,\frac{\epsilon_i\epsilon_j}{\prod_{s=1}^n \epsilon_s}.$$

This factorization also shows that the binomial relations $\sigma_{ij}\sigma_{kl} = \sigma_{il}\sigma_{jk}$ hold on M_n^1. These same binomial relations are valid for the covariances in factor analysis [**14**, Theorem 16], thus further underlining the analogies between the Gaussian case and the binary case. Theorem 20 in [**34**] extends the covariance equations $\sigma_{ij}\sigma_{kl} = \sigma_{il}\sigma_{jk}$ to a collection of quadratic binomial equations in all tree-cumulants, which in turn can be expressed in terms of higher order correlations. For the star tree, these equations are equivalent on Δ_{2^n-1} to the rank ≤ 2 constraints. However, for general tree models, the binomial equations in the tree-cumulants are necessary conditions for distributions to lie in these models.

We now turn to the tropical versions of the RBM model for $k = 1$. The variety V_n^1 is cut out by the 3×3-minors of all flattenings of the table $\big(p(v)\big)_{v\in\{0,1\}^n}$. It is known that the 3×3-minors of **one** fixed two-dimensional matrix form a tropical basis. Recall (e.g. from [**5**, §2]) that a *tropical basis* of a polynomial ideal is a generating set with property that the intersection of the corresponding tropical hypersurfaces equals the tropical variety of the ideal. The tropical basis property of the 3×3-minors is equivalent to [**11**, Theorem 6.5].

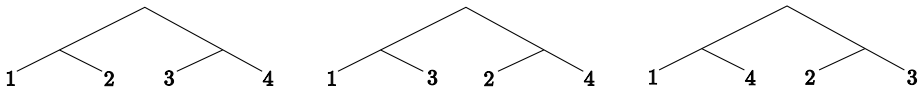
It is natural to ask whether this property continues to hold for the set of **all** 3×3-determinants in Theorem 3.1. Since each flattening of our table corresponds to a non-trivial edge split of a tree on n taxa (i.e. a partition of the set of taxa into two sets each of cardinality ≥ 2), our question can be reformulated as follows:

QUESTION 3.1. Is the tropical RBM variety TV_n^1 equal to the intersection of the tropical rank 2 varieties associated to non-trivial edge splits on a collection of trees on n taxa?

The tropical rank two varieties associated to each of the edge splits have been studied recently by Markwig and Yu [**25**]. They endow this determinantal variety with a simplicial fan structure that has the virtue of being shellable. The cones of this simplicial fan correspond to weighted bicolored trees on 2^{n-1} taxa with no monochromatic cherries. The interior points in a cone can be viewed as a matrix encoding the distances between leaves with different colors in the associated weighted bicolored tree.

Question 3.1 is void for $n \leq 3$, so the first relevant case concerns $n = 4$ taxa. We were surprised to learn that the answer is negative already in this case:

EXAMPLE 3.1. The prime ideal of the variety V_4^1 is generated by the sixteen 3×3-minors of the three flattenings of the $2\times2\times2\times2$-table p. As a statistical model, each one of the three flattenings corresponds to the graphical model associated to each one of the quartet trees (12|34), (13|24) and (14|23), as depicted in Figure 3.1.



(a) (12|34) (b) (13|24) (c) (14|23)

Figure 3.1: Quartet trees associated to the flattenings for $n = 4$.

Algebraically, each flattening corresponds to the variety cut out by the sixteen 3×3-minors of a 4×4-matrix of unknowns. These minors form a tropical basis. The tropical variety they define is a pure fan of dimension 11 in $\mathbb{TP}^{15}$ with a 6-dimensional lineality space. The simplicial fan structure on this variety given by [**25**] has the f-vector $\big(98, 1152, 4248, 6072, 2952\big)$. Combinatorially, this object is a shellable 4-dimensional simplicial complex which is the bouquet of 73 spheres. However, this determinantal variety admits a different fan structure, induced from the Gröbner fan as in [**5**], or from the fact that the sixteen 3×3-minors form a tropical basis. Its f-vector is $\big(50, 360, 1128, 1680, 936\big)$.

The tropical variety TV_4^1 is a pure fan of dimension 9 in $\mathbb{TP}^{15}$. Its lineality space has dimension 4, and the cones of various dimensions are tallied in its f-vector

$$f(TV_4^1) \quad = \quad \big(382, 3436, 11236, 15640, 7680\big).$$

Question 3.1 asks whether the 9-dimensional tropical variety TV_4^1 is the intersection of the three 11-dimensional tropical determinantal varieties associated with the three trees in Figure 3.1. The answer is "no". Using the software `Gfan` [**21**], we computed the tropical prevariety cut out by the union of all forty-eight 3×3-minors. The output is a *non-pure* polyhedral fan of dimension 10 with a 4-dimensional lineality space (the same one as of TV_4^1), having f-vector $(298, 2732, 9440, 13992, 7304, 96)$. The tropical variety TV_4^1 is a triangulation of a proper subfan, and each of the 96 10-dimensional maximal cones lies in the prevariety but not in the variety. An example of a such a vector in the relative interior of a maximal cone is

$$q \;=\; (59, 1, 80, 86, 102, 108, 107, 113, 109, 115, 100, 106, 78, 84, 21, 43).$$

(Here, coordinates are indexed in lexicographic order $p_{0000}, p_{0001}, \ldots, p_{1111}$). Given the weights q, the initial form of each 3×3-minor of each flattening is a binomial, however, the initial form of the following polynomial in the ideal of V_4^1 is the underlined monomial:

$$\begin{matrix} \underline{p_{0000}p_{0110}p_{1010}p_{1101}} - p_{0010}p_{0100}p_{1000}p_{1111} + p_{0010}p_{0100}p_{1001}p_{1110} \\ -p_{0000}p_{0110}p_{1001}p_{1110} - p_{0001}p_{0110}p_{1010}p_{1100} + p_{0000}p_{0010}p_{1100}p_{1111} \\ -p_{0000}p_{0010}p_{1101}p_{1110} + p_{0001}p_{0110}p_{1000}p_{1110}. \end{matrix}$$

Anders Jensen performed another computation, using `Gfan` and `SoPlex` [**33**], which verified that we get a tropical basis by augmenting the 3×3-minors with the above quartic and its images under the symmetry group of the 4-cube. This is a non-trivial computation because the corresponding fan structure on TV_4^1 has the f-vector

$$(37442, 321596, 843312, 880488, 321552).$$

Using the language of [**11**], we may conclude from our computational results that the notions of tropical rank and Kapranov rank disagree for 2×2×2×2-tensors. □

Last but not least, we examine the tropical model TM_n^1. This is a proper subfan of the tropical variety TV_n^1, namely, TM_n^1 is the image of the tropical morphism $\Phi : \mathbb{R}^{2n+1} \to \mathbb{TP}^{2^n-1}$ which is the specialization of (2.5) for $k = 1$. Equivalently, Φ is the tropicalization of the map (3.2), and its coordinates are written explicitly as

$$q(v) \;=\; b^\top v \;+\; \max\big\{\, 0\,,\, \omega v + c \big\}. \tag{3.3}$$

This concave function is the maximum of two linear functions. The $2n+1$ parameters are given by a column vector $b \in \mathbb{R}^n$, a row vector $\omega \in \mathbb{R}^n$, and a scalar $c \in \mathbb{R}$. A different tropical map which has the same image as Φ can be derived from (3.1). As v ranges over $\{0,1\}^n$, there are 2^n such concave functions, and these form the coordinates of the tropical morphism Φ. We note that Φ made its first explicit appearance in [**28**, Equation (10)], where it was discussed in the context of *ancestral reconstruction* in statistical phylogenetics. Subsequently, Develin [**9**] and Draisma [**12**, §7.2] introduced a tropical approach to secant varieties of toric varieties, and our model fits well into the context developed by these two authors.

REMARK 3.1. The first tropical RBM model TM_n^1 is the image of the tropical secant map for the Segre variety $(\mathbb{P}^1)^n$ in the sense of Develin [**9**] and Draisma [**12**].

The linear space for their constructions has basis $\{\sum_{\alpha\in\{0,1\}^n,\alpha_i=1} e_\alpha : i = 1,\dots,n\}$, and the underlying point configuration consists of the vertices of the n-cube.

In light of Example 3.1, it makes sense to say that the $2\times\cdots\times 2$-tensors in the tropical variety TV_n^1 are precisely those that have *Kapranov (tensor) rank* ≤ 2. This would be consistent with the results and nomenclature in [**9, 11**]. A proper subset of the tensors of Kapranov rank ≤ 2 are those that have *Barvinok (tensor) rank* ≤ 2. These are precisely the points in the first tropical RBM model TM_n^1.

We close this section by showing that TM_n^1 has the expected dimension:

PROPOSITION 3.2. *The dimension of the tropical RBM model* TM_n^1 *is* $2n+1$.

PROOF. Each region of linearity of the map Φ is defined by a partition C of $\{0,1\}^n$ into two disjoint subsets C^- and C^+, according to the condition $\omega v + c < 0$ or $\omega v + c > 0$. Thus, the corresponding region is an open convex polyhedral cone, possibly empty, in the parameter space $\mathbb{R}^{2n+1}$. It consists of all triples (b,ω,c) such that $\omega v + c < 0$ for $v \in C^-$ and $\omega v + c > 0$ for $v \in C^+$. Assuming $n \geq 3$, we can choose a partition C of $\{0,1\}^n$ such that this cone is non-empty and both C^- and C^+ affinely span $\mathbb{R}^n$. The image of the cone under the map Φ spans a space isomorphic to the direct sum of the images of $b \mapsto (b^\top v : v \in C)$ and $(\omega, c) \mapsto (\omega v + c : v \in C^+)$. Hence this image has dimension $2n+1$, as expected. □

An illustration of the proof of Proposition 3.2 is given in Figure 3.2. The technique of partitioning the vertices of the cube will be essential in our dimension computations for general k in the next section. In Section 5 we return to the small models TM_n^1 and take a closer look at their geometric and statistical properties.

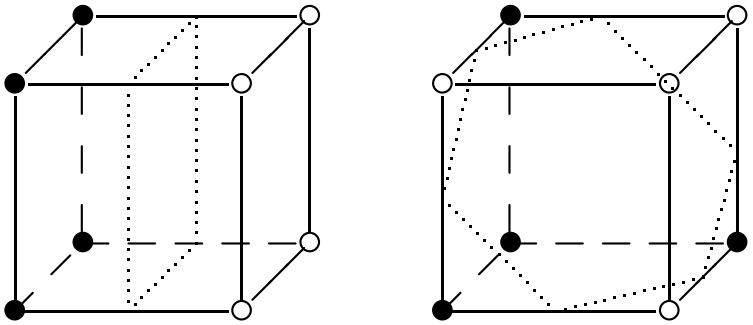


Figure 3.2: Partitions of $\{0,1\}^3$ that define non-empty cones on which Φ is linear. Here C^+ and C^- are indicated by black ($\bullet$) and white ($\circ$) vertices of the 3-cube. The slicing on the right represents a cone in the parameter space whose image under Φ is full-dimensional, while the one on the left does not.

4. The tropical model and its dimension

This section is concerned with Conjecture 2.1 which states that the tropical RBM model has the expected dimension. Namely, our aim is to show that

$$\dim(TM_n^k) \;=\; kn + k + n \qquad \text{for} \qquad k \;\leq\; \frac{2^n - 1 - n}{n+1}.$$

For $k = 1$ this is Proposition 3.2, and we now consider the general case $k \geq 2$. Our main tool towards this goal is the dimension formula in Theorem 4.1 below. As in the previous section, we study the regions of linearity of the tropical morphism Φ.

Let A denote the matrix of format $2^n \times n$ whose rows are the vectors in $\{0,1\}^n$. A subset C of the vertices of the n-cube is a *slicing* if there exists a hyperplane that has the vertices in C on the positive side and the remaining vertices of the n-cube on the other side. In the notation in the proof of Proposition 3.2, the subset C was denoted by C^+. Two examples of slicings for $n = 3$ are shown in Figure 3.2.

For any slicing C of the n-cube, let A_C be the $2^n \times (n+1)$-matrix whose rows v indexed by the vertices in C are $(1, v) \in \{0,1\}^{n+1}$ and whose other rows are all identically zero. The following result extends the argument used for Proposition 3.2.

LEMMA 4.1. *On each region of linearity, the tropical morphism Φ in (2.5) coincides with the linear map represented by a $2^n \times (nk+n+k)$-matrix of the form*

$$\mathcal{A} \;=\; \big(A \mid A_{C_1} \mid A_{C_2} \mid \cdots \mid A_{C_k} \big),$$

for some slicings $C_1, C_2, \ldots, C_k$ of the n-cube.

PROOF. The tropical map $\Phi : \mathbb{R}^{nk+n+k} \to \mathbb{TP}^{2^n-1}$ can be written as follows:

$$\Phi(W, b, c) \quad = \quad \big(\max_{h \in \{0,1\}^k} \{ h^\top (Wv + c), 0 \} \,+\, b^\top v \;\big)_{v \in \{0,1\}^n}.$$

Consider a parameter vector θ with coordinates

$$\theta := (b_1, b_2, \ldots, b_n, c_1, \omega_{11}, \ldots, \omega_{1n}, c_2, \omega_{21}, \ldots, \omega_{2n}, \ldots, c_k, \omega_{k1}, \ldots, \omega_{kn}).$$

We associate to this vector the k hyperplanes $H_i(\theta) = \{v \in \mathbb{R}^n : \omega_{i1} v_1 + \ldots + \omega_{in} v_n + c_i = 0\}$ for $i = 1, 2, \ldots, k$. Let us assume that θ is chosen generically. Then, for each index i, we have $\{0,1\}^n \cap H_i(\theta) = \emptyset$, and we obtain a slicing of the n-cube with $C_i(\theta) := \big\{ v \in \{0,1\}^n : \sum_{j=1}^n \omega_{ij} v_j + c_i > 0 \big\}$. The generic parameter vector θ lies in a unique open region of linearity of the tropical morphism Φ. More precisely, this region corresponds to the cone of all θ' in $\mathbb{R}^{nk+n+k}$ such that $C_i(\theta) = C_i(\theta')$ for $i = 1, 2, \ldots, k$. By construction, the map $\Phi : \mathbb{R}^{nk+n+k} \to \mathbb{R}^{2^n}$ is linear on this cone. Following the definition of Φ we see that this linear map is left multiplication of the vector θ by a matrix whose rows are indexed by the observed states v and columns indexed by the coordinates of θ. This matrix is precisely the matrix $\mathcal{A}$ above, where $C_i = C_i(\theta)$ for $i = 1, 2, \ldots, k$. The result follows by continuity of the map Φ. □

As an immediate consequence of Lemma 4.1 we obtain the following result:

THEOREM 4.1. *The dimension of the tropical RBM model TM_n^k equals the maximum rank of any matrix of size $2^n \times \big(nk + n + k\big)$ of the form*

$$\mathcal{A} \;=\; \big(A \mid A_{C_1} \mid A_{C_2} \mid \cdots \mid A_{C_k} \big),$$

where $\{C_1, C_2, \ldots, C_k\}$ is any set of k slicings of the n-cube.

Theorem 4.1 furnishes a tool to attack Conjecture 2.1. What remains is the combinatorial problem of finding a suitable collection of slicings of the n-cube. In what follows we shall apply existing results from coding theory to this problem.

There are two quantities from the coding theory literature [**2, 6, 7, 20**] that are of interest to us. The first one is $A_2(n, 3)$, the size (number of codewords) of the largest binary code on n bits with each pair of codewords at least Hamming distance (number of bit flips) 3 apart. The second one is $K_2(n, 1)$, the size of the

smallest *covering code* on n bits. In other words, $K_2(n,1)$ is the least number of codewords such that every string of n bits lies within Hamming distance one of some codeword. We obtain:

COROLLARY 4.1. *The dimension of the tropical RBM model satisfies*

- $\dim TM_n^k = nk+n+k$ *for* $k < A_2(n,3)$,
- $\dim TM_n^k = \min\{nk+n+k, 2^n-1\}$ *for* $k = A_2(n,3)$,
- $\dim TM_n^k = 2^n-1$ *for* $k \geq K_2(n,1)$.

PROOF. For the first statement, let $k \leq A_2(n,3)-1$ and fix a code in n bits of size $k+1$ with minimum distance ≥ 3. For each codeword let C_j denote its Hamming neighborhood, that is, the codeword together with all strings that are at Hamming distance 1. These $k+1$ sets C_j are pairwise disjoint, and each of them corresponds to a slicing of the cube as in Theorem 4.1. The disjointness of the $k+1$ neighborhoods means that $nk+n+k \leq 2^n-1$. Elementary row and column operations can now be used to see that the corresponding $2^n \times (nk+n+k)$ matrix $\mathcal{A} = (A|A_{C_1}|\cdots|A_{C_k})$ has rank $nk+n+k$. This is because, after such operations, $\mathcal{A}$ consists of a block of format $n \times n$ and k blocks of format $(n+1)\times(n+1)$ along the diagonal. The first block has rank n and the remaining k blocks have rank $n+1$ each. The same reasoning is valid for $k = A_2(n,3)$ except that it may now happen that $nk+k+n \geq 2^n$. In this case, the k blocks have total rank $k(n+1)$ and together with the first $n \times n$ block they give a matrix of maximal rank $\min\{nk+n+k, 2^n-1\}$.

For the third statement, we suppose $C_1, \ldots, C_k$ are slicings with subslicings $C_i' \subseteq C_i$ such that the C_i' are disjoint and no $n+1$ of the vertices in a given C_i lie in a hyperplane. Then $\operatorname{rank}(\mathcal{A}) \geq n + \sum_{i=1}^k |C_i'|$ by similar arguments. This is because we may construct the C_i' by pruning neighbors from codewords, and are left with a lower-dimensional Hamming neighborhood which is a slicing. □

The computation of $A_2(n,3)$ and $K_2(n,1)$, both in general and for specific values of n, has been an active area of research since the 1950s. In Table 4.1 we summarize some of the known results for specific values of n. This table is based on [**6, 24**]. For general values of n, the following bounds can be obtained.

PROPOSITION 4.1. *For binary codes with* $n \geq 3$, *the Varshamov bound*

$$A_2(n,3) \geq 2^{n-\lceil \log_2(n+1)\rceil}$$

holds, whereas for covering codes,

$$K_2(n,1) \leq 2^{n-\lfloor \log_2(n+1)\rfloor}.$$

For $n = 2^\ell - 1$ *with* $\ell \geq 3$, *we have the equality* $A_2(n,3) = K_2(n,1) = 2^{2^\ell-\ell-1}$.

PROOF. A proof of the Varshamov bound on $A_2(n,3)$ may be found in [**20**]. The last statement holds because $A_2(n,3) = K_2(n,1)$ for perfect Hamming codes: for every $\ell \geq 3$ there is a perfect $(2^\ell-1, 2^\ell-\ell-1, 3)$ Hamming code (i.e. a perfect Hamming code on $2^\ell-1$ bits, of size $2^\ell-\ell-1$, and with Hamming distance 3). For a proof of this result, see [**7**]. Additionally, we have $K_2(2^m-1,1) = 2^{2^m-m-1}$ for $m \geq 3$; see [**6**].

The simple upper bound on $K_2(n,1)$ can be obtained by using overlapping copies of the next smallest Hamming code. Suppose $n \neq 2^{\ell'}-1$ for any ℓ', i.e. n is strictly between two integers of the form $2^\ell-1$ (*Hamming integer numbers).*

n	$k \leq$	$k \geq$
5	2^2	**7**
6	2^3	**12**
7	2^4	2^4
8	$\mathbf{2^2 \cdot 5}$	2^5
9	$\mathbf{2^3 \cdot 5}$	**62**
10	$\mathbf{2^3 \cdot 9}$	**120**
11	$\mathbf{2^4 \cdot 9}$	**192**
12	2^8	**380**
13	2^9	**736**
14	2^{10}	**1408**
15	2^{11}	2^{11}
16	$\mathbf{2^5 \cdot 85}$	2^{12}
17	$\mathbf{2^6 \cdot 83}$	2^{13}
18	$\mathbf{2^8 \cdot 41}$	2^{14}
19	$\mathbf{2^{12} \cdot 5}$	**31744**
20	$\mathbf{2^{12} \cdot 9}$	**63488**
21	$\mathbf{2^{13} \cdot 9}$	**122880**
22	$\mathbf{2^{14} \cdot 9}$	**245760**
23	$\mathbf{2^{15} \cdot 9}$	**393216**
24	2^{19}	**786432**
25	2^{20}	**1556480**
26	2^{21}	**3112960**
27	2^{22}	**6029312**
28	2^{23}	**12058624**
29	2^{24}	**23068672**
30	2^{25}	**46137344**
31	2^{26}	2^{26}
32	$\mathbf{2^{20} \cdot 85}$	2^{27}
33	$\mathbf{2^{21} \cdot 85}$	2^{28}

n	$k \leq$
35	$\mathbf{2^{23} \cdot 83}$
37	$\mathbf{2^{26} \cdot 41}$
39	$\mathbf{2^{31} \cdot 5}$
47	$\mathbf{2^{38} \cdot 9}$
63	2^{57}
70	$\mathbf{2^{43} \cdot 1657009}$
71	$\mathbf{2^{63} \cdot 3}$
75	$\mathbf{2^{63} \cdot 41}$
79	$\mathbf{2^{70} \cdot 5}$
95	$\mathbf{2^{85} \cdot 9}$
127	2^{120}
141	$\mathbf{2^{113} \cdot 1657009}$
143	$\mathbf{2^{134} \cdot 3}$
151	$\mathbf{2^{138} \cdot 41}$
159	$\mathbf{2^{149} \cdot 5}$
163	$\mathbf{2^{151} \cdot 19}$
191	$\mathbf{2^{180} \cdot 9}$
255	2^{247}
270	$\mathbf{2^{202} \cdot 1021273028302258913}$
283	$\mathbf{2^{254} \cdot 1657009}$
287	$\mathbf{2^{277} \cdot 3}$
300	$\mathbf{2^{220} \cdot 3348824985082075276195}$
303	$\mathbf{2^{289} \cdot 41}$
319	$\mathbf{2^{308} \cdot 5}$
327	$\mathbf{2^{314} \cdot 19}$
383	$\mathbf{2^{371} \cdot 9}$
511	2^{502}
512	$\mathbf{2^{443} \cdot 1021273028302258913}$

Table 4.1: Special cases where Conjecture 2.1 holds, based on [**6, 24**] and Corollary 4.1. Bold entries show improvements made by various researchers on the bounds provided by Corollary 4.2. For example, for $n = 19$, TM_n^k has the expected dimension if $k \leq 2^{12} \cdot 5 = 20480$ and dimension $2^n - 1$ if $k \geq 31744$, while Corollary 4.2 bounds are $2^{14} = 16384$ and $2^{15} = 32768$, respectively. The "$k \leq$" columns list lower bounds on $A_2(n, 3)$ while the "$k \geq$" column lists upper bounds on $K_2(n, 1)$.

Let $\underline{n}$ be the largest Hamming integer smaller than n, with $\ell = \lfloor \log_2(n+1) \rfloor$, so $\underline{n} = 2^\ell - 1$. The number of hidden nodes needed to cover the $\underline{n}$-cube is exactly $K_2(\underline{n}, 1) = 2^{2^\ell - \ell - 1}$. We may use the $\underline{n}$ codes to cover each of the $2^{n-\underline{n}}$ faces of the n-cube with $2^{\underline{n}}$ vertices, although we will have overlaps. That is,

$$K_2(n, 1) \;\leq\; K_2(\underline{n}, 1) \cdot 2^{n-\underline{n}}. \tag{4.1}$$

Taking $\log_2$ in the inequality (4.1), we obtain

$$\log_2 K_2(n, 1) \;\leq\; \log_2(K_2(\underline{n}, 1) 2^{n-\underline{n}}) \;=\; n - \lfloor \log_2(n+1) \rfloor.$$

This implies $K_2(n, 1) \;\leq\; 2^{n - \lfloor \log_2(n+1) \rfloor}$. □

Our method results in the following upper and lower bounds for arbitrary values of n. Note that the bound is tight if $n+1$ is a power of 2. Otherwise there might be a multiplicative gap of up to 2 between the lower and upper bound. In addition to these general bounds, we have the specific results recorded in Table 4.1.

COROLLARY 4.2. *The coding theory argument leads to the following bounds:*

- *If* $k < 2^{n-\lceil \log_2(n+1) \rceil}$, *then* $\dim TM_n^k = nk+n+k$.
- *If* $k = 2^{n-\lceil \log_2(n+1) \rceil}$, *then* $\dim TM_n^k = \min\{nk+n+k, 2^n-1\}$.
- *If* $k \geq 2^{n-\lfloor \log_2(n+1) \rfloor}$, *then* $\dim TM_n^k = 2^n - 1$.

PROOF OF THEOREM 1.1. This is now easily completed by combining Corollary 4.2 with the inequalities in (2.6). □

We close this section with the remark that the use of Hamming codes is a standard tool in the study of dimensions of secant varieties. We learned this technique from Tony Geramita and his collaborators [**4**]. For a review of the relevant literature see Draisma's paper [**12**]. It is important to note that, in spite of the combinatorial similarities, the varieties we study here are different from and more complicated than higher secant varieties of Segre varieties. This may be because the varieties here involve both the secant construction and Hadamard products.

5. Polyhedral geometry of parametric inference

The tropical model TM_n^k is not just a convenient tool for estimating the dimension of the statistical model M_n^k. It is also of interest as the geometric object that organizes the space of inference functions which the model can compute. This statistical interpretation of tropical spaces was introduced in [**28**] and further developed in [**16, 29**]. We shall now discuss this perspective for the RBM model.

Given an RBM model with fixed parameters learned by some estimation procedure and an observed state v, we want to infer which value $\hat{h}$ of the hidden data maximizes $\text{Prob}(h \mid v)$. The inferred string $\hat{h}$ might be used in classification or as the input data for another RBM in a deep architecture. Such a vector of hidden states is called an *explanation* of the observation v. Each choice of parameters $\theta = (b, W, c)$ defines an *inference function* I_θ sending $v \mapsto \hat{h}$. The value $I_\theta(v)$ equals the hidden string $h \in \{0,1\}^k$ that attains the maximum in the tropical polynomial

$$\max_{h\in\{0,1\}^k} \{h^\top W v + c^\top h + b^\top v\} \;=\; b^\top v \;+ \max_{h\in\{0,1\}^k} \{h^\top W v + c^\top h\}. \tag{5.1}$$

In order for the inference function I_θ to be well-defined, it is necessary (and sufficient) that $\theta = (b, W, c)$ lies in an open cone of linearity of the tropical morphism Φ. In that case, the maximum in Equation (5.1) is attained for a unique value of h. That h can be recovered from the expression of Φ as we vary the parameters in the fixed cone of linearity. Thus, the inference functions are in one-to-one correspondence with the regions of linearity of the tropical morphism Φ.

The RBM model grew out of work on artificial neurons modeled as linear threshold functions [**26, 30**]. We pause our geometric discussion to offer remarks about these functions and the types of inference functions that our model can represent.

A *linear threshold function* is a function $\{0,1\}^n \to \{0,1\}$ defined by choosing a weight vector $\omega \in \mathbb{R}^n$ and a target weight $\pi \in \mathbb{R}$. For any point $v \in \{0,1\}^n$ we compute the value ωv, we test if this quantity is at most π or not, and we assign value 0 or 1 to V depending on $\pi \geq \omega v$ or $\pi < \omega v$. The weights ω, π define a

hyperplane in $\mathbb{R}^n$ such that the vertices of the n-cube lie on the "true" or "false" side of the hyperplane. Using the linear threshold functions, we construct a k-valued function $\{0,1\}^n \to \{0,1\}^k$ where we replace the weight vector ω by a $k \times n$ matrix W and the target weight π by a vector $\pi \in \mathbb{R}^k$. More precisely, the function assigns a vertex of the k-cube where the i-th coordinate equals 0 if $(Wv)_i \geq \pi_i$ and 1 if not. Our discussion of slicings of the n-cube in Section 4 implies the following observation:

PROPOSITION 5.1. *The inference functions for the restricted Boltzmann machine model M_n^k are precisely those Boolean functions $\{0,1\}^n \to \{0,1\}^k$ for which each of the k coordinate functions $\{0,1\}^n \to \{0,1\}$ is a linear threshold function.*

Most Boolean functions are not linear threshold functions, that is, are not inference functions for the model M_n^1. For example, the parity function cannot be so represented. To be precise, while the number of all Boolean functions is 2^{2^n}, it is known [**27**] that for $n \geq 8$ the number $\lambda(n)$ of linear threshold functions satisfies

$$2^{\binom{n}{2}+16} < \lambda(n) \leq 2^{n^2}.$$

The exact number $\lambda(n)$ of linear threshold functions has been computed for up to $n = 8$. The *On-Line Encyclopedia of Integer Sequences* [**31**, A000609] reveals

$$\lambda(1\ldots 8) = 4, 14, 104, 1882, 94572, 15028134, 8378070864, 17561539552946. \tag{5.2}$$

Combining k such functions for $k \geq 2$ yields $\lambda(n)^k = 2^{\Theta(kn^2)}$ possible inference functions for the RBM model M_n^k. This number grows exponentially in the number of model parameters. This is consistent with the result of Elizalde and Woods in [**16**] which states that the number of inference functions of a graphical model grows polynomially in the size of the graph when the number of parameters is fixed.

In typical implementations of RBMs using IEEE 754 doubles, the size in bits of the representation is $64(nk+n+k)$. Thus the number $2^{\Theta(kn^2)}$ of inference functions representable by a theoretical RBM M_n^k will eventually outstrip the number $2^{64(nk+n+k)}$ representable in a fixed-precision implementation; for example with $k = 100$ hidden nodes, this happens at $n \geq 132$. As a result, the size of the regions of linearity will shrink to single points in floating point representation. This is one possible contributor to the difficulties that have been encountered in scaling RBMs.

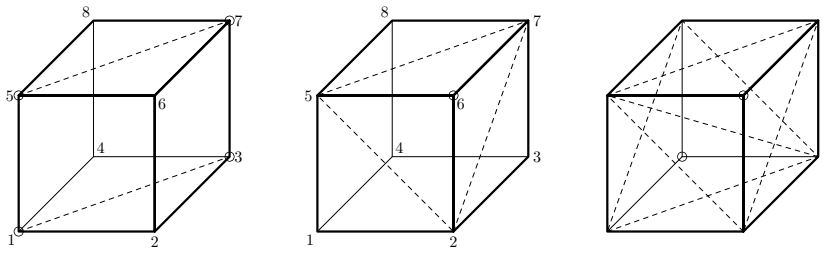
The tropical point of view allows us to organize the geometric information of the space of inference functions into the tropical model TM_n^k, which can then be analyzed with the tools of tropical and polyhedral geometry. We now describe this geometry in the case $k = 1$. Geometrically, we can think of the linear threshold functions as corresponding to the vertices of the $(n+1)$-dimensional zonotope corresponding to the n-cube. This zonotope is the Minkowski sum in $\mathbb{R}^{n+1}$ of the 2^n line segments $[(1,0,\ldots,0),(1,v)]$ where v ranges over the set $\{0,1\}^n$.

The quantity $\lambda(n)$ is the number of vertices of these zonotopes, and their facet numbers were computed by Aichholzer and Aurenhammer [**1**, Table 2]. They are

$$4, 12, 40, 280, 6508, 504868, 142686416, 172493511216, \ldots \tag{5.3}$$

For example, the second entry in (5.2) and (5.3) refers to a 3-dimensional zonotope known as the *rhombic dodecahedron*, which has 12 facets and $\lambda(2) = 14$ vertices. Likewise, the third entry in (5.2) and (5.3) refers to a 4-dimensional zonotope with 40 facets and $\lambda(3) = 104$ vertices. The normal fan of that zonotope is an arrangement of eight hyperplanes, indexed by $\{0,1\}^3$, which partitions $\mathbb{R}^4$ into 104 open

convex polyhedral cones. That partition lifts to a partition of the parameter space $\mathbb{R}^7$ for M_3^1 whose cones are precisely the regions on which the tropical morphism Φ is linear. The image of that morphism is the first non-trivial tropical RBM model TM_3^1. This model has the expected dimension 7 and it happens to be a pure fan.



(a) D_{1357} (b) V_6 (c)

Figure 5.1: Subdivisions of the 3-cube that represent vertices and facets of TM_3^1

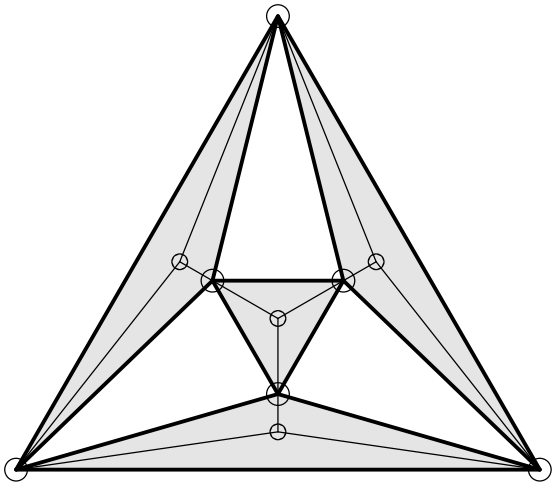


Figure 5.2: The tropical model TM_3^1 is glued from four triangulated bipyramids. In this octahedron graph, each of the bipyramids is represented by a shaded triangle.

.

EXAMPLE 5.1. The tropical RBM model TM_3^1 is a 7-dimensional fan whose lineality space is 3-dimensional. It is a subfan of the secondary fan of the 3-cube [**9**, Corollary 2.2]. The secondary fan of the 3-cube can be represented as a 3-dimensional polyhedral sphere with f-vector $(22, 100, 152, 74)$. The 74 facets of that 3-sphere correspond to triangulations of the 3-cube. The tropical model TM_3^1 consists of all regular subdivisions of the 3-cube with two regions covering all eight vertices. It sits inside the polyhedral 3-sphere as a *simplicial* subcomplex with f-vector $(14, 40, 36, 12)$. Its 12 facets (tetrahedra) correspond to a single triangulation

type of the 3-cube as depicted in Figure 5.1c. The 14 vertices of TM_3^1 come in two families: six vertices D_j corresponding to diagonal cuts, as in Figure 5.1a, and eight vertices V_i representing corner cuts, as in Figure 5.1b. The edges come in three families: four edges V_iV_j corresponding to pairs of corner cuts at antipodal vertices of the cube, twenty-four edges V_iD_j, and twelve edges D_iD_j. Finally, of the four possible triangles, only two types are present: the ones with two vertices of different type. Thus, they are 12 triangles $V_iV_jD_k$ and 24 triangles $V_iD_jD_k$.

Figure 5.2 depicts the simplicial complex TM_3^1 which is pure of dimension 3. The six vertices D_i and the twelve edges D_jD_k form the edge graph of an octahedron. The four nodes interior to the shaded triangles represent pairs of vertices V_i that are joined by an edge. Each of the shaded triangles represents three tetrahedra that are glued together along a common edge V_iV_j. Thus the twelve tetrahedra in TM_3^1 come as four triangulated bypiramids. The four bypiramids are then glued into four of the triangles in the octrahedron graph. Our analysis shows that the simplicial complex TM_3^1 has reduced homology concentrated in degree 1 and it has rank 3. □

The previous example is based on the fact that the image of the tropical map $\Phi : \mathbb{R}^{2n+1} \to \mathbb{R}^{2^n}$ is a subfan of the secondary fan of the n-cube. However, it is important to note that Φ is **not** a morphism of fans with respect to the natural fan structure on the parameter space $\mathbb{R}^{2n+1}$ given by the slicings of the n-cube.

EXAMPLE 5.2. Consider the case $n = 2$. Here M_2^1 equals $\mathbb{R}^4$ with its secondary fan structure coming from the two triangulations of the square. Modulo lineality, this fan is simply the standard fan structure $\{\mathbb{R}_{\leq 0}, \{0\}, \mathbb{R}_{\geq 0}\}$ on the real line. The fan structure on the parameter space $\mathbb{R}^7$ has 14 maximal cones. Modulo lineality, this is the normal fan of the rhombic dodecahedron, i.e. a partition of $\mathbb{R}^3$ into 14 open convex cones by an arrangement of four planes through the origin. Ten of these 14 open cones are mapped onto cones, namely, four are mapped onto $\mathbb{R}_{\leq 0}$, two are mapped onto $\{0\}$, and four onto $\mathbb{R}_{\geq 0}$. The remaining four cones are mapped onto $\mathbb{R}^1$, so Φ does not respect the fan structures relative to these four cones.

The situation is analogous for $n = 3$ but more complicated. The tropical map Φ is injective on precisely eight of the 104 maximal cones in the parameter space. These eight cones are the slicings shown on Figure 5.1a. The map Φ is injective on such a cone, but the cone is divided into three subcones by the secondary fan structure on M_3^1. The resulting $24 = 3 \cdot 8$ maximal cells in the parameter space are mapped in a 2-to-1 fashion onto the 12 tetrahedra in Figure 5.2. It would be worthwhile to study the combinatorics of the graph of Φ for $n \geq 3$. □

Acknowledgments

We thank Jan Draisma, J.M. Landsberg, Honglak Lee, Andrew Ng, Sergey Norin, Lior Pachter, Seth Sullivant, Ilya Sutskever, Jenia Tevelev, and Piotr Zwiernik for helpful discussions. Special thanks go to Anders Jensen for the computations he did for us.

References

[1] O. Aichholzer and F. Aurenhammer, *Classifying hyperplanes in hypercubes*, SIAM Journal on Discrete Mathematics **9** (1996) 225–232.

[2] M.R. Best and A.E. Brouwer, *The triply shortened binary Hamming code is optimal*, Discrete Mathematics **17** (1977) 235–245.

[3] A.E. Brouwer and J. Draisma, *Equivariant Gröbner bases and the Gaussian two-factor model*, `arXiv:0908.1530`.
[4] M.V. Catalisano, A. Geramita, and A. Gimigliano, *Secant varieties of* $\mathbb{P}^1 \times \cdots \times \mathbb{P}^1$ *(n-times) are not defective for* $n \geq 5$, `arXiv:0809.1701`.
[5] T. Bogart, A.N. Jensen, D. Speyer, B. Sturmfels, and R. Thomas, *Computing tropical varieties*, Journal of Symbolic Computation **42** (2007) 54–73.
[6] G. Cohen, I. Honkala, S. Litsyn, and A. Lobstein, *Covering Codes*, North Holland (2005).
[7] T. Cover and J. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley and Sons, Inc. (2006).
[8] M.A. Cueto and J. Yu, *An implicitization challenge for binary factor analysis*, presented at MEGA 2009 (Effective Methods in Algebraic Geometry, Barcelona, June 2009).
[9] M. Develin, *Tropical secant varieties of linear spaces*, Discrete and Computational Geometry **35** (2006) 117–129.
[10] J. Draisma, *Finiteness for the k-factor model and chirality varieties*, Advances in Mathematics **223** (2010) 243–256
[11] M. Develin, F. Santos and B. Sturmfels: *On the tropical rank of a matrix*, in Discrete and Computational Geometry, (eds. J.E. Goodman, J. Pach and E. Welzl), Mathematical Sciences Research Institute Publications **52**, Cambridge University Press (2005) 213–242.
[12] J. Draisma, *A tropical approach to secant dimensions*, Journal of Pure and Applied Algebra **212** (2008) 349–363.
[13] M. Drton and S. Sullivant, *Algebraic statistical models*, Statistica Sinica **17** (2007) 1273–1297.
[14] M. Drton, B. Sturmfels and S. Sullivant, *Algebraic factor analysis: tetrads, pentads and beyond*, Probability Theory and Related Fields **138** (2007) 463–493.
[15] M. Drton, B. Sturmfels and S. Sullivant, *Lectures on Algebraic Statistics*, Oberwolfach Seminars **40**, Birkhäuser, Basel (2009).
[16] S. Elizalde and K. Woods, *Bounds on the number of inference functions of a graphical model*, Statistica Sinica **17** (2007) 1395–1415.
[17] L. Garcia, M. Stillman and B. Sturmfels, *Algebraic geometry of Bayesian networks*, Journal of Symbolic Computation **39** (2005) 331–355.
[18] G.E. Hinton, S. Osindero and Y.-W. Teh, *A fast learning algorithm for deep belief nets*, Neural Computation **18** (2006) 1527–1554.
[19] G.E. Hinton and R.R. Salakhutdinov, *Reducing the dimensionality of data with neural networks*, Science **313** (2006) 504–507.
[20] W.C. Huffman and V. Pless, *Fundamentals of Error Correcting Codes*, Cambridge University Press (2003).
[21] A.N. Jensen, *Gfan, a software system for Gröbner fans*. Available at `http://www.math.tu-berlin.de/~jensen/software/gfan/gfan.html`
[22] J.M. Landsberg and L. Manivel, *On the ideals of secant varieties of Segre varieties*, Foundations of Computational Mathematics **4** (2004) 397–422.
[23] N. Le Roux and Y. Bengio, *Representational power of restricted Boltzmann machines and deep belief networks*, Neural Computation **20** (2008) 1631–1649.
[24] S. Litsyn, E.M. Rains, and N.J.A. Sloane, Table of Nonlinear Binary Codes at `http://www.eng.tau.ac.il/~litsyn/tableand/index.html`. Last updated November 24, 1999.
[25] H. Markwig and J. Yu, *The space of tropically collinear points is shellable*, Collectanea Mathematica, **60**(1) (2009) 63–77.
[26] M. Minsky and S. Papert, *Perceptrons, An Introduction to Computational Geometry*. MIT Press, Cambridge, MA. (1969).
[27] P.C. Ojha, *Enumeration of linear threshold functions from the lattice of hyperplane intersections,* IEEE Trans. Neural Networks, **11**(4) (2000) 839–850.
[28] L. Pachter and B. Sturmfels, *Tropical geometry of statistical models*, PNAS **101** (2004) 16132–16137.
[29] L. Pachter and B. Sturmfels (editors), *Algebraic Statistics for Computational Biology*, Cambridge University Press (2005).
[30] F. Rosenblatt, *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*, Spartan Books (1962).
[31] N.J.A. Sloane, *The On-Line Encyclopedia of Integer Sequences* (2008), `www.research.att.com/~njas/sequences/`
[32] D. Speyer and B. Sturmfels, *Tropical mathematics*, Mathem. Magazine **82** (2009) 163–173.

[33] R. Wunderling, *Paralleler und Objektorientierter Simplex-Algorithmus*, ZIB Technical Report TR 96-09, Berlin (1996).
[34] P. Zwiernik and J.Q. Smith, *The geometry of conditional independence tree models with hidden variables*, `arXiv:0904.1980.`

Department of Mathematics, University of California, Berkeley, CA 94720, USA
E-mail address: `macueto@math.berkeley.edu`

Departments of Mathematics and Statistics, The Pennsylvania State University, University Park, PA 16802, USA
E-mail address: `morton@math.psu.edu`

Department of Mathematics, University of California, Berkeley, CA 94720, USA
E-mail address: `bernd@math.berkeley.edu`