GRÖBNER BASES FOR STAGED TREES

LAMPRINI ANANIADI AND ELIANA DUARTE

We consider the problem of finding generators of the toric ideal associated to a combinatorial object called a staged tree. Our motivation to consider this problem originates from the use of staged trees to represent discrete statistical models such as conditional independence models and discrete Bayesian networks. The main theorem in this article states that toric ideals of staged trees that are balanced and stratified are generated by a quadratic Gröbner basis whose initial ideal is square-free. We apply this theorem to construct Gröbner bases of a subclass of discrete statistical models represented by staged trees. The proof of the main result is based on Sullivant’s toric fiber product construction (J. Algebra 316:2 (2007), 560–577).

1. Introduction

The study of toric ideals associated to statistical models was pioneered by the work of Diaconis and Sturmfels [4], who first used the generators of a toric ideal to formulate a sampling algorithm for discrete distributions. Since then, and with the subsequent work of [2; 17; 8], the study of toric ideals of discrete statistical models has been an active area of research in algebraic statistics. The books by Sullivant [18, Chapter 9] and Aoki, Hara and Takemura [1] are good references to learn about the role of toric ideals in statistics. A recent introduction to the topic, from the point of view of binomial ideals, can be found in [13, Chapter 9], which also contains a thorough list of references of previous contributions to this topic.

In 2008, Smith and Anderson [15] introduced a new graphical discrete statistical model called a staged tree model. This model is represented by an event tree together with an equivalence relation on its vertices. Staged tree models are useful to represent conditional independence relations among events and random variables, such as those coming from discrete graphical models. Hence, any discrete Bayesian network or decomposable model is also a staged tree model [15]. There are two properties that make staged tree models more general than Bayesian networks or decomposable models. First, the state space of a staged tree model does not have to be a cartesian product. Second, using staged tree models it is possible to represent extra context-specific conditional independence between events. The book of Collazo, Görgen and Smith [3] is a good reference to learn about these models.

In this article we define the toric ideal associated to a staged tree and study its properties from an algebraic and combinatorial point of view. We present Theorem 2.14, which states that toric ideals of staged trees that are balanced and stratified have quadratic Gröbner basis with square-free initial ideal. We apply Theorem 2.14 in Section 5 to obtain Gröbner bases for toric ideals of staged tree models. Our


Keywords: graphical model, toric ideals, staged tree, Markov bases, toric fiber product.
results provide a new point of view on the construction of Gröbner bases for decomposable graphical models, some conditional independence models, as well as the construction of Gröbner bases for staged tree models whose underlying tree is asymmetric.

This article is organized as follows. In Section 2, we define the toric ideal associated to a staged tree. In Section 3, we formulate a toric fiber product construction for balanced and stratified staged trees. In Section 4, we prove our main result Theorem 2.14. Finally, in Section 5 we apply our results to compute Gröbner bases for several statistical models.

2. Staged trees

We start by defining our two objects of interest: a staged tree and its associated toric ideal. First, we set up the graph-theoretic notation and conventions. Let $T = (V, E)$ be a directed rooted tree, with vertex set $V$ and set $E$ of directed edges. We only consider trees $T = (V, E)$ where all elements in $E$ are oriented away from the root. Since we only consider directed paths, we refer to any directed path in $T$ simply as a path. For $v, w \in V$ the directed edge in $E$ from $v$ to $w$ is denoted by $(v, w)$, the set of children of $v$ is $\text{ch}(v) = \{u \mid (v, u) \in E\}$, and the set of outgoing edges from $v$ is $E(v) = \{(v, u) \mid u \in \text{ch}(v)\}$. A vertex $v \in V$ is a leaf if $\text{ch}(v) = \emptyset$ and it is an nonroot vertex if it is different from the root.

**Definition 2.1.** Let $T = (V, E)$ be a tree, $L$ a finite set of labels, and $\theta : E \rightarrow L$ a surjective function. For each $v \in V$, $\theta_v := \{\theta(e) \mid e \in E(v)\}$ is the set of labels associated to $v$. The pair $(T, \theta)$ is a staged tree if

(i) for each $v \in V$, we have $|\theta_v| = |E(v)|$, and

(ii) for any two vertices $v, w \in V$ either $\theta_v = \theta_w$ or $\theta_v \cap \theta_w = \emptyset$.

Condition (ii) in Definition 2.1 defines an equivalence relation on the set of nonleaf vertices of $T$. Namely, two nonleaf vertices $v, w \in V$ are equivalent if and only if $\theta_v = \theta_w$. We refer to the partition induced by this equivalence relation on the set of nonleaf vertices as the set of stages of $T$ and to a single element in this partition as a stage. Condition (i) in Definition 2.1 guarantees that all edge labels associated to a single vertex are distinct. We use $(T, \theta)$ to denote a staged tree with labeling rule $\theta$. For simplicity we will often drop the use of $\theta$ and write $T$ for a staged tree.

To define the toric ideal associated to $(T, \theta)$ we define two polynomial rings. The first ring is $\mathbb{R}[p]_T := \mathbb{R}[p_\lambda \mid \lambda \in \Lambda]$, where $\Lambda$ is the set of root-to-leaf paths in $T$. The second ring is $\mathbb{R}[\Theta]_T := \mathbb{R}[z, L]$, where the labels in $L$ are indeterminates together with a homogenizing variable $z$. For a path $\gamma$ in $T$, $E(\gamma)$ is the set of edges in $\gamma$.

**Definition 2.2.** The toric staged tree ideal associated to $(T, \theta)$ is the kernel of the ring homomorphism $\varphi_T : \mathbb{R}[p]_T \rightarrow \mathbb{R}[\Theta]_T$ defined as

\[ p_\lambda \mapsto z \cdot \prod_{e \in E(\lambda)} \theta(e). \quad (1) \]

The ideal $\ker(\varphi_T)$ defines the toric variety specified as the closure of the image of the monomial parametrization $\Phi_T : (\mathbb{C}^*)^{|L|} \rightarrow \mathbb{P}^{|\Lambda|-1}$ given by $(\theta(e) \in L) \mapsto (z \cdot \prod_{e \in E(\lambda)} \theta(e))_{\lambda \in \Lambda}$. We use the homogenizing variable $z$ in Definition 2.2 to consider the projective toric variety in $\mathbb{P}^{|\Lambda|-1}$. 
It is often useful to encode a monomial map between polynomial rings by using an exponent matrix. Let $B = (b_{ij})$ be a $d \times n$ matrix with nonnegative integer entries. The columns of $B$ define a monomial map $\phi_B : \mathbb{R}[z_1, \ldots, z_n] \to \mathbb{R}[t_1, \ldots, t_d]$, $z_j \mapsto \prod_{i=1}^d t_i^{b_{ij}}$. The matrix $B$ is the exponent matrix of $\phi_B$. We will use this notation in Sections 3 and 4.

**Example 2.3.** The staged tree $T_1$ in Figure 1 has label set $L = \{s_0, \ldots, s_{13}\}$. Each vertex in $T_1$ is denoted by a string of 0s and 1s, and each edge has a label associated to it. The root-to-leaf paths in $T_1$ are denoted by $p_{ijkl}$, where $i, j, k, l \in \{0, 1\}$. A vertex in $T_1$ represented with a blank circle indicates a stage consisting of a single vertex. We use colors in the vertices of $T_1$ to indicate which vertices are in the same stage. For instance, the set of purple vertices $\{000, 010, 100, 110\}$ are in the same stage and therefore they have the same set $\{s_{10}, s_{11}\}$ of associated edge labels. The map $\phi_{T_1}$ sends $(s_0, \ldots, s_{13})$ to

$$(s_0 s_2 s_6 s_{10}, s_0 s_2 s_6 s_{11}, s_0 s_2 s_7 s_{12}, s_0 s_2 s_7 s_{13}, s_0 s_3 s_8 s_{10}, s_0 s_3 s_8 s_{11}, s_0 s_3 s_9 s_{12}, s_0 s_3 s_9 s_{13}, s_1 s_4 s_6 s_{10}, s_1 s_4 s_6 s_{11}, s_1 s_4 s_7 s_{12}, s_1 s_4 s_7 s_{13}, s_1 s_5 s_8 s_{10}, s_1 s_5 s_8 s_{11}, s_1 s_5 s_9 s_{12}, s_1 s_5 s_9 s_{13}).$$

The toric ideal $\ker(\phi_{T_1})$ is generated by a quadratic Gröbner basis with square-free initial ideal.

**Example 2.4.** Consider the two staged trees $T_2, T_3$ depicted in Figure 1. For the staged tree $T_2$, $\ker(\phi_{T_2})$ is generated by a quadratic Gröbner basis with square-free initial ideal. For $T_3$, the ideal $\ker(\phi_{T_1})$ also has a Gröbner basis with square-free initial ideal but its elements are of degree 2 and degree 4.

We are interested in relating the combinatorial properties of the staged tree $(T, \theta)$ with the properties of the toric ideal $\ker(\phi_T)$. The two definitions that are relevant for the statement of the main theorem
are the definition of balanced staged tree and of stratified staged tree. Before stating the main theorem, Theorem 2.14, we look into the definition and consequences of these two notions.

**Definition 2.5.** Let $T$ be a tree. For $v \in V$, the level of $v$ is denoted by $\ell(v)$ and it is equal to the number of edges in the unique path from the root of $T$ to $v$. If all the leaves in $T$ have the same level then the level of $T$ is equal to the level of any of its leaves. The staged tree $(T, \theta)$ is stratified if all its leaves have the same level and if every two vertices in the same stage have the same level.

It is easy to check that all trees in Figure 1 are stratified. Namely, we only need to verify that every two vertices with the same color are also in the same level. Notice that the combinatorial condition of $(T, \theta)$ being stratified implies the algebraic condition that the map $\phi_T$ is square-free.

We now turn our attention to the definition of a balanced staged tree. This definition is formulated in terms of polynomials associated to each vertex of the tree. We proceed to explain their notation and basic properties.

**Definition 2.6.** Let $(T, \theta)$ be a staged tree, $v \in V$, and $T_v$ the subtree of $T$ rooted at $v$. The tree $T_v$ is a staged tree with the induced labeling from $T$. Let $\Lambda_v$ be the set of $v$-to-leaf paths in $T$. The interpolating polynomial of $T_v$ is

$$t(v) := \sum_{\lambda \in \Lambda_v} \prod_{e \in E(\lambda)} \theta(e).$$

When $v$ is the root of $T$, the polynomial $t(v)$ is called the interpolating polynomial of $T$. Two staged trees $(T, \theta)$ and $(T, \theta')$ with the same label set $L$ are polynomially equivalent if their interpolating polynomials are equal.

The interpolating polynomial of a staged tree is useful to capture symmetries among subtrees. It is also an important tool in the study of the statistical properties of staged tree models. This polynomial was defined by Görgen and Smith in [10] and further studied by Görgen et al. in [11]. Although these two articles are written for a statistical audience, we would like to emphasize that their symbolic algebra approach to the study of statistical models proves to be very important for the use of these models in practice. We will define the statistical model associated to a staged tree and connect to other results in algebraic statistics in Section 5.

If $(T, \theta)$ is a staged tree, the polynomials $t(\cdot)$ satisfy a recursive relation. This relation is useful to prove statements about the algebraic and combinatorial properties of $T$. We state this property as a lemma.

**Lemma 2.7** [11, Theorem 1]. Let $(T, \theta)$ be a staged tree, $v \in V$ and $\text{ch}(v) = \{v_0, \ldots, v_k\}$. Then the polynomial $t(v)$ admits the recursive representation $t(v) = \sum_{i=0}^{k} \theta(v, v_i) t(v_i)$.

**Example 2.8.** Consider the staged tree $T_1$ in Figure 1. If $v$ and $w$ are orange and blue vertices in $T_1$ respectively and $r$ is the root of $T_1$ then

$$t(v) = s_6(s_{10} + s_{11}) + s_7(s_{12} + s_{13}),$$
$$t(w) = s_8(s_{10} + s_{11}) + s_9(s_{12} + s_{13}),$$
$$t(r) = (s_0s_2 + s_1s_4)t(v) + (s_0s_3 + s_1s_5)t(w).$$
Definition 2.9. Let \((T, \theta)\) be a staged tree and let \(v, w\) be two vertices in the same stage, with \(\text{ch}(v) = \{v_0, \ldots, v_k\}\) and \(\text{ch}(w) = \{w_0, \ldots, w_k\}\). After a possible reindexing of the elements in \(\text{ch}(w)\), we may assume that \(\theta(v, v_i) = \theta(w, w_i)\) for all \(i \in \{0, \ldots, k\}\). The pair \(v, w\) is balanced if
\[
t(v_i)t(w_j) = t(w_i)t(v_j) \quad \text{in} \; \mathbb{R}[[\Theta]]_T \quad \text{for all} \; i \neq j \in \{0, \ldots, k\}.
\]

The staged tree \((T, \theta)\) is **balanced** if every pair of vertices in the same stage is balanced.

Example 2.10. The two staged trees \(T_2, T_3\) in Figure 1 are not balanced. The pair of pink vertices in \(T_2\) is not balanced because \((s_{10} + s_{11})(s_{12} + s_{13}) \neq (s_{10} + s_{11})^2\). By a similar argument we can check that \(T_3\) is also not balanced.

Although the balanced condition in Definition 2.9 seems to be algebraic and hard to check, in many cases it is very combinatorial. To formulate a precise instance where this is true we need the following definition.

Definition 2.11. Let \((T = (V, E), \theta)\) be a staged tree. We say that two vertices \(v, w \in V\) are in the same **position** if they are in the same stage and \(t(v) = t(w)\).

The notion of position for vertices in the same stage was formulated in [15]. Intuitively it means that if we regard the subtrees \(T_v\) and \(T_w\) as representing the unfolding of a sequence of events, then the future of \(v\) and \(w\) is essentially the same. In the next lemma we use positions of vertices to provide a sufficient condition on a stratified staged tree \((T, \theta)\) to be balanced.

Lemma 2.12. Let \((T, \theta)\) be a stratified staged tree. Suppose that every two vertices in \(T\) that are in the same stage are also in the same position. Then \((T, \theta)\) is balanced.

Proof. Following Definition 2.9, it suffices to prove that any pair of vertices in the same position is balanced. Let \(v, w\) be two vertices in the same position. We use the same notation in Definition 2.9 and assume without loss of generality that \(\theta(v, v_i) = \theta(w, w_i)\). Using Lemma 2.7 we write
\[
t(v) = t(w) \iff \sum_{i=0}^k \theta(v, v_i)t(v_i) = \sum_{i=0}^k \theta(w, w_i)t(w_i) \iff \sum_{i=0}^k \theta(v, v_i)(t(v_i) - t(w_i)) = 0.
\]

Since \((T, \theta)\) is stratified, the variables appearing in the polynomials \(t(v_i), t(w_i)\) are disjoint from the set of variables \([\theta(v, v_0), \ldots, \theta(v, v_k)]\). Thus \(t(v_i) = t(w_i)\) for all \(i \in \{0, \ldots, k\}\). It follows that for all \(i, j \in \{0, \ldots, k\}\) the equality \(t(v_i)t(w_j) = t(w_i)t(v_j)\) is true. Hence \((T, \theta)\) is balanced. \(\square\)

Example 2.13. The staged tree \(T_1\) in Figure 1 is balanced. This can be readily checked by noting that the blue vertices are in the same position and that the same is true for the orange vertices. The two trees in Figure 3 are examples of balanced staged trees in which the blue vertices are not in the same position.

We are now ready to state the main theorem.

Theorem 2.14. If \((T, \theta)\) is a balanced and stratified staged tree then \(\ker(\varphi_T)\) is generated by a quadratic \(\text{Gröbner basis with square-free initial ideal.}\)
We clarify that the conditions of \((\mathcal{T}, \theta)\) being balanced and stratified in Theorem 2.14 are sufficient for \(\ker(\varphi_\mathcal{T})\) to have a quadratic Gröbner basis but are not necessary. In the examples of staged trees in Figure 1, all of the trees \(\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3\) are stratified but only \(\mathcal{T}_1\) is balanced. Even though \(\mathcal{T}_2\) is not balanced, it has a quadratic Gröbner basis with square-free initial terms.

3. Toric fiber products

In this section we review toric fiber products following the exposition in [17]. We then use these results in Section 4 to prove Theorem 2.14.

Given a positive integer \(m\), set \([m] = \{1, 2, \ldots, m\}\). Let \(r\) be a positive integer, and let \(s\) and \(t\) be two vectors of positive integers in \(\mathbb{Z}^r_{>0}\). Consider the multigraded polynomial rings
\[
\mathbb{K}[x] := \mathbb{K}[x^i_j | i \in [r], j \in [s_i]] \quad \text{and} \quad \mathbb{K}[y] := \mathbb{K}[y^i_k | i \in [r], k \in [t_i]]
\]
graded by the same set \(A = \{a^1, \ldots, a^r\} \subset \mathbb{Z}^d\), where
\[
\deg(x^i_j) = \deg(y^i_k) = a^i
\]
and such that there exists a vector \(w \in \mathbb{Q}^d\) such that \(\langle w, a^i \rangle = 1\) for any \(a^i \in A\). A polynomial in \(\mathbb{K}[x]\) or \(\mathbb{K}[y]\) is \(A\)-homogeneous whenever it is homogeneous with respect to the multigrading given by \(A\). An ideal in \(\mathbb{K}[x]\) or \(\mathbb{K}[y]\) is \(A\)-homogeneous if it is generated by \(A\)-homogeneous elements. If \(I \subseteq \mathbb{K}[x]\) and \(J \subseteq \mathbb{K}[y]\) are \(A\)-homogeneous ideals, then the quotient rings \(R = \mathbb{K}[x]/I\) and \(S = \mathbb{K}[y]/J\) are also multigraded rings. Let
\[
\mathbb{K}[z] := \mathbb{K}[z^i_{jk} | i \in [r], j \in [s_i], k \in [t_i]]
\]
and consider the ring homomorphism
\[
\phi_{I,J} : \mathbb{K}[z] \rightarrow R \otimes_{\mathbb{K}} S,
\]
\[
z^i_{jk} \mapsto \bar{x}^i_j \otimes \bar{y}^i_k,
\]
where \(\bar{x}^i_j\) and \(\bar{y}^i_k\) are the equivalence classes of \(x^i_j\) and \(y^i_k\) respectively.

**Definition 3.1.** The toric fiber product of \(I\) and \(J\) is \(I \times_A J := \ker(\phi_{I,J})\).

We now recall results from [17] about the generators of the ideal \(I \times_A J\). The generators of \(I \times_A J\) come in two flavors; new quadratic generators that are created by the toric fiber product construction and new generators that are lifts of generators from \(I\) and \(J\).

Consider the monomial parametrization
\[
\phi_B : \mathbb{K}[z^i_{jk} | i \in [r], j \in [s_i], k \in [t_i]] \rightarrow \mathbb{K}[x^i_j, y^i_k | i \in [r], j \in [s_i], k \in [t_i]],
\]
\[
z^i_{jk} \mapsto x^i_jy^i_k,
\]
where \(B\) is the exponent matrix of \(\phi_B\). Let
\[
\text{Quad}_B := \langle z^i_{jk}z^j_{jk} - z^i_{jk}z^i_{jk} | 1 \leq i \leq r, 1 \leq j_1 < j_2 \leq s_i, 1 \leq k_1 < k_2 \leq t_i \rangle.
\]
By [17, Proposition 10], the elements in \( \text{Quad}_B \) are a Gröbner basis of the ideal \( I_B := \ker(\phi_B) \) with respect to any term order that selects the underlined terms as leading terms. The elements in \( \text{Quad}_B \) are new quadratic generators created by the toric fiber product construction of \( I \) and \( J \).

The construction of the generators of \( I \times_A J \) that are lifts to the ring \( \mathbb{K}[z] \) of elements in \( I \) and \( J \) is explained in full generality in [17]. Since we will only consider lifts of pure quadratic binomials, we restrict the definition from [17] to this case. We define lifts of \( A \)-homogeneous elements in \( \mathbb{K}[x] \), an analogous construction works to define lifts of elements in \( \mathbb{K}[y] \).

Consider the \( A \)-homogeneous polynomial

\[
f = x^{i_1}_{a_1}x^{i_2}_{a_2} - x^{i_3}_{a_3}x^{i_2}_{a_4} \in \mathbb{K}[x],
\]

where \( i_1, i_2 \in [r], \, a_1, a_3 \in [s_{i_1}], \, a_2, a_4 \in [s_{i_2}] \) and \( f \in I \). Set \( k = (k_1, k_2) \) with \( k_1 \in [t_{i_1}], \, k_2 \in [t_{i_2}] \) and consider \( f_k \in \mathbb{K}[z] \) defined by

\[
f_k = x^{i_1}_{a_1k_1}x^{i_2}_{a_2k_2} - x^{i_3}_{a_3k_1}x^{i_2}_{a_4k_2},
\]

The new \( A \)-homogeneous polynomial \( f_k \) is in \( I \times_A J \) for all \( k \) because \( f \in I \).

**Definition 3.2.** Let \( A \) be linearly independent and let \( F \subset I \) be a collection of pure and quadratic \( A \)-homogeneous polynomials. We associate to each \( f \in F \) the set \( T_f = [i_{r_1}] \times [i_{r_2}] \) of indices and define

\[
\text{Lift}(F) = \{ f_k | f \in F, \, k \in T_f \}.
\]

The set \( \text{Lift}(F) \) is called the lifting of \( F \) to \( I \times_A J \). For a collection \( H \) of \( A \)-homogeneous elements of \( J \), we define \( \text{Lift}(H) \) in a similar way.

We are now ready to state the result from [17] that we will use in the proof of Theorem 2.14. The important part of this theorem is that we can construct Gröbner basis of the toric fiber product \( I \times_A J \) by using lifts of Gröbner bases for \( I \) and \( J \) together with the elements in \( \text{Quad}_B \).

**Theorem 3.3** [17, Theorem 12]. Suppose that \( A \) is linearly independent. Let \( F \subset I \) be a homogeneous Gröbner basis for \( I \) with respect to the weight vector \( \omega_1 \) and let \( H \subset J \) be a homogeneous Gröbner basis for \( J \) with respect to the weight vector \( \omega_2 \). Let \( \omega \) be a weight vector such that \( \text{Quad}_B \) is a Gröbner basis for \( I_B \). Then the set \( \text{Lift}(F) \cup \text{Lift}(H) \cup \text{Quad}_B \) is a Gröbner basis for \( I \times_A J \) with respect to the weight order \( \phi_B^*(\omega_1, \omega_2) + \varepsilon \omega \) for sufficiently small \( \varepsilon > 0 \).

4. **Proof of main theorem**

In the first part of this section we explain how to construct staged trees from smaller pieces and relate this construction to toric fiber products in Proposition 4.4. Then, in Proposition 4.11 we use this construction repeatedly for balanced and stratified staged trees. Finally, we use these results together with Theorem 3.3 to prove our main theorem.

Let \((T, \theta)\) be a staged tree. We recursively define an indexing on the set of nonroot vertices of \( T \). The children of the root are indexed by \( [0, 1, \ldots, k] \). If \( a \) is the index of a vertex \( v \) and \( |E(v)| = j + 1 \),
then we index the children of $a$ by $a0, \ldots, aj$. This way each nonroot vertex in $V$ is indexed by a finite sequence of nonnegative integers $a = a1a2 \cdots a\ell$,

where $\ell$ is the level of the vertex indexed by $a$. From this point on we refer to any nonroot vertex in $V$ via its index $a$. All vertices of the trees in Figure 1 are indexed following this rule. In Figure 1 the index of each vertex is displayed immediately above each vertex and on the side for the leaves. We denote by $i_T$ the set of indices of the leaves in $T$.

**Definition 4.1.** If a staged tree has level one we call it a *level-one tree*. We reserve for it the special notation $(B, \epsilon)$, where $B = (V, E)$ is the tree and $\epsilon$ its labeling rule. By condition (i) in Definition 2.1, the size of the label set of $B$ is equal to $|E|$. Let $E = \{e0, \ldots, em\}$. We use $e_k$ to denote the image of the $k$-th element in $E$ under $\epsilon$. We also use the notation $(B, \{e0, \ldots, em\})$ when we wish to emphasize the label set of the level-one tree.

**Definition 4.2.** Let $(T, \theta)$ be a staged tree and $G = \{G1, \ldots, G_r\}$ be a partition of the set of leaves $i_T$. We consider a collection $\{(Bi, \epsilon^{(i)}) \mid i \in [r]\}$ of level-one trees such that their label sets are pairwise disjoint and disjoint from the label set of $(T, \theta)$. The *gluing component* $T_G$ associated to $T$ and $G$ is

$$T_G := \bigsqcup_{i \in [r]} (Bi, \epsilon^{(i)}).$$

The gluing component $T_G$ is a forest of level-one trees; its label set is the union of the label sets of each $(Bi, \epsilon^{(i)})$. We denote by $[T, T_G]$ the tree obtained by gluing $Bi$ to the leaf $a$ for all $a \in Gi$ and all $i \in [r]$.

**Remark 4.3.** The tree $[T, T_G]$ is a staged tree. Its label set is the union of the labels of each $(Bi, \epsilon^{(i)})$. We denote by $[T, T_G]$ the tree obtained by gluing $Bi$ to the leaf $a$ for all $a \in Gi$ and all $i \in [r]$.

We relate $\ker(\varphi_{(T, T_G)})$ to the toric fiber product of the two ideals $\ker(\varphi_T)$ and the zero ideal $(0)$. Fix the notation for $T, G, [T, T_G]$ as in Definition 4.2. We associate to $T_G$ the rings

$$\mathbb{R}[p]_{T_G} := \mathbb{R}[p^i_j \mid j \in i_B, i \in [r]] \quad \text{and} \quad \mathbb{R}[\Theta]_{T_G} := \mathbb{R}[\epsilon^{(i)}_k \mid i \in [r], k \in i_B],$$

and the ring map

$$\varphi_{T_G} : \mathbb{R}[p]_{T_G} \to \mathbb{R}[\Theta]_{T_G}, \quad p^i_j \mapsto \epsilon^{(i)}_k.$$

Since there is a one-to-one correspondence between the variables $p^i_j$ and $\epsilon^{(i)}_k$, we see that $\varphi_{T_G}$ is an isomorphism. In particular, $\ker(\varphi_{T_G}) = (0)$. Now, using $G$ we regroup the variables in $\mathbb{R}[p]_T$ by

$$\mathbb{R}[p]_T = \mathbb{R}[p^i_j \mid j \in Gi, i \in [r]].$$
We define multigradings on the polynomial rings $\mathbb{R}[p]_T$ and $\mathbb{R}[p]_{T_G}$ by
\[
\deg(p^j_i) = \deg(p^j_k) = e_i \quad \text{for} \quad j \in G_i, \ k \in i_B, \ i \in [r].
\]
Here $e_i$ is the $i$-th standard unit vector in $\mathbb{Z}^r$. If $A$ is the set of all these multidegrees, then $T$ is linearly independent as it is the collection of standard unit vectors in $\mathbb{Z}^r$.

Suppose $\ker(\varphi_T)$ is $A$-homogeneous and fix $R = \mathbb{R}[p]_T / \ker(\varphi_T)$, $S = \mathbb{R}[p]_{T_G} / \ker(\varphi_{T_G})$. Let $\mathbb{R}[p]_{[T,T_G]} = \mathbb{R}[p^j_k \mid j \in G_i, \ k \in i_B, \ i \in [r]]$ and consider the ring homomorphism
\[
\psi : \mathbb{R}[p]_{[T,T_G]} \rightarrow R \otimes_{\mathbb{R}} S,
\]
\[
p^j_k \mapsto \tilde{p}^j_k \otimes \tilde{p}^j_k \quad \text{for} \quad j \in G_i, \ k \in i_B, \ i \in [r].
\]

The ideal $\ker(\psi) = \ker(\varphi_T) \times_A \langle 0 \rangle$ is the toric fiber product of $\ker(\varphi_T)$ and $\langle 0 \rangle$.

**Proposition 4.4.** Let $T$, $G$, and $T_G$ be as in Definition 4.2. Suppose that $\ker(\varphi_T)$ is $A$-homogeneous. Then
\[
\ker(\varphi_{[T,T_G]}) = \ker(\varphi_T) \times_A \langle 0 \rangle.
\]

**Proof.** Consider the tensor product of maps $\tilde{\varphi}_T \otimes \tilde{\varphi}_{T_G} : R \otimes_{\mathbb{R}} S \rightarrow \mathbb{R}[\Theta]_T \otimes_{\mathbb{R}} \mathbb{R}[\Theta]_{T_G}$, where $\tilde{\varphi}_T : R \rightarrow \mathbb{R}[\Theta]_T$ and $\tilde{\varphi}_{T_G} : S \rightarrow \mathbb{R}[\Theta]_{T_G}$ are induced by $\varphi_T$ and $\varphi_{T_G}$ on the quotient rings $R$ and $S$ respectively. Note that there is a canonical isomorphism $\mathbb{R}[\Theta]_T \otimes_{\mathbb{R}} \mathbb{R}[\Theta]_{T_G} \cong \mathbb{R}[\Theta]_{[T,T_G]}$. Under this isomorphism,
\[
\tilde{\varphi}_T \otimes \tilde{\varphi}_{T_G}(p^j_k \otimes \tilde{p}^j_k) = \varphi_T(p^j_k) \cdot \varphi_{T_G}(p^j_k) = \left( z \cdot \prod_{e \in E(\lambda_j)} \theta(e) \right) e^{(i)}_k = \varphi_{[T,T_G]}(p^j_k).
\]

The last equality follows by the construction of $[T, T_G]$. Hence $\varphi_{[T,T_G]} = (\tilde{\varphi}_T \otimes \tilde{\varphi}_{T_G}) \circ \psi$. Since $\tilde{\varphi}_T \otimes \tilde{\varphi}_{T_G}$ is injective, $\ker(\varphi_{[T,T_G]}) = \ker(\psi)$. By (3), and since $\ker(\varphi_T)$ is $A$-homogeneous, we conclude $\ker(\psi) = \ker(\varphi_T) \times_A \langle 0 \rangle$. Thus $\ker(\varphi_{[T,T_G]}) = \ker(\varphi_T) \times_A \langle 0 \rangle$. \qed

We make several remarks on the scope of Proposition 4.4 via the next set of examples.

**Definition 4.5.** Let $(T, \theta)$ be a stratified staged tree of level $m$. For $1 \leq q \leq m$ we define $V_{\leq q} := \bigcup_{i=0}^q V_i$, where $V_q := \{ v \in V : \ell(v) = q \}$. The tree $T^{(q)} = (V_{\leq q}, E_{\leq q})$ is the induced subtree of $T$ on the vertex set $V_{\leq q}$. The restriction $\theta|_{E_{\leq q}}$ defines a labeling on $T^{(q)}$. The staged tree $(T^{(q)}, \theta|_{E_{\leq q}})$ is the level-$q$ subtree of $(T, \theta)$.

**Example 4.6.** Consider the staged tree $T_1$ in Figure 1 and let $T = T_1^{(3)}$ be the level-three subtree of $T$. The label set of $T$ is $\{s_0, \ldots, s_9\}$. Fix
\[
G = \{(000, 010, 100, 110), \ (001, 011, 101, 111)\},
\]
\[
T_G = (B_1, \{s_{10}, s_{11}\}) \cup (B_2, \{s_{12}, s_{13}\}).
\]

With this choice of $T$, $G$, and $T_G$ we see that $T_1 = [T, T_G]$. Now $\mathbb{R}[p]_T = \mathbb{R}[p^i_a \mid a \in G_i, \ i \in [1, 2]]$; hence $\deg(p^0_{000}, p^1_{010}, p^1_{100}, p^1_{110}) = e_1$ and $\deg(p^2_{001}, p^2_{011}, p^2_{101}, p^2_{111}) = e_2$ so $A = \{e_1, e_2\} \subset \mathbb{Z}^2$ is of full rank. The ideal
\[
\ker(\varphi_T) = \langle p^1_{000}, p^2_{010} - p^1_{100}, p^2_{011}, p^1_{010} p^2_{111} - p^1_{110} p^2_{011} \rangle
\]
is $A$-homogeneous. Hence by Proposition 4.4 $\ker(\varphi_{T_1}) = \ker(\varphi_T) \times_A \langle 0 \rangle$. 

---

**Gröbner Bases for Staged Trees**

9

---
Example 4.7. Let $T_2$ be the staged tree from Figure 1. We proceed in a similar fashion as in Example 4.6. Set $T = T_2^{(2)}$.

$G = \{(00, 01, 10), (11, 31), (20, 21, 30)\}$,

$T_G = (B_1, \{s_8, s_9\}) \cup (B_2, \{s_{12}, s_{13}\}) \cup (B_3, \{s_{10}, s_{11}\})$.

Then $T_2 = [T, T_G]$. The set $G$ defines a multigrading on $\mathbb{R}[p]_T$ with $A = \{e_1, e_2, e_3\} \subset \mathbb{Z}^3$. The ideal

$$\ker(\phi_T) = \langle p_{10}^1 p_{11}^2 - p_{10}^1 p_{01}^1, p_{20}^1 p_{31}^1 - p_{30}^2 p_{21}^3 \rangle$$

is not $A$-homogeneous. Thus in this case $\ker(\phi_{T_2}) \neq \ker(\phi_T) \times_A (0)$.

Example 4.8. Let $T_3$ be as in Figure 1 and $T = T_3^{(2)}$. Fix

$G = \{(00, 01, 10, 31), (20, 21, 30, 11)\}$,

$T_G = (B_1, \{s_8, s_9\}) \cup (B_2, \{s_{10}, s_{11}\})$

so $T_3 = [T, T_G]$. The set $G$ defines a multigrading on $\mathbb{R}[p]_T$ with $A = \{e_1, e_2\} \subset \mathbb{Z}^2$. The ideal

$$\ker(\phi_T) = \langle p_{10}^1 p_{11}^2 - p_{10}^1 p_{01}^1, p_{20}^1 p_{31}^1 - p_{30}^2 p_{21}^2 \rangle$$

is not $A$-homogeneous. However there is a nonempty principal subideal of $\ker(\phi_T)$ that is $A$-homogeneous. This principal ideal $Q$ is generated by the quartic $p_{10}^1 p_{11}^2 p_{20}^1 p_{31}^1 - p_{01}^1 p_{10}^1 p_{21}^2 p_{30}^2$. In this case $\ker(\phi_{T_3}) = Q \times_A (0)$. This does not fall in the context of Proposition 4.4 since $\ker(\phi_{T_3}) \neq \ker(\phi_T) \times_A (0)$.

Start with a level-one probability tree $T_1$. Let $G^1$ be a partition of $i_{T_1}$, $T_{G^1}$ a gluing component and set $T_2 = [T_1, T_{G^1}]$. In the inductive step, $T_{j+1} = [T_j, T_{G^j}]$, with $r_j := |G^j|$. At each step $j$ we also require that the label set of $T_{G^j}$ is disjoint from the label set of $T_j$. Note that $G^j$ is a partition of the set $i_{T_j}$. After $n$ iterations, we obtain a stratified staged tree $T_n$ whose set of stages is exactly $\bigcup_{j=1}^{n-1} G^j$. Whenever a staged tree $(T, \theta)$ is constructed in this way, so $T = T_n$ for some $n$, we say $T$ is an \textit{inductively constructed} staged tree.

Any stratified staged tree can be inductively constructed. This follows because the set of edge labels associated to the vertices in different levels are disjoint and because all leaves have the same level. Staged trees such as the one in Figure 3 (left) where the leaves have different levels do not fall into the definition of inductively constructed staged tree, even though the edge labels associated to vertices in two different levels are disjoint.

We will use Theorem 3.3 in Section 3 to write down the generators of inductively constructed staged trees that are balanced. The balanced condition is the combinatorial ingredient that makes the algebra of the toric fiber product work when considering iteratively constructed staged trees. This is already evident by looking at the trees in Figure 1 together with the Examples 4.6, 4.7, and 4.8. Although, all of the trees are stratified and can be inductively constructed, only $\ker(\phi_{T_1})$ can be constructed in steps by using toric fiber products. This is because $T_1$ is balanced.

Let $T_j$ be an inductively constructed staged tree and $T_{j+1} = [T_j, T_{G^j}]$, with $r_j = |G^j|$. Consider the monomial map

$$\phi_{B_j} : \mathbb{R}[p]_{T_{j+1}} \to \mathbb{R}[p_{i}^j, p_{k}^j] \mid a \in G^j_i, k \in i_{B^j_i}, i \in [r_j]),$$

$$p_{a k}^j \mapsto p_{a}^i p_{k}^j,$$

(4)
where $B_j$ denotes the exponent matrix of the monomial map $\phi_{B_j}$. Set $I_{B_j} = \ker(\phi_{B_j})$ and
\[
\text{Quad}_{B_j} = \{p_{ak_1}^i p_{bk_2}^i - p_{ak_1}^{i_1} p_{ak_2}^{i_2} \mid a, b \in G_i^j, \ k_1 \neq k_2 \in i_{E_j}, \ i \in [r_j]\}.
\]
By Proposition 10 in [17], $\text{Quad}_{B_j}$ is a Gröbner basis for $I_{B_j}$ with respect to any term order that selects the underlined terms as leading terms.

We now write down the of the elements in $\text{Lift}(F)$ for $F \subset \mathbb{R}[p]_T$. Fix $T, G$, and $T_G$ as in Definition 4.2 and denote by $A$ the multigrading of the rings $\mathbb{R}[p]_T$, $\mathbb{R}[p]_{|T, T_G|}$ defined by $G$. Following Definition 3.2 for lifts of elements in the ring $\mathbb{R}[p]_T$, we consider the $A$-homogeneous polynomial
\[
f = p_{a_1}^{i_1} p_{a_2}^{i_2} - p_{a_3}^{i_1} p_{a_2}^{i_2} \in \mathbb{R}[p]_T,
\]
where $a_1, a_2, a_3 \in G_1$, $a_2, a_4 \in G_2$ and $i_1, i_2 \in [r]$. Set $k = (k_1, k_2)$ with $k_1 \in i_{B_1}, k_2 \in i_{B_2}$ and consider $f_k \in \mathbb{R}[p]_{|T, T_G|}$ defined by
\[
f_k = p_{a_1 k_1}^{i_1} p_{a_2 k_2}^{i_2} - p_{a_3 k_1}^{i_1} p_{a_2 k_2}^{i_2}.
\]
For each $f \in F$, the set $T_f = i_{B_1} \times i_{B_2}$ is the set of associated indices and
\[
\text{Lift}(F) = \{f_k \mid f \in F, k \in T_f\}.
\]

**Definition 4.9.** Let $T_j$ be an inductively constructed staged tree with $T_{j+1} = [T_i, T_{G_i}]$ for $1 \leq i \leq n - 1$ and let $A_i$ be the grading in $\mathbb{R}[p]_{|T_i}$ determined by $G_i$. Fix two nonnegative integers $i, q$ with $1 \leq i + q \leq n - 1$. We define
\[
\text{Lift}^i(\text{Quad}_{B_i}) := \text{Lift}_{A_{i+q}}(\cdots(\text{Lift}_{A_{i+2}}(\text{Lift}_{A_{i+1}}(\text{Quad}_{B_i}))\cdots),
\]
where the subscript in $\text{Lift}_{A}(\cdot)$ indicates the grading of the argument is with respect to $A$.

We formulate a lemma that says that level-$q$ subtrees of balanced and stratified staged trees are also balanced. This leads us to consider interpolating polynomials of a vertex in two different rings. For a staged tree $(T = (V, E), \theta)$ and a vertex $v \in V$ of level $q$, we write $t_{(j)}(v)$ for the interpolating polynomial of $v$ in the level-$j$ subtree $(T^{(j)}, \theta|_{E_{(j)}})$, where $q \leq j \leq m$ and $m$ is the level of $T$. Thus $t_{(j)}(v)$ is an element of $\mathbb{R}[\Theta]_{|T^{(j)}}$.

**Lemma 4.10.** Let $(T, \theta)$ be a staged tree of level $m$ and let $q$ be a positive integer with $1 \leq q \leq m - 1$. If $(T, \theta)$ is balanced and stratified, then the level-$q$ subtree $T^{(q)}$ of $T$ is also balanced and stratified.

**Proof.** We must prove that if $a, b$ are two vertices in $T^{(q)}$ that are in the same stage, then they are a balanced pair in $\mathbb{R}[\Theta]_{|T^{(q)}}$. Since $T$ is balanced, $a, b$ are a balanced pair in $\mathbb{R}[\Theta]_T$. Namely,
\[
t_{(m)}(ak_1)t_{(m)}(bk_2) = t_{(m)}(ak_2)t_{(m)}(bk_1) \quad \text{in} \quad \mathbb{R}[\Theta]_T \text{ for all } k_1, k_2 \in \{0, \ldots, |\theta(a)| - 1\}.
\]
For a vertex $v$ in $T^{(q)}$ define $[v] = \{u \in i_{T^{(q)}} \mid$ the root-to-$u$ path in $T^{(q)}$ goes through $v\}$. Then by a repeated use of Lemma 2.7, for $e \in \{ak_1, bk_2, ak_2, bk_1\}$,
\[
t_{(m)}(e) = \sum_{u \in [e]} \prod_{e \in E(\lambda_u)} \theta(e)t_{(m)}(u),
\]
where $\lambda_u$ is the $e$ to $u$ path in $T^{(q)}$. Here $t_{(m)}(e)$ is an element of $\mathbb{R}[\Theta]_T$. Denote by $t_{m}(e)|_{T^{(q)}}$ the polynomial obtained from $t_{(m)}(e)$ by specializing $t_{(m)}(u) = 1$ for all $u \in [e]$. This specialization is a
polynomial in $\mathbb{R}[\Theta]_{\mathcal{T}(q)}$. Since $\mathcal{T}$ is stratified, $t(e)|_{\mathcal{T}(q)}$ is the interpolating polynomial $t_{(q)}(e)$ of $e$ as a vertex in $\mathcal{T}(q)$. Applying this specialization to (5) yields the balanced condition for the pair $a, b$ in $\mathbb{R}[\Theta]_{\mathcal{T}(q)}$. □

**Proposition 4.11.** Let $\mathcal{T}_i$ be a balanced and inductively constructed staged tree. Suppose $\mathcal{T}_{i+1} = [\mathcal{T}_i, \mathcal{T}_{G'}]$ and $\mathcal{T}_{i+1}$ is balanced. Then the elements in

$$\text{Lift}^{i-2}(\text{Quad}_{B_i}), \text{Lift}^{i-3}(\text{Quad}_{B_2}), \ldots, \text{Lift}(\text{Quad}_{B_{i-2}}), \text{Quad}_{B_{i-1}}$$

are $\mathcal{A}_i$-homogeneous.

**Proof.** Since $\mathcal{T}_i$ is inductively constructed, there is a sequence of stratified trees and gluing components $(\mathcal{T}_1, \mathcal{T}_{G'}), \ldots, (\mathcal{T}_{i-1}, \mathcal{T}_{G^{i-1}})$ from which $\mathcal{T}_i$ is constructed. Moreover, by Lemma 4.10 each of $\mathcal{T}_1, \ldots, \mathcal{T}_{i-1}$ is also balanced. Fix $q \in \{0, 1, \ldots, i - 2\}$ and $j = i - q - 1$, we show that the binomials in Lift$(\text{Quad}_{B_j})$ are $\mathcal{A}_j$-homogeneous. To this end we prove that for $m$ such that $0 \leq m \leq q$, the elements in Lift$(m)(\text{Quad}_{B_j})$ are $\mathcal{A}_{j+m+1}$-homogeneous. The proof is by induction on $m$.

Fix $m = 0$. We will show that the elements in Quad$_{B_j}$ are $\mathcal{A}_{j+1}$-homogeneous. The multidegrees in $\mathcal{A}_{j+1}$ are defined according to the partition $G^{j+1}$ of the leaves of $\mathcal{T}_{j+1}$. If two leaves $c, d \in \mathcal{T}_{j+1}$ in $\mathcal{T}_{j+1}$ are in the same set $G^{j+1}_\beta$ of the partition $G^{j+1}$, then deg$(p_c) = \text{deg}(p_d)$ in $\mathbb{R}[p]_{\mathcal{T}_{j+1}}$.

Since $\mathcal{T}_{j+2}$ is balanced, every pair of vertices in $\mathcal{T}_{j+2}$ in the same stage is balanced. In particular, this means that for all $a \in \{1, \ldots, r_j\}$ and $a, b \in G^j_a$

$$t_{(j+2)}(ak_1) t_{(j+2)}(bk_2) = t_{(j+2)}(bk_1) t_{(j+2)}(ak_2) \quad \text{for } k_1, k_2 \in b_{B^j_i} \quad \text{in } \mathbb{R}[\Theta]_{\mathcal{T}_{j+2}}, \quad (6)$$

where $\text{ch}(a) = \{ak \mid k \in b_{B^j_i}\}$ and $\text{ch}(b) = \{bk \mid k \in b_{B^j_i}\}$. Using the construction of $\mathcal{T}_{j+2}$ from $\mathcal{T}_{j+1}$ and $\mathcal{T}_{G^{j+1}}$, we know that for any index $c \in \{ak, bk \mid k \in b_{B^j_i}\}$, we have $t_{(j+2)}(c) = e_0^{(j+1)}k + \cdots + e_{k'}^{(j+1)}k$, where $\{e_0^{(j+1)}, \ldots, e_{k'}^{(j+1)}\}$ is the set of labels of some level-one probability tree $B^{j+1}_\beta$ in $\mathcal{T}_{G^{j+1}}$. It follows that (6) can only involve at most two sets of variables associated to two level-one probability trees in $\mathcal{T}_{G^{j+1}}$, say $B^{j+1}_\beta, B^{j+1}_\gamma$. This implies that either $\{ak_1, bk_1\} \subset G^{j+1}_\beta$ and $\{ak_2, bk_1\} \subset G^{j+1}_\gamma$ or $\{ak_1, ak_2\} \subset G^{j+1}_\beta$ and $\{bk_2, bk_1\} \subset G^{j+1}_\gamma$. We use this fact to determine the multigrading of the elements in Quad$_{B_j}$ with respect to $\mathcal{A}_{j+1}$. By definition,

$$\text{Quad}_{B_j} = \bigcup_{a=1}^{r_j} \{p_{ak_1}p_{bk_2} - p_{bk_1}p_{ak_2} \mid a, b \in G^j_a, k_1, k_2 \in b_{B^j_i}\}.$$

Thus we calculate that the element $p_{ak_1}p_{bk_2} - p_{bk_1}p_{ak_2}$ in Quad$_{B_j}$ is $\mathcal{A}_{j+1}$-homogeneous of degree $e_\beta + e_\gamma$, where $e_\beta$ and $e_\gamma$ are the multidegrees in $\mathcal{A}_{j+1}$ associated to the sets $G^{j+1}_\beta$ and $G^{j+1}_\gamma$, respectively. This completes the proof for $m = 0$. As a result, all the equations in Quad$_{B_j}$ can be lifted to elements in $\text{ker}(\varphi_{\mathcal{T}_{j+2}})$.

Suppose we have constructed Lift$(m-1)(\text{Quad}_{B_j})$ inductively by lifting the equations in Quad$_{B_j}$ and at each step all equations lift. An element in Lift$(m-1)(\text{Quad}_{B_j})$ is a binomial of the form

$$f = p_{ak_1}p_{bk_2} - p_{bk_1}p_{ak_2}, \quad (7)$$

where $a \in \{1, \ldots, r_j\}, a, b \in G^j_a, k_1, k_2 \in b_{B^j_i}$ and $s, s', u, u'$ are sequences of nonnegative integers of length $m - 1$ that arise as subindices after lifting $m - 1$ times. Note that $ak_1s, bk_2s', bk_1u, ak_2s' \in b_{\mathcal{T}_{j+m}}$. The claim is that (7) is $\mathcal{A}_{j+m}$-homogeneous.
Following a similar argument as for $m=0$, we know that two elements in the same set of the partition $G^{j+m}$ have the same multidegree with respect to $A_{j+m}$. As before, this condition can be verified for $f$ by checking that

$$\sum_{1}^{n} t_{(j+m+1)(ak_1 s)} t_{(j+m+1)(bk_2 u')} = t_{(j+m+1)(bk_1 u)} t_{(j+m+1)(ak_2 s')},$$  \quad \text{in } \mathbb{R}[\Theta]_{T_{j+m+1}}. \quad (8)$$

For $c \in \{ak_1, bk_2, ak_2, bk_1\}$, we have $[c] := \{w \in T_{j+m} \mid \text{the root-to-w path in } T_{j+m} \text{ goes through } c\}$. To check that (8) holds, consider (2) from Definition 2.9 for the vertices $a, b \in G^{j}_\alpha$. This equation is

$$t_{(j+m+1)(ak_1 s)} t_{(j+m+1)(bk_2)} = t_{(j+m+1)(bk_1)} t_{(j+m+1)(ak_2)},$$

where $k_1, k_2 \in i_{b_1}$. We use Lemma 2.7 to rewrite this equation as

$$\left( \sum_{ak_1 \in [ak_1]} \prod_{e \in E(ak_1 \to ak_1 s)} \theta(e) \right) \left( \sum_{bk_2 u' \in [bk_2]} \prod_{e \in E(bk_2 \to bk_2 u')} \theta(e) \right) = \left( \sum_{bk_1 u \in [bk_1]} \prod_{e \in E(bk_1 \to bk_1 u)} \theta(e) \right) \left( \sum_{ak_2 s' \in [ak_2]} \prod_{e \in E(ak_2 \to ak_2 s')} \theta(e) \right). \quad (9)$$

When we specialize $t_{(j+m+1)(ak_1 s)} = t_{(j+m+1)(bk_2 u')} = t_{(j+m+1)(bk_1 u)} = t_{(j+m+1)(ak_2 s')} = 1$ in each sum in (9) we get the interpolating polynomials $t_{(j+m)(ak_1)}$, $t_{(j+m)(bk_2)}$, $t_{(j+m)(bk_1)}$, $t_{(j+m)(ak_2)}$ in $\mathbb{R}[\Theta]_{T_{j+m}}$. By Lemma 4.10, $T_{j+m}$ is balanced; therefore

$$\left( \sum_{ak_1 \in [ak_1]} \prod_{e \in E(ak_1 \to ak_1 s)} \theta(e) \right) \left( \sum_{bk_2 u' \in [bk_2]} \prod_{e \in E(bk_2 \to bk_2 u')} \theta(e) \right) = \left( \sum_{bk_1 u \in [bk_1]} \prod_{e \in E(bk_1 \to bk_1 u)} \theta(e) \right) \left( \sum_{ak_2 s' \in [ak_2]} \prod_{e \in E(ak_2 \to ak_2 s')} \theta(e) \right). \quad (10)$$

The factors in the above equality are sums of monomials all with coefficients equal to 1. Thus for every pair $ak_1 s \in [ak_1]$, $bk_2 u' \in [bk_2]$ in the product of the left-hand side of the equation, there exists a pair $ak_2 s' \in [ak_2]$, $bk_2 u \in [bk_1]$ in the product of the right-hand side of the equation such that

$$\left( \prod_{e \in E(ak_1 \to ak_1 s)} \theta(e) \right) \left( \prod_{e \in E(bk_2 \to bk_2 u')} \theta(e) \right) = \left( \prod_{e \in E(bk_1 \to bk_1 u)} \theta(e) \right) \left( \prod_{e \in E(ak_2 \to ak_2 s')} \theta(e) \right). \quad (11)$$

Hence (5) for the vertices $a, b$ in $T_{j+m+1}$ can be rewritten as

$$\sum_{ak_1 \in [ak_1]} \prod_{e \in E(ak_1 \to ak_1 s)} \theta(e) \left( t_{(j+m+1)(ak_1 s)} t_{(j+m+1)(bk_2 u')} - t_{(j+m+1)(bk_1 u)} t_{(j+m+1)(ak_2 s')} \right) = 0.$$

Since $T_{j+m+1}$ is stratified, the variables that appear in the factored monomials above are different from the variables that appear in the factors of the form

$$t_{(j+m+1)(ak_1 s)} t_{(j+m+1)(bk_2 u')} - t_{(j+m+1)(bk_1 u)} t_{(j+m+1)(ak_2 s')}.$$
Hence this last equation is true only if
\[ t_{(j+m+1)}(ak_1s)t_{(j+m+1)}(bk_2u') - t_{(j+m+1)}(bk_1u)t_{(j+m+1)}(ak_2s') = 0 \]
for each summand. Following a similar argument as in the case for \( m = 0 \), this proves that the elements in \( \text{Lift}^{n-1}(\text{Quad}_B) \) are \( A_{j+m} \)-homogeneous.

We are now ready to prove our main result, Theorem 2.14, using toric fiber products for balanced and inductively constructed staged trees.

**Proof of Theorem 2.14.** If \( \mathcal{T} \) is stratified, then \( \mathcal{T} \) is an iteratively constructed staged tree and \( \mathcal{T} = \mathcal{T}_n \) for some \( n \). Set \( F_n = \text{Lift}^{n-2}(\text{Quad}_{B_1}) \cup \text{Lift}^{n-3}(\text{Quad}_{B_2}) \cup \cdots \cup \text{Quad}_{B_{n-1}} \). We prove by induction on \( n \) that \( \ker(\phi_{\mathcal{T}_n}) \) is generated by \( F_n \) and that \( F_n \) is a Gröbner basis with square-free initial ideal. The first nontrivial case is \( n = 2 \). We have \( F_2 = \text{Quad}_{B_1} \) and from Proposition 10 in [17], \( F_2 \) is a Gröbner basis for the ideal \( \ker(\phi_{\mathcal{T}_2}) = \ker(\phi_{\mathcal{T}_1}) \times A_i(0) \). Suppose the statement is true for \( i \), so the elements in \( F_i \) are a Gröbner basis for \( \ker(\phi_{\mathcal{T}_i}) \). Since \( \mathcal{T}_n \) is balanced, by Lemma 4.10 the trees \( \mathcal{T}_i \) and \( \mathcal{T}_{i+1} \) are also balanced. From Proposition 4.11 the elements in \( F_i \) are \( A_i \)-homogeneous, so by Theorem 3.3 the set \( F_{i+1} \) is a Gröbner basis for \( \ker(\phi_{\mathcal{T}_{i+1}}) \). Since the elements in \( F_n \) are all extensions of elements in \( \text{Quad}_{B_j} \) for \( j \) with \( 1 \leq j \leq n - 1 \) we see that all the terms in these binomials are square-free. Hence the initial ideal of \( F_n \) is square-free. □

**Corollary 4.12.** Let \( (\mathcal{T}, \theta) \) be a balanced and stratified staged tree. Fix \( \Delta \) to be the polytope defined by the convex hull of the lattice points in the exponent matrix of the map \( \phi_{\mathcal{T}} \). Then \( \Delta \) has a regular unimodular triangulation. In particular the toric variety defined by \( \ker(\phi_{\mathcal{T}}) \) is Cohen–Macaulay.

**Proof.** The ideal \( \ker(\phi_{\mathcal{T}}) \) has a square-free quadratic Gröbner basis with respect to a term order \( \prec \). From [16, Corollary 8.9], this induces a regular unimodular triangulation of \( \Delta \). □

## 5. Connections to discrete statistical models

Staged tree models are a class of graphical discrete statistical models introduced by Anderson and Smith in [15]. While Bayesian networks and decomposable models are defined via conditional independence statements on random variables corresponding to the vertices of a graph, staged tree models encode independence relations on the events of an outcome space represented by a tree. In the statistical literature these models are also referred to as chain event graphs. We refer the reader to the book [3] and to [19] to find out more about their statistical properties, practical implementation, and causal interpretation. In this section we give a formal definition of staged tree models and recall results from [6; 9] about their defining equations.

Given a discrete random variable \( X \) with state space \( \{0, \ldots, n\} \), a probability distribution on \( X \) is a vector \((p_0, \ldots, p_n) \in \mathbb{R}^{n+1}\), where \( p_i = P(X = i), \ i \in \{0, \ldots, n\}, \ p_i \geq 0 \) and \( \sum_{i=0}^{n} p_i = 1 \). The open probability simplex
\[ \Delta_n^0 = \{(p_0, \ldots, p_n) \in \mathbb{R}^{n+1} \mid p_i > 0, \ p_0 + \cdots + p_n = 1\} \]
consists of all the strictly positive probability distributions for a discrete random variable with state space \( \{0, \ldots, n\} \). A **discrete statistical model** is a subset of \( \Delta_n^0 \). In the next definition we associate a discrete statistical model to a given staged tree.
Definition 5.1. Let \((\mathcal{T}, \theta)\) be a staged tree. We define the parameter space

\[
\Theta_{\mathcal{T}} := \{ x \in \mathbb{R}^{|\mathcal{E}|} \mid \text{for all } e \in \mathcal{E}, \ x_{\theta(e)} \in (0, 1) \text{ and for all } a \in \mathcal{V}, \ \sum_{e \in E(a)} x_{\theta(e)} = 1 \}.
\]

Note that \(\Theta_{\mathcal{T}}\) is a product of open probability simplices. A staged tree model \(\mathcal{M}_{(\mathcal{T}, \theta)}\) is the image of the map \(\Psi_{\mathcal{T}} : \Theta_{\mathcal{T}} \to \Delta_{|\mathcal{T}|-1}^\circ\) defined by

\[
x \mapsto p_x = \left( \prod_{e \in E(\lambda_j)} x_{\theta(e)} \right)_{j \in i_{\mathcal{T}}}.
\]

We can check that, for every \(x \in \Theta_{\mathcal{T}}\), \(p_x\) is a probability distribution and therefore \(\Psi(\Theta_{\mathcal{T}}) \subset \Delta_{|\mathcal{T}|-1}^\circ\).

Two staged trees \((\mathcal{T}, \theta)\) and \((\mathcal{T'}, \theta')\) are statistically equivalent if there exists a bijection between \(\Lambda_{\mathcal{T}}\) and \(\Lambda_{\mathcal{T'}}\) in such a way that the image of \(\Psi_{\mathcal{T}}\) is equal to the image of \(\Psi_{\mathcal{T'}}\) under this bijection.

Example 5.2. The staged tree \(\mathcal{T}_1\) in Figure 1 is the staged tree representation of the decomposable model associated to the undirected graph \(G = [12][23][34]\) on four nodes.

Remark 5.3. For staged tree models, the root-to-leaf paths in the tree represent the possible unfoldings of a sequence of events. Given an edge \((v, w)\) in \(\mathcal{T}\), the label \(\theta(v, w)\) is the transition probability from \(v\) to \(w\) given arrival at \(v\).

Remark 5.4. A staged tree model \(\mathcal{M}_{(\mathcal{T}, \theta)}\) is a discrete statistical model parametrized by polynomials. The domain of this model is a semialgebraic set given by a product of simplices. As a consequence the \((\mathcal{T}, \theta)\) is equal \([6]\). However, Theorem 10 in \([6]\) states that if a staged tree \((\mathcal{T}, \theta)\) is balanced and stratified, then \(\ker(\varphi_{\mathcal{T}}) = \ker(\bar{\varphi}_{\mathcal{T}})\). Combining this result with Theorem 2.14 we can obtain Gröbner bases for staged tree model ideals whose staged tree is balanced and stratified.

Corollary 5.6. If \((\mathcal{T}, \theta)\) is a balanced and stratified staged tree, then the ideal \(\ker(\bar{\varphi}_{\mathcal{T}})\) has a quadratic Gröbner basis with square-free initial ideal.

Example 5.7. Consider the staged tree model defined by the tree \(\mathcal{T}_1\) in Figure 1 as in Example 5.2. Since this staged tree model is equal to the decomposable model given by \(G = [12][23][34]\), from \([8]\) we know it has a quadratic Gröbner basis. We recover the same result from the perspective of staged trees by using Corollary 5.6.
Corollary 5.6 is relevant in statistics because of the connection of Gröbner bases to sampling [1]. We presented Example 5.7, where a balanced and stratified staged tree represents an instance of a decomposable graphical model. We now provide more examples of staged tree models for which Corollary 5.6 holds. The first one is an explanation of the contraction axiom for conditional independence statements through the lens of staged trees. Before we present our examples we do a quick overview of discrete conditional independence models. Our exposition follows that in [18, Chapter 4]; for more details we refer the reader to [14; 5].

Let $X = (X_1, \ldots, X_n)$ be a vector of discrete random variables, where $X_i$ has state space $[d_i]$ for $i \in [n]$. The vector $X$ has state space $\mathcal{X} = [d_1] \times \cdots \times [d_n]$ and we write $p_{u_1 \cdots u_n}$ for the probability $P(X_1 = u_1, \ldots, X_n = u_n)$. We consider only positive probability distributions of the random vector $X$. For each subset $A \subset [n]$, $X_A$ is the subvector of $X$ indexed by the elements in $A$. Similarly, $X_A = \prod_{i \in A}[d_i]$ and for a vector $x \in \mathcal{X}$, $x_A$ denotes the restriction of $x$ to the indexes in $A$.

**Definition 5.8.** Let $A, B, C$ be pairwise disjoint subsets of $[n]$. The random vector $X_A$ is conditionally independent of $X_B$ given $X_C$ if for every $a \in X_A$, $b \in X_B$ and $c \in X_C$

$$P(X_A = a, X_B = b|X_C = c) = P(X_A = a|X_C = c) \cdot P(X_B = b|X_C = c).$$

The notation $X_A \perp X_B \mid X_C$ is used to denote that the random vector $X$ satisfies the conditional independence statement that $X_A$ is conditionally independent on $X_B$ given $X_C$. When $C$ is the empty set this reduces to marginal independence between $X_A$ and $X_B$.

If $\mathcal{C}$ is a list of conditional independence statements among variables in a vector $X$, the conditional independence model $\mathcal{M}_\mathcal{C}$ is the set of all probability distributions inside the open probability simplex $\Delta^\circ_{|\mathcal{X}|-1}$ that satisfy the conditional independence statements in $\mathcal{C}$. A conditional independence statement $X_A \perp X_B \mid X_C$ translates into the condition that the joint probability distribution of the variables in $X$ satisfies a set of quadratic equations. For elements $a \in X_A$, $b \in X_B$ and $c \in X_C$ we set $p_{a,b,c,+} = P(X_A = a, X_B = b, X_C = c)$.

**Proposition 5.9** [18]. If $X$ is a discrete random vector, then the independence statement $X_A \perp X_B \mid X_C$ holds for $X$ if and only if the probability distribution of $X$ satisfies

$$p_{a_1,b_1,c,+}p_{a_2,b_2,c,+} - p_{a_1,b_2,c,+}p_{a_2,b_1,c,+} = 0$$

for all $a_1, a_2 \in X_A$, $b_1, b_2 \in X_B$ and $c \in X_C$.

Let $\mathbb{R}[p_x \mid x \in \mathcal{X}]$ be the polynomial ring with one indeterminate for each element in the state space of $X$. The conditional independence ideal $I_{A \perp B \mid C}$, is the ideal in $\mathbb{R}[p_x \mid x \in \mathcal{X}]$ generated by all quadratic relations in Proposition 5.9. If $\mathcal{C}$ is a list of conditional independence statements then we define $I_\mathcal{C}$ as the sum of all conditional independence ideals associated to statements in $\mathcal{C}$.

**Example 5.10.** We consider the contraction axiom for positive distributions using staged tree models. Fix three discrete random variables $X_1, X_2, X_3$ with state spaces $[d_1 + 1]$, $[d_2 + 1]$, $[d_3 + 1]$ respectively. The contraction axiom states that the set of conditional independence statements $\mathcal{C} = \{X_1 \perp X_2 \mid X_3, X_2 \perp X_3\}$ implies the statement $X_2 \perp (X_1, X_3)$. A primary decomposition of the ideal $I_\mathcal{C}$ was obtained in [7, Theorem 1]. Here we provide a proof, using staged trees, that one of the primary components of $I_\mathcal{C}$ is
the prime binomial ideal $I_{X_2 \perp \perp (X_1, X_3)}$. As mentioned in [7] this is a well-known fact. First we explain how to represent the two statements in $C$ with a staged tree. Consider the tree $T$ in Figure 2. This tree represents the state space of the vector $(X_3, X_2, X_1)$ as a sequence of events where $X_3$ takes place first, $X_2$ second and $X_1$ third. The vertices of $T$ are indexed recursively as defined at the beginning of Section 4. The statement $X_2 \perp \perp X_3$ is represented by the stage consisting of the vertices $\{0, \ldots, d_3\}$; these are colored gray in $T$. The statement $X_1 \perp \perp X_2 \mid X_3$ is represented by the stages $S_0, \ldots, S_{d_3}$, where $S_i = \{ij \mid j \in \{0, \ldots, d_2\}\}$ and $i \in \{0, \ldots, d_3\}$. These stages mean that for a given outcome of $X_3$, the unfolding of the event $X_2$ followed by $X_1$ behaves as an independence model on two random variables. In Figure 2 the stage $S_0$ is colored in pink and the stage $S_{d_3}$ is colored in purple. Although the gray vertices are not in the same position, we can easily check that $T$ is balanced and stratified. Therefore $\ker(\varphi_T)$ has a quadratic Gröbner basis. Following the proof of Theorem 2.14 we can construct this basis explicitly. It consists of a set of quadratic equations given by the elements in $\text{Quad}_{B_2}$ coming from the stages in $S_0, \ldots, S_{d_3}$ and the lifts of the equations $\text{Quad}_{B_1}$ coming from the stage $\{0, \ldots, d_3\}$. If we swap the order of $X_1$ and $X_2$ in $T$, we obtain the staged tree $T'$ in Figure 2. This tree represents the same statistical model as $T$ now with the unfolding of events $X_3, X_1, X_2$. The gray stages in $T'$ represent the statement $X_2 \perp \perp (X_1, X_3)$. Hence, after establishing the evident bijection between the leaves of $T$ and $T'$ we see that $I_{X_2 \perp \perp (X_1, X_3)} = \ker(\varphi_{T'}) = \ker(\varphi_T)$.

One of the main differences between staged tree models and discrete Bayesian networks is that the state space of a Bayesian network is equal to the product of the state spaces of the random variables in the vertices of the graph, while the state space of a staged tree model does not necessarily have to equal a cartesian product. When $T$ is not equal to the cartesian product of some finite sets we call the tree $T$ asymmetric. The lemmas that follow are important to show that Theorem 2.14 also holds for the case when $T$ is asymmetric. This implies that we can use Theorem 2.14 to construct quadratic Gröbner bases for staged tree models whose underlying tree does not necessarily represents the outcomes of a vector of discrete random variables.
The definition of staged tree in [9] requires that each vertex in $T$ has either no or at least two outgoing edges from $v$. We stepped away from making this requirement for the staged trees we consider in Section 2. In the next lemmas we explain how this mild extension of the definition behaves with respect to the balanced condition for a pair of vertices, and how trees defined according to [9] are recovered from the more general trees we consider. Throughout the next lemmas, we fix a staged tree $(T, \theta)$ with edge set $E$ and define $E_1 = \{ e \in E \mid E(v) = \{ e \} \text{ for some } v \in V \}$. For the trees in Figure 3, $T$ has $|E_1| = 6$, while for $T'$, $|E_1| = 0$.

**Lemma 5.11.** Suppose $(T, \theta)$ is a staged tree. Let $T'$ be the staged tree obtained from $T$ by contracting the edges in $E_1$. Then $\mathcal{M}_{(T, \theta)} = \mathcal{M}_{(T', \theta)}$ and $\ker(\overline{\phi}_T) = \ker(\overline{\phi}_{T'})$.

**Proof.** First, note that the number of root-to-leaf paths in $T'$ is the same as in $T$. Moreover, each root-to-leaf path $\lambda'$ in $T'$ is obtained from a unique root-to-leaf path $\lambda$ in $T$ by contracting the edges in $E_1$. Now let $\lambda$ be a root-to-leaf path in $T$. The $\lambda$-coordinate of the map $\Psi_T$ applied to an element $\theta \in \Theta_T$ is

$$[\Psi_T(\theta)]_\lambda = \prod_{e \in E(\lambda)} \theta(e) = \prod_{e \in E(\lambda')} \theta(e) = [\Psi_{T'}(\theta |_{T'})]_{\lambda'}.$$ 

The second equality in the previous equation follows from taking a closer look at $\Theta_T$. Indeed for all $e \in E_1$ we have $\theta(e) = 1$ because of the sum-to-1 conditions imposed on $\Theta_T$ in Definition 5.1. For the third equality, $\theta |_{T'}$ denotes the restriction of the vector $\theta$ to the edge labels of $T'$. It follows from the equalities above that the coordinates of $\Psi_T$ and $\Psi_{T'}$ are equal. Therefore $\mathcal{M}_{(T, \theta)} = \mathcal{M}_{(T', \theta)}$. A similar argument applied to the maps $\overline{\phi}_T$ and $\overline{\phi}_{T'}$ shows that $\ker(\overline{\phi}_T) = \ker(\overline{\phi}_{T'})$. To carry out this argument we need to reindex the leaves of the trees; this can be done by dropping the index of the elements in $E_1$. □

We illustrate Lemma 5.11 in Figure 3 where $T'$ is obtained from $T$ by contracting the six edges in $E_1$. The two staged trees in this figure define the same statistical model.
Remark 5.12. To prove Corollary 5.6 we used [6, Theorem 10]. The proof of Theorem 10 in [6] is presented for trees such that \( E_1 = 0 \). However the result still holds when \( |E_1| > 1 \) because the ideal \( I_{\text{Paths}} \) (from [6]) is contained in \( \text{ker}(\varphi_T) \) in this case also; see [6] for more details.

Lemma 5.13. Suppose \((T, \theta)\) is a balanced and stratified staged tree. Let \( T' \) be the tree obtained from \( T \) by contracting the edges in \( E_1 \). Then \((T', \theta)\) is also balanced.

Proof. Suppose \( T \) is balanced and \( a, b \) are in the same stage. Following the notation from Definition 2.9, we have \( t(a_i) t(b_j) = t(b_j) t(a_i) \) for all \( i \neq j \in \{0, 1, \ldots, k\} \). We specialize \( \theta(e) = 1 \) in this equation for all \( e \in E_1 \) to obtain \( t(a_i) t(b_j) |_{\theta(e)=1, e \in E_1} = t(b_j) t(a_i) |_{\theta(e)=1, e \in E_1} \) in \( \mathbb{R}[\Theta]_{T'} \). Therefore \( T' \) is also balanced.

Corollary 5.14. Suppose \( T \) is a balanced and stratified staged tree. Let \( T' \) be the staged tree obtained from \( T \) by contracting the edges in \( E_1 \). Then \( \text{ker}(\bar{\varphi}_T) \) is a toric ideal with a quadratic Gröbner basis whose initial ideal is square-free.

Proof. From Corollary 5.6 it follows that \( \text{ker}(\bar{\varphi}_T) \) is a toric ideal with a quadratic Gröbner basis and square-free initial ideal. After an appropriate bijection, by Lemma 5.11, \( \text{ker}(\bar{\varphi}_T) = \text{ker}(\bar{\varphi}_T') \).

We illustrate the result in Corollary 5.14 with an example.

Example 5.15. Fix \( T \) and \( T' \) to be the staged trees in Figure 3. The staged tree \( T' \) is considered in [6, Section 6] as an example of the possible unfolding of events in a cell culture. A thorough discussion of this example and its difference with other graphical models is also contained in [6, Section 6]. Here we explain how to obtain a Gröbner basis for \( \text{ker}(\varphi_{T'}) \) using Corollary 5.14. The tree \( T' \) is balanced and statistically equivalent to \( T \). By Corollary 5.6, \( T \) has a quadratic Gröbner basis with square-free initial ideal. Using the lemmas preceding this example, there is a bijection between the root-to-leaf paths in \( T \) and \( T' \); thus \( \mathbb{R}[p]_T \) and \( \mathbb{R}[p]_{T'} \) are isomorphic. Under this isomorphism, the Gröbner basis for \( \text{ker}(\varphi_T) \) is a Gröbner basis for \( \text{ker}(\bar{\varphi}_T') \); its generators are

\[
\begin{align*}
p_{0111} p_{10} - p_{0011} p_{110}, & \quad p_{0011} p_{0110} - p_{0010} p_{0111}, & \quad p_{0110} p_{10} - p_{0010} p_{110}, \\
p_{0010} p_{010} - p_{0000} p_{0110}, & \quad p_{0011} p_{010} - p_{0000} p_{0111}, & \quad p_{010} p_{10} - p_{0000} p_{110}.
\end{align*}
\]

Acknowledgements

Computations using the symbolic algebra software Macaulay2 [12] were crucial for the development of this paper. We thank Thomas Kahle for his support in the completion of this project and Bernd Sturmfels for a careful reading of an earlier version of this manuscript. Both authors were supported by the Deutsche Forschungsgemeinschaft (314838170, GRK 2297 MathCoRe).

References


Received 2019-10-07. Revised 2020-12-21. Accepted 2021-01-05.

LAMPRINI ANANIADI: lamprini.ananiadi@ovgu.de
Otto-Von-Guericke Universität, Magdeburg, Germany

ELIANA DUARTE: eliana.duarte@ovgu.de
Otto-Von-Guericke Universität, Magdeburg, Germany

and

Max-Planck-Institute for Mathematics in the Sciences, Leipzig, Germany
EQUIVARIANT HILBERT SERIES FOR HIERARCHICAL MODELS

AIDA MARAJ AND UWE NAGEL

Toric ideals to hierarchical models are invariant under the action of a product of symmetric groups. Taking the number of factors, say, \( m \), into account, we introduce and study invariant filtrations and their equivariant Hilbert series. We present a condition that guarantees that the equivariant Hilbert series is a rational function in \( m+1 \) variables with rational coefficients. Furthermore we give explicit formulas for the rational functions with coefficients in a number field and an algorithm for determining the rational functions with rational coefficients. A key is to construct finite automata that recognize languages corresponding to invariant filtrations.

1. Introduction

Hierarchical models are used in algebraic statistics to determine dependencies among random variables; see, e.g., [17]. Such a model is determined by a simplicial complex and the number of states each random variable can take. The Markov basis to any hierarchical model corresponds to a generating set of an associated toric ideal; see [3]. This toric ideal is rather symmetric; that is, it is invariant under the action of a product of symmetric groups. The number of minimal generators of the toric ideals grows rapidly when the number of states of the considered random variables increases. However, the independent set theorem (see Theorem 2.4) shows that the symmetry can be leveraged to describe, for a fixed simplicial complex, simultaneously the generating sets and thus Markov bases for all numbers of states of the random variables. The conceptional proof of this result by Hillar and Sullivant [7] introduces the notion of an \( S_\infty \)-invariant filtration. Informally, this is a sequence \((I_n)_{n \in \mathbb{N}}\) of compatible ideals \(I_n\) in polynomial rings \(R_n\) whose number of variables increases with \(n\) and where each \(I_n\) is invariant under the action of a symmetric group that permutes the variables of \(R_n\). To such a filtration, the second author and Römer [14] introduced an equivariant Hilbert series in order to analyze simultaneously quantitative properties of the ideals in the filtration. It is a formal power series in two variables and they showed that it is rational with rational coefficients [14, Theorem 7.8].

The variables occurring in the elements of a toric ideal to a hierarchical model can naturally be grouped into \(m\) sets of variables, where \(m\) is the number of random variables. Permuting the variables in any one of these groups gives a group action that leaves the ideal invariant. This suggests the introduction of an \(S_\infty^m\)-invariant filtration (see Definition 2.2). For \(m = 1\) it specializes to the filtrations mentioned above. Every \(S_\infty^m\)-invariant filtration naturally gives rise to an equivariant Hilbert series defined as a formal power series in \(m+1\) variables (see Definition 3.1). Our main result gives a condition guaranteeing...
that this power series is a rational function in \( m + 1 \) variables with rational coefficients (see Theorem 3.5). Furthermore, we present two methods to determine this rational function. One approach is more special and produces an explicit rational function, but with coefficients in a suitable extension field of the rational numbers (see Proposition 5.4). The other approach is much more general and gives directly a formula for the rational function with rational coefficients. It determines the equivariant Hilbert series as the generating function of a regular language (see Section 5).

The remaining part of this paper is organized as follows. In Section 2, we discuss the symmetry of toric ideals to hierarchical models and introduce \( S^m_{\infty} \)-invariant filtrations. Their equivariant Hilbert series in \( m + 1 \) variables are studied in Section 3. Our main result about such Hilbert series is stated as Theorem 3.5. We reduce its proof to a special case in that section, but postpone the argument for the special case to the following section. In Section 4 we use regular languages and finite automata to establish the special case. The idea is to encode the monomials that determine the Hilbert series by a language. We then construct a deterministic finite automaton that recognizes this language. Thus, the language is regular. Using a suitable weight function we then show that the corresponding generating function of the language is essentially the desired Hilbert series. Since generating functions of regular languages are rational, this completes the argument of our main result. Furthermore, using the finite automaton that describes a regular language, there is an algorithm that determines the generating function of the language explicitly as a rational function with rational coefficients. This is explained and illustrated in Section 5. We also describe in that section a more limited direct approach that gives an explicit formula for the rational function, but with coefficients in a number field.

\section{Symmetry and filtrations}

After reviewing needed concepts and notation we introduce \( S^m_{\infty} \)-invariant filtrations in this section.

Throughout this paper we use \( \mathbb{N} \) to denote the set of positive integers and \( \mathbb{N}_0 \) to denote the set of nonnegative integers. For any \( q \in \mathbb{N} \), we set \( \{q\} = \{1, 2, \ldots, q\} \), and so \( \{0\} = \emptyset \). We use \#\( T \) to denote the number of elements in a finite set \( T \).

A hierarchical model \( \mathcal{M} = \mathcal{M}(\Delta, \mathbf{r}) \) with \( m \) parameters is given by a collection \( \Delta = \{F_1, F_2, \ldots, F_q\} \) of nonempty subsets \( F_j \subset [m] \) with \( \bigcup_{j \in \{q\}} F_j = [m] \) and a vector \( \mathbf{r} = (r_1, r_2, \ldots, r_m) \in \mathbb{N}^m \). Each parameter corresponds to a random variable, and \( r_i \) denotes the number of values parameter \( i \) can take. We refer to \( \mathbf{r} \) as the \textit{vector of states}. Every set \( F_j \) indicates a dependency among the parameters corresponding to its vertices. Thus, we may assume that no \( F_j \) is contained in some \( F_i \) with \( i \neq j \) and refer to the sets \( F_j \) as \textit{facets}.

Diaconis and Sturmfels [3] pioneered the use of algebraic methods in order to study statistical models. We need some notation. For any subset \( F = \{i_1, i_2, \ldots, i_s\} \subset [m] \), we write

\[ \mathbf{r}_F = (r_{i_1}, r_{i_2}, \ldots, r_{i_s}) \in \mathbb{N}^s \quad \text{and} \quad [\mathbf{r}_F] = [r_{i_1}] \times [r_{i_2}] \times \cdots \times [r_{i_s}] \subset \mathbb{N}^s. \]

In particular, \([\mathbf{r}_{\{m\}}] = [\mathbf{r}] \subset \mathbb{N}^m\). Given a field \( \mathbb{K} \) and a hierarchical model \( \mathcal{M} = \mathcal{M}(\mathbf{r}, \Delta) \), consider the ring homomorphism

\[ \Phi_{\mathcal{M}} : \mathbb{K}\mathbf{r} = \mathbb{K}[x_i \mid i \in [\mathbf{r}]] \longrightarrow \mathcal{S}_{\mathcal{M}} = \mathbb{K}[y_{j,i_F} \mid F_j \in \Delta, i_F \in [\mathbf{r}_{F_j}]], \]

\[ x_i \longmapsto \prod_{F_j \in \Delta} y_{j,i_{F_j}}. \] (2-1)
We also refer to $R_I$ whose homogeneous ideal is $I$. We also refer to $R/I_M$ as the coordinate ring of the model $M$.

In the simplest cases explicit sets of generators of such ideals are known. We use the standard partial order $\leq$ on $\mathbb{Z}^n$ given by $i = (i_1, \ldots, i_n) \leq j = (j_1, \ldots, j_n)$ if $i_1 \leq j_1$, $\ldots$, $i_n \leq j_n$. If $q = 1$ then $\Phi_M$ is an isomorphism, and so $I_M$ is zero.

**Example 2.1.** Let $q = 2$; i.e., $\Delta = \{F_1, F_2\}$.

(i) Suppose first that $F_1$ and $F_2$ are disjoint. Possibly permuting the positions of the entries of a vector $i \in [r] = [r_{F_1} \cup F_2]$, we write $x_{i_{F_1}, i_{F_2}}$ instead of $x_i$. This corresponds to a bijection $[r_{F_1} \cup F_2] \to [r_{F_1}] \times [r_{F_2}]$. Using this notation, a generating set of $I_M$ is (see, e.g., [2; 3])

$$G(M(r, \{F_1, F_2\})) = \{x_{i_{F_1}, i_{F_2}} x_{i'_{F_1}, i'_{F_2}} - x_{i_{F_1}, i'_{F_2}} x_{i'_{F_1}, i_{F_2}} \mid i_{F_1} < i'_{F_1} \in [r_{F_1}], \ i_{F_2} < i'_{F_2} \in [r_{F_2}]\},$$

In the special case, where $m = 2$ and, say, $F_1 = \{1\}$, $F_2 = \{2\}$, this set becomes

$$\{x_{i_{F_1}, i_{F_2}} x_{i'_{F_1}, i'_{F_2}} - x_{i_{F_1}, i'_{F_2}} x_{i'_{F_1}, i_{F_2}} \mid 1 \leq i_1 \leq i'_1 \leq r_1, \ 1 \leq i_2 \leq i'_2 \leq r_2\},$$

which is the set of $2 \times 2$ minors of a generic $r_1 \times r_2$ matrix with entries $x_{i_1, i_2}$. The image of the map $\Phi_M$ in this case is known in algebraic geometry as the coordinate ring of the Segre product $\mathbb{P}^{n-1} \times \mathbb{P}^{m-1}$ whose homogeneous ideal is $I_M$.

(ii) Consider now the general case, where $F_1$ and $F_2$ are not necessarily disjoint. Note that $[m]$ is the disjoint union of $F_1 \setminus F_2$, $F_2 \setminus F_1$ and $F_1 \cap F_2$. Thus, possibly permuting the positions of the entries of $i \in [r]$ as above, we write $x_{i_{F_1 \setminus F_2}, i_{F_1 \cap F_2}, i_{F_2 \setminus F_1}}$ for $x_i$. Fixing a vector $c \in [r_{F_1 \cap F_2}]$, we define a set $G^c(M(r_{[m] \setminus F_1 \cap F_2}, \{F_1 \setminus F_2, F_2 \setminus F_1\}))$ whose elements are

$$x_{i_{F_1 \setminus F_2}, c} i_{F_1 \cap F_2} x_{i'_{F_1 \setminus F_2}, c} i'_{F_2 \setminus F_1} - x_{i_{F_1 \setminus F_2}, c} i_{F_1 \cap F_2} x_{i'_{F_2 \setminus F_1}, c} i'_{F_2 \setminus F_1},$$

where

$$i_{F_1 \setminus F_2} < i'_{F_1 \setminus F_2} \in [r_{F_1 \setminus F_2}] \quad \text{and} \quad i_{F_2 \setminus F_1} < i'_{F_2 \setminus F_1} \in [r_{F_2 \setminus F_1}].$$

The collection

$$G(M(r, \{F_1, F_2\})) = \bigcup_{c \in r_{F_1 \cap F_2}} G^c(M(r_{[m] \setminus F_1 \cap F_2}, \{F_1 \setminus F_2, F_2 \setminus F_1\}))$$

is a generating set for the ideal $I_M(r, \{F_1, F_2\})$; see [4; 8].

Even in the simple cases of Example 2.1, the number of minimal generators of a toric ideal $I_M$ is large if the entries of $r$ are large. However, many of these generators have similar shape. This can be made precise using symmetry.

Indeed, denote by $\mathcal{S}_n$ the symmetric group in $n$ letters. Set $\mathcal{S}_r = \mathcal{S}_{r_1} \times \mathcal{S}_{r_2} \times \cdots \times \mathcal{S}_{r_m}$. This group acts on the polynomial ring $R_r$ by permuting the indices of its variables, that is,

$$(\sigma_1, \ldots, \sigma_m) \cdot x_i = x_{(\sigma_1(i_1), \ldots, \sigma_m(i_m))}.$$
whose objects are finite sets and whose morphisms are injections. This approach can be extended to any
\( r \in S_{[r]} \) and \( f \in I_M \). In some cases, this invariance can be used to obtain all minimal generators of \( I_M \)
from a subset by using symmetry. For example, in the special case \( m = q = 2 \), \( F_1 = \{1\}, F_2 = \{2\} \), with
\( r_1, r_2 \geq 2 \), the set \( G(M(r, \{F_1, F_2\})) \) can be obtained from
\[
x_{1,1}x_{2,2} - x_{1,2}x_{2,1}
\]
using the action of \( S_{r_1} \times S_{r_2} \). Note that this is true for every vector \( r = (r_1, r_2) \). There is a vast generalization
of this observation using the concept of an invariant filtration.

The symmetric group \( S_n \) is naturally embedded into \( S_{n+1} \) as the stabilizer of \( \{n + 1\} \). Using this
construction componentwise, we get an embedding of \( S_{[r]} \) into \( S_{[r']} \) if \( r \leq r' \). Set
\[
S^m_{\infty} = \bigcup_{r \in \mathbb{N}^m} S_{[r]}.
\]

**Definition 2.2.** An \( S^m_{\infty} \)-invariant filtration is a family \((I_r)_{r \in \mathbb{N}^m}\) of ideals \( I_r \subset R_r \) such that every ideal \( I_r \)
is \( S_{[r]} \)-invariant and, as subsets of \( R_{r'} \),
\[
S_{[r']} : I_r \subset I_{r'} \quad \text{whenever} \quad r \leq r'.
\]

Note that fixing \( \Delta \), the ideals \((I_{M(\Delta,r)})_{r \in \mathbb{N}_0^m}\) form an \( S^m_{\infty} \)-invariant filtration. It is useful to extend these ideas.

**Remark 2.3.** Let \( T \) be any nonempty subset of \([m]\). For vectors \( r \in \mathbb{N}^m \), we want to fix the entries in
positions supported at \( T \), but vary the other entries. To this end write \((r_{[m] \setminus T}, T)\) instead of \( r \).

Fix a vector \( c \in \mathbb{N}^m \setminus \#T \). Let \( I_{c, r_T} \subset R_T \) be an \( S^m_{\infty} \)-invariant filtration. Restricting \( S_{[r]} \) and its action to
components supported at \( T \), we get an \( S^m_{\infty} \)-invariant filtration of ideals \( I_{c, r_T} \subset R_{c, r_T} \) with \( r_T \in \mathbb{N}^T \).

Note that this idea applies to the ideals \( I_{M(\Delta,r)} \) with fixed \( \Delta \). We can now state the mentioned extension
of the example given above Definition 2.2. It is called independent set theorem and has been established
by Hillar and Sullivant in [7, Theorem 4.7]; see also [5].

**Theorem 2.4.** Fix \( \Delta \) and consider a subset \( T \subset [m] \) such that \#(\( F_j \cap T \)) \leq 1 \) for every \( j \in [q] \). Assume
the number of states of every parameter \( j \in [m] \setminus T \) is fixed, and consider the hierarchical models
\( M(\Delta, r_T) = M(\Delta, (c, r_T)) \), where \( c \in \mathbb{N}^m - \#T \). Then the ideals \( I_{M(\Delta,r)} \) form an \( S^m_{\infty} \)-invariant filtration
\( J_{\Delta,r, m \setminus T} = (I_{M(\Delta,r)})_{r_T \in \mathbb{N}^T} \), that is, there is some \( d \in \mathbb{N}^T \) such that \( S_{[r_T]} : I_{M(\Delta,d)} \) generates in \( R_{c, r_T} \)
the ideal \( I_{M(\Delta,r)} \) whenever \( r_T \geq d \).

In other words, this result says that a generating set of the ideal \( I_{M(\Delta,r)} \) can be obtained from a fixed
finite minimal generating set of \( I_{M(\Delta,(c,d))} \) by applying suitable permutations whenever the number of
states of every parameter in \([m] \setminus T \) is large enough.

Theorem 2.4 is not true without an assumption on the set \( T \); see [7, Example 4.3].

**Remark 2.5.** An \( S^m_{\infty} \)-invariant filtration can also be described using a categorical framework. Indeed, if
\( m = 1 \) this approach has been used in [15] to study also sequences of modules by using the category FI,
whose objects are finite sets and whose morphisms are injections. This approach can be extended to any
\( m \geq 1 \) using the category \( \text{FI}^m \) (see, e.g., [12] in the case of modules over a fixed ring). For conceptional
simplicity we prefer to use invariant filtrations in this paper.
3. Equivariant Hilbert series

In order to study asymptotic properties of ideals in an \(S_\infty\)-invariant filtration, an equivariant Hilbert series was introduced in [14]. Here we study an extension of this concept for \(S_m^m\)-invariant filtrations.

We begin by recalling some basic facts. Let \(I\) be a homogeneous ideal in a polynomial ring \(R\) in finitely many variables over some field \(\mathbb{k}\). We will always assume that any variable has degree 1. Thus, \(R/I = \bigoplus_{j \geq 0} [R/I]_j\) is a standard graded \(\mathbb{k}\)-algebra. Its Hilbert series is the formal power series

\[
H_{R/I}(t) = \sum_{j \geq 0} \dim_{\mathbb{k}} [R/I]_j t^j.
\]

By Hilbert’s theorem (see, e.g., [1, Corollary 4.1.8]), it is rational and can be uniquely written as

\[
H_{R/I}(t) = \frac{g(t)}{(1-t)^{\dim R/I}},
\]

with \(g(t) \in \mathbb{Z}[t]\) and \(g(1) > 0\), unless \(I = R\). The number \(g(1)\) is called the degree of \(I\).

**Definition 3.1.** The equivariant Hilbert series of an \(S_m^m\)-invariant filtration \(\mathcal{F} = (I_r)_{r \in \mathbb{N}^m}\) of ideals \(I_r \subset R_r\) is the formal power series in variables \(s_1, \ldots, s_m, t\)

\[
equivH_{\mathcal{F}}(s_1, \ldots, s_m, t) = \sum_{r \in \mathbb{N}^m} H_{R_r/I_r}(t) \cdot s_1^{r_1} \cdots s_m^{r_m}
\]

\[
= \sum_{r \in \mathbb{N}^m} \sum_{j \geq 0} \dim_{\mathbb{k}} [R_r/I_r]_j \cdot s_1^{r_1} \cdots s_m^{r_m} t^j.
\]

If \(m = 1\), that is, \(\mathcal{F}\) is an \(S_\infty\)-invariant filtration, the Hilbert series of \(\mathcal{F}\) is always rational by [14, Theorem 7.8] or [11, Theorem 4.3]. For \(m \geq 1\), one can also consider another formal power series by focusing on components whose degree is on the diagonal of \(\mathbb{N}^m\). This gives

\[
\sum_{r \in \mathbb{N}^m} H_{R_{(r, \ldots, r)}/I_{(r, \ldots, r)}}(t) \cdot s^r.
\]

It is open whether this formal power series is rational if \(m \geq 2\), even if the ideals are trivial.

**Example 3.2.** Let \(m = 2\) and consider the filtration \(\mathcal{F} = (I_r)\), where every ideal \(I_r\) is zero. Since the ring \(R_{(r_1, r_2)}\) has dimension \(r_1 r_2\), one obtains

\[
equivH_{\mathcal{F}}(s_1, s_2, t) = \sum_{(r_1, r_2) \in \mathbb{N}^2} H_{R_{(r_1, r_2)}}(t) \cdot s_1^{r_1} s_2^{r_2} = \sum_{(r_1, r_2) \in \mathbb{N}^2} \frac{1}{(1-t)^{r_1} r_2} \cdot s_1^{r_1} s_2^{r_2} = \sum_{r_1 \geq 1} \left[-1 + \frac{(1-t)^{r_1}}{(1-t)^{r_1} - s_2^{r_1}}\right].
\]

We do not know if this is a rational function in \(s_1, s_2\) and \(t\). However, if one considers the more standard Hilbert series with \(r = r_1 = r_2\), one gets

\[
\sum_{r \geq 0} H_{R_{(r, r)}}(t) \cdot s^r = \sum_{n \geq 1} \frac{1}{(1-t)^n} \cdot s^r.
\]

This is not a rational function because the sequence \((1/(1-t)^r)_{r \in \mathbb{N}}\) does not satisfy a finite linear recurrence relation with coefficients in \(\mathbb{Q}(t)\).
For the remainder of this section we restrict ourselves to considering ideals of hierarchical models $\mathcal{M}(\Delta, r)$. As pointed out in Remark 2.3, for any subset $T \neq \emptyset$ of $[m]$, these ideals give rise to $\mathcal{S}_m^{\infty}$-invariant filtrations. To study their equivariant Hilbert series, it is convenient to simplify notation. We may assume that $T = \{m - #T + 1, \ldots, m\}$ and fix the entries of $r$ in positions supported on $[m] \setminus T$; that is, we fix $c \in \mathbb{N}^{m - #T}$ and set $n = (n_1, \ldots, n_{m - #T}) = r_T$ for $r \in \mathbb{N}^m$ to obtain $r = (c, n)$. We write $\mathcal{M}(\Delta, n)$ instead of $\mathcal{M}(\Delta, (c, n))$ and denote the resulting $\mathcal{S}_m^{m - #T}$-invariant filtration $(I_{\mathcal{M}(\Delta, n)})_{n \in \mathbb{N}^{m - #T}}$ by $\mathcal{J}_{\Delta, r_{[m] \setminus T}}$, as in the independent set theorem. Its equivariant Hilbert series is

$$\text{equivH}_{\mathcal{J}_{\Delta, r_{[m] \setminus T}}} (s_1, s_2, \ldots, s_{#T}, t) = \sum_{n \in \mathbb{N}^{#T}} H_{R_{(c, n)}/I_{\mathcal{M}(\Delta, n)}} (t) \cdot s_1^{n_1} \cdots s_{#T}^{n_{#T}}.$$ 

The independent set theorem (Theorem 2.4) guarantees stabilization of the filtration. This suggests the following problem.

**Question 3.3.** If $T \subset [m]$ satisfies $\#(F \cap T) \leq 1$ for every facet $F$ of $\Delta$, is the equivariant Hilbert series of $\mathcal{J}_{\Delta, r_{[m] \setminus T}}$ rational?

The answer is affirmative if $T$ consists of exactly one element.

**Proposition 3.4.** If $\#T = 1$, then the equivariant Hilbert series of $\mathcal{J}_{\Delta, r_{[m] \setminus T}}$ is rational.

**Proof.** The assumption means $T = \{m\}$ and $r = (c, n)$ with $c \in \mathbb{N}^{m - 1}$ and $n \in \mathbb{N}$. Set $c = c_1 \cdots c_{m - 1}$ and fix a bijection

$$\psi : [c] = [c_1] \times \cdots \times [c_{m - 1}] \to [c].$$

For every $n \in \mathbb{N}$, it induces a ring isomorphism

$$R_{(c, n)} = \mathbb{K}[x_{i, j} \mid (i, j) \in [c] \times [n]] \to \mathbb{K}[x_{i, j} \mid (i, j) \in [c] \times [n]] = R'_{n},$$

$$x_{i, j} \mapsto x_{\psi(i), j}.$$ 

This isomorphism maps every ideal $I_{\mathcal{M}(\Delta, n)}$ corresponding to the model $\mathcal{M}(\Delta, (c, n))$ onto an $S_n$-invariant ideal $I_n$. In particular, the rings $R_{(c, n)}/I_{\mathcal{M}(\Delta, n)}$ and $R'_n/I_n$ have the same Hilbert series and the family $(I_n)_{n \in \mathbb{N}}$ is an $\mathcal{S}_\infty$-invariant filtration. Thus, its equivariant Hilbert series is rational by [14, Theorem 7.8] or [11, Theorem 4.3].

Our main result in this section describes further cases in which the answer to Question 3.3 is affirmative.

**Theorem 3.5.** The equivariant Hilbert series of $\mathcal{J}_{\Delta, r_{[m] \setminus T}}$ is a rational function with rational coefficients if

1. $F_i \cap F_j = \emptyset$ for any distinct $F_i$, $F_j \in \Delta$, and
2. $|F \cap T| \leq 1$ for any $F \in \Delta$.

This results applies in particular to the independence model, where it takes an attractive form.

**Example 3.6.** A hierarchical model describing $m$ independent parameters is called independence model. Its collection of facets is $\Delta = \{1\}, \{2\}, \ldots, \{m\}$. Thus, we may apply Theorem 3.5 with any subset $T$ of $[m]$. Using $T = [m]$, we show in Example 5.5 below that

$$\text{equivH}_{\mathcal{J}_{\Delta, r_{[m] \setminus T}}} (s_1, s_2, \ldots, s_m, t) = \sum_{n \in \mathbb{N}^m} H_{R_{n}/I_{\mathcal{M}(\Delta, n)}} (t) \cdot s_1^{n_1} \cdots s_m^{n_m} = \frac{s_1 \cdots s_m}{(1 - s_1) \cdots (1 - s_m) - t}.$$
The proof of Theorem 3.5 will be given in two steps. First we show that it is enough to prove the result in a special case where every facet consists of two elements. Second, we use regular languages to show the desired rationality in the following section.

In the remainder of this section we establish the reduction step.

**Lemma 3.7.** Consider a collection $\Delta = \{F_1, \ldots, F_q\}$ on vertex set $[m]$ and a subset $T$ of $[m]$ satisfying

1. $F_i \cap F_j = \emptyset$ for any $F_i, F_j \in \Delta$, and
2. $|F \cap T| = 1$ for any $F \in \Delta$.

Then there is a collection $\Delta' = \{F'_1, \ldots, F'_q\}$ on vertex set $[m']$ consisting of two element facets and also satisfying conditions (1) and (2) with the property that for every $c \in \mathbb{N}^{m-\#T}$ there is some $c' \in \mathbb{N}^{m'-\#T}$ such that the filtrations corresponding to the models $\mathcal{M}(\Delta, (c, n))$ and $\mathcal{M}(\Delta', (c', n))$ with $n \in \mathbb{N}^{\#T}$ have the same equivariant Hilbert series.

**Proof.** The assumptions imply that $T$ must have $q$ elements. We may assume that every facet in $\Delta$ has at least two elements. Indeed, if $F \in \Delta$ has only one element then we may replace $F$ by the union $F'$ of $F$ and a new vertex. Assigning to the parameter corresponding to the new vertex exactly one possible state gives a new model whose coordinate ring has the same Hilbert series as the original model.

Given such a hierarchical model $\mathcal{M}_n = \mathcal{M}(\Delta, (c, n))$ on vertex set $[m]$, we will construct a new hierarchical model $\mathcal{M}'_n = \mathcal{M}(\Delta', (c', n))$ on $m' = 2q$ vertices that has the same Hilbert series. The new vertex set is the disjoint union of the $q$ vertices in $F_j \cap T$ with $j \in [q]$ and a set $V$ of $q$ other vertices, say $V = [q]$. For $j \in [q]$, set $F'_j = \{j\} \cup (F_j \cap T)$. Thus, the sets $F'_j$ are pairwise disjoint because $F_1, \ldots, F_q$ have this property, and each $F'_j$ has exactly two elements. In particular, $\Delta' = \{F'_1, \ldots, F'_q\}$ and $T$ satisfy conditions (1) and (2).

Now let $c'_j = \prod_{e \in F'_j \setminus T} c_e = |c_{F'_j \setminus T}|$ be the number of states of the parameter corresponding to the vertex $j \in F'_j$. Furthermore, for every $j \in [q]$, let the parameter corresponding to the vertex $F'_j \cap T$ have the same number of states as $F_j \cap T$ in $\mathcal{M}_n$. This completes the definition of a new hierarchical model $\mathcal{M}'_n = \mathcal{M}(\Delta', (c', n))$. The passage from $\mathcal{M}_n$ to $\mathcal{M}'_n$ is illustrated in the example below:

\[
\begin{array}{ccc}
\begin{array}{ccc}
c_1 & c_2 & c_3 \\
n_1 & n_2 & n_3
\end{array}
& \longrightarrow & \begin{array}{ccc}
c'_1 = c_1 c_2 & c'_2 = 1 & c'_3 = c_3 \\
n_1 & n_2 & n_3
\end{array}
\end{array}
\]

$\Delta = \{124, 5, 36\}$, $r = (c_1, c_2, c_3, n_1, n_2, n_3)$ \quad $\longrightarrow$ \quad $\Delta' = \{14, 25, 36\}$, $r' = (c'_1, 1, c'_3, n_1, n_2, n_3)$.

Varying $n \in \mathbb{N}^q$, the ideals $I_{\mathcal{M}_n}$ form an $S^q_\infty$-invariant filtration. Thus, to establish the assertion it is enough to prove that for every $n \in \mathbb{N}^q$, the quotient rings $R_n/I_{\mathcal{M}_n}$ and $R'_n/I_{\mathcal{M}'_n}$ are isomorphic.

For every $F_j \in \Delta$, the sets $[c_{F_j \setminus T}]$ and $[c'_j]$ have the same finite cardinality. Choose a bijection $\psi_j : [c_{F_j \setminus T}] \longrightarrow [c'_j]$.
These choices determine two further bijections
\[
(\psi_1, \ldots, \psi_q, \text{id}_{[n]}) : [c_{F_1} \setminus T] \times \cdots \times [c_{F_q} \setminus T] \times [n] \longrightarrow [c_1'] \times \cdots \times [c_q'] \times [n], \tag{3-1}
\]
\[
(\psi_j, \text{id}_{[n_j]}) : [c_{F_j} \setminus T] \times [n_j] \longrightarrow [c_j'] \times [n_j]. \tag{3-2}
\]

Bijection (3-1) induces the isomorphism of polynomial rings
\[
\Psi : R(c, n) = \mathbb{K}[x_{i_{F_1} \setminus T}, \ldots, x_{i_{F_q} \setminus T}, k] \mid i_{F_j} \in [c_{F_j} \setminus T], k \in [n]] \longrightarrow \mathbb{K}[x_{i_1}, \ldots, x_{i_q}, k \mid i_j \in [c_j'], k \in [n]] = R_n',
\]
\[
\quad x_{i_{F_1} \setminus T}, \ldots, x_{i_{F_q} \setminus T}, k \longmapsto x_{\psi_1(i_{F_1} \setminus T)}, \ldots, x_{\psi_q(i_{F_q} \setminus T)}, k.
\]

Similarly, bijection (3-2) induces the isomorphism of polynomial rings
\[
\Psi' : S_n = \mathbb{K}[y_{j, i_{F_j} \setminus T}, k_j \mid 1 \leq j \leq q, i_{F_j} \in [c_{F_j} \setminus T], k_j \in [n_j]] \longrightarrow \mathbb{K}[y_{j, i}, k_j \mid 1 \leq j \leq q, i_j \in [c_j'], k_j \in [n_j]] = S_n',
\]
\[
\quad y_{j, i_{F_j} \setminus T}, k_j \longmapsto y_{j, \psi_j(i_{F_j} \setminus T), k_j}.
\]

We claim that the following diagram is commutative:
\[
\begin{array}{ccc}
R(c, n) & \xrightarrow{\Phi_M} & S_n \\
\downarrow{\Psi} & & \downarrow{\Psi'} \\
R_n' & \xrightarrow{\Phi_{M'}} & S_n'
\end{array}
\tag{3-3}
\]

Indeed, it suffices to check this for variables. In this case commutativity is shown by the following diagram:
\[
\begin{array}{ccc}
x_{i_{F_1} \setminus T}, \ldots, x_{i_{F_q} \setminus T}, k & \xrightarrow{\Phi_M} & \prod_{j=1}^q y_{j, i_{F_j} \setminus T}, k_j \\
\downarrow{\Psi} & & \downarrow{\Psi'} \\
x_{\psi_1(i_{F_1} \setminus T)}, \ldots, x_{\psi_q(i_{F_q} \setminus T)}, k & \xrightarrow{\Phi_{M'}} & \prod_{j=1}^q y_{j, \psi_j(i_{F_j} \setminus T)}, k_j
\end{array}
\]

Since \(\Psi\) and \(\Psi'\) are isomorphisms, the commutativity of diagram (3-3) implies that \(\text{im}(\Phi) \cong \text{im}(\Phi')\), which concludes the proof.

We also need the following result.

**Proposition 3.8.** Let \(\mathcal{I} = \{I_n\}_{n \in \mathbb{N}}\) be the \(S_G^q\)-invariant filtration corresponding to hierarchical models \(\mathcal{M}(\Delta, (c, n))\) with \(\Delta\) consisting of \(q\) 2-element disjoint facets \(F_1, \ldots, F_q\), each meeting \(T\) in exactly one vertex. Then the equivariant Hilbert series of \(\mathcal{I}\) is a rational function in \(s_1, \ldots, s_q, t\) with rational coefficients.

This will be shown in the following section. Assuming the result, we complete the argument for establishing Theorem 3.5.

**Proof of Theorem 3.5.** Let \(\nu\) be the number of facets in \(\Delta\) whose intersection with \(T\) is empty. We use induction on \(\nu \geq 0\). If \(\nu = 0\), the claimed rationality follows by combining Lemma 3.7 and Proposition 3.8.

Let \(\nu \geq 1\). We may assume that \(F_1 \cap T = \emptyset\) and that vertex 1 is in \(F_1\). By assumption, it has \(c_1\) states. Set \(\tilde{n} = (n_1, n), \ c = (c_2, \ldots, c_{#T})\) and \(\tilde{T} = T \cup \{1\}\). Then the hierarchical models \(\tilde{\mathcal{M}}(\Delta, (\tilde{c}, \tilde{n}))\) give rise to a
filtration $\mathcal{F} = \mathcal{F}_\Delta, r_{(n)}$, $\mathcal{F}$. By induction on $n$, it has a rational equivariant Hilbert series. By definition, it is

$$\text{equivH}_\mathcal{F}(s_1, s_2, \ldots, s_{q-v+1}, t) = \sum_{n_1 \geq 1} s_1^{n_1} \cdot \text{equivH}_\mathcal{F}(s_2, \ldots, s_{q-v+1}, t).$$

Hence $\text{equivH}_\mathcal{F}$ is obtained by evaluating $(1/c_1!) (\partial^{c_1} \text{equivH}_\mathcal{F} / \partial s_1^{c_1})$ at $s_1 = 0$. It follows that also $\text{equivH}_\mathcal{F}$ is rational. \hfill \Box

4. Regular languages

The goal of this section is to establish Proposition 3.8. We adopt its notation.

Fix $c \in \mathbb{N}^q$. As above, we write $x_{i,k}$ for $x_{i_1, i_2, \ldots, i_q, k_1, k_2, \ldots, k_q}$, where $(i, k) = (i_1, \ldots, i_q, k_1, \ldots, k_q) \in [c] \times [n] \subset \mathbb{N}^{2q}$. Thus, $y_{j, i, j, k_j}$ is simply $y_{j, i, j}$. For any $n \in \mathbb{N}^q$, the homomorphism associated to the model $M_n = M(\Delta, (c, n))$ is

$$\Phi_n : R_n = \mathbb{K}[x_{i,k} | (i, k) \in [c] \times [n]] \longrightarrow \mathbb{K}[y_{j, i, j, k_j} | j \in [q], i_j \in [c_j], k_j \in [n_j]] = S_n.$$

Set

$$A_n = \text{im} \Phi_n = \mathbb{K}\left[\prod_{j=1}^{q} y_{j, i, j, k_j} | i_j \in [c_j], k_j \in [n_j]\right].$$

We denote the sets of monomials in $A_n$ and $S_n$ by $\text{Mon}(A_n)$ and $\text{Mon}(S_n)$, respectively. Define $\text{Mon}(A)$ as the disjoint union of the sets $\text{Mon}(A_n)$ with $n \in \mathbb{N}^q$ and similarly define $\text{Mon}(S)$, where $S = \mathbb{K}[y_{j, i, j} | j \in [q], i_j \in [c_j], k \in \mathbb{N}]$. Our next goal is to show that the elements of $\text{Mon}(A)$ are in bijection to the words of a suitable formal language.

Consider a set

$$\Sigma = \{\xi_i, \tau_j | i \in [c], j \in [q]\}$$

with $q + \prod_{j=1}^{q} c_j$ elements. Let $\Sigma^*$ be the free monoid on $\Sigma$. A formal language with words in the alphabet $\Sigma$ is a subset of $\Sigma^*$. We refer to the elements of $\Sigma$ as letters. The empty word is denoted by $\varepsilon$.

In order to compare subsets of $\Sigma^*$ with $\text{Mon}(A)$ we need suitable maps. For $j \in [q]$, define a shift operator $T_j : \text{Mon}(S) \rightarrow \text{Mon}(S)$ by

$$T_j(y_{l,i,k}) = \begin{cases} y_{l,i,k+1} & \text{if } l = j, \\ y_{l,i,k} & \text{if } l \neq j, \end{cases}$$

extended multiplicatively to $\text{Mon}(S)$. Define a map $m : \Sigma^* \rightarrow \text{Mon}(S)$ inductively using the three rules

(a) $m(\varepsilon) = 1$,

(b) $m(\xi_i w) = \prod_{j=1}^{q} y_{j,i,j} m(w)$,

(c) $m(\tau_j w) = T_j(m(w))$,

where $w \in \Sigma^*$. In particular, this gives $m(\xi_i) = \Phi_n(x_{i,1})$ for any $n \in \mathbb{N}^q$, where $1$ is the $q$-tuple whose entries are all equal to $1$. 

Example 4.1. If $c_1 = c_2 = q = 2$, one has $\Sigma = \{\xi_{1,1}, \xi_{1,2}, \xi_{2,1}, \xi_{2,2}, \tau_1, \tau_2\}$, and, for any $n \geq (2, 3),$
\[
m(\tau_1 \tau_2 \xi_{1,2} \tau_2 \xi_{1,1} \tau_1) = T_1(T_2(y_{1,1,1}y_{2,2,1}T_2(y_{1,1,1}y_{2,1,1}T_1(1))))
\]
\[
= T_1(T_2(y_{1,1,1}y_{2,2,1}y_{1,1,1}y_{2,1,2}))
\]
\[
= y_{1,1,2}y_{2,2,2}y_{1,1,2}y_{2,1,3}
\]
\[
= \Phi_n(x_{(1,2),(2,2)})\Phi_n(x_{(1,1),(2,3)}).
\]

The map $m$ is certainly not injective because the variables $y_{j,i,k}$ commute. For example, if $q = 2$ one has $m(\tau_1 \tau_2) = m(\tau_2 \tau_1)$ and $m(\xi_{2,1} \xi_{1,2}) = m(\xi_{1,2} \xi_{2,1})$ and $m(\tau_1 \xi_{1,2} \tau_2 \xi_{2,1}) = m(\tau_1 \xi_{2,2} \tau_2 \xi_{1,1})$. Thus, we introduce a suitable subset of $\Sigma^*$.

Definition 4.2. Let $\mathcal{L}$ be the set of words in $\Sigma^*$ that satisfy the following conditions:

1. Every substring $\tau_i \tau_j$ has $i \leq j$.
2. In every substring with no $\tau_j$, if $\xi_i$ occurs to the left of some $\xi_i'$, then the $j$-th entry of $i$ is less than or equal to the $j$-th entry of $i'$.

To avoid triple subscripts below, we denote the $j$-th entry of a $q$-tuple $k_i$ by $k_{(l,j)}$; that is, we write
\[
k_l = (k_{(1,1)}, k_{(1,2)}, \ldots, k_{(l,q)}) \in \mathbb{N}^q.
\]

Using multi-indices, we write $\tau^a$ for $\tau_1^{a_1} \tau_2^{a_2} \cdots \tau_q^{a_q}$ with $a = (a_1, a_2, \ldots, a_q)$. A string consisting only of $\tau$-letters can be written as $\tau^k$ if and only if it satisfies condition (1) in Definition 4.2. With this notation, one gets immediately the following explicit description of the words in $\mathcal{L}$.

Lemma 4.3. The elements of the formal language $\mathcal{L}$ are precisely the words of the form
\[
\tau^{k_1} \xi_{i_1}^{k_2} \xi_{i_2}^{k_3} \cdots \xi_{i_d}^{k_{d+1}}
\]
where $i_1, \ldots, i_d \in \mathbb{Z}$, $k_1, \ldots, k_{d+1} \in \mathbb{N}_{0}^q$, and $i_{(l-1,j)} \leq i_{(l,j)}$ whenever $k_{(l,j)} = 0$ for some $(l, j)$ with $2 \leq l \leq d$ and $j \in [q]$.

The following elementary observation is useful.

Lemma 4.4. Every monomial in $\text{Mon}(A)$ can be uniquely written as a string of variables such that one has the variable in any position $l$ is of the form $y_{j,i,k}$ with $j = l \mod q$ and, for each $j \in [q]$, if a variable $y_{j,i_1,k_j}$ appears to the left of $y_{j,i_2,k_j}'$ then either $k_j < k_j'$ or $k_j = k_j'$ and $i_j \leq i_j'$.

Proof. If for some $j$, two variables $y_{j,i_1,k_j}$ and $y_{j,i_2,k_j}'$ appearing in a monomial do not satisfy the stated condition, then swap their positions. Repeating this step as long as needed results in a string meeting the requirement. It is unique, because the given condition induces an order on the variables $y_{j,i,k}$ with fixed $j$. In the desired string, for each fixed $j$, the variables $y_{j,i,k}$ occur in this order when one reads the string from left to right.

We illustrate the above argument.
Example 4.5. Let \( q = 2 \). To simplify notation write \( y_{jk} \) instead of \( y_{1,j,k} \) and \( z_{jk} \) instead of \( y_{2,j,k} \). Then one gets, for example,

\[
\begin{align*}
y_{22}z_{21}y_{14}z_{11}y_{31}z_{21} & \quad y_{22}y_{14}y_{31} \quad y_{31}y_{22}y_{14} \quad y_{31}z_{11}y_{22}z_{21}y_{14}z_{21}.
\end{align*}
\]

We observed above that the map \( m \) sends each letter \( \zeta_i \) to the monomial \( \Phi_n(x_{i+1}) \). It follows that \( m(\Sigma^*) \) is a subset of \( \text{Mon}(A) \). In fact, one has the following result.

**Proposition 4.6.** For any \( n \in \mathbb{N}_0^q \), denote by \( \mathcal{L}_n \) the set of words in \( \mathcal{L} \) in which, for each \( j \in [q] \), the letter \( \tau_j \) occurs precisely \( n_j \) times. Then \( m \) induces for every \( n \in \mathbb{N}_0^q \) a bijection

\[
m_n : \mathcal{L}_n \to \text{Mon}(A_{n+1}), \quad w \mapsto m(w).
\]

**Proof.** The definition of \( m \) readily implies \( m(w) \in \text{Mon}(A_{n+1}) \) if \( w \in \mathcal{L}_n \).

First we show that \( m_n \) is surjective. Let \( m \in \text{Mon}(A_{n+1}) \) be any monomial. Its degree is \( dq \) for some \( d \in \mathbb{N}_0 \). By Lemma 4.4, \( m \) can be written as

\[
m = \prod_{l=1}^d \left( \prod_{j=1}^q y_{j,i(l,j),k(l,j)} \right) = \prod_{l=1}^d \Phi_n(x_{i_l,k_l})
\]

such that, for each \( j \in [q] \), one has

\[
1 \leq k(1,j) \leq \cdots \leq k(d,j) \leq n_j + 1, \\
i(l-1,j) \leq i(l,j) \quad \text{if} \; k(l,j) = 0 \; \text{for some} \; l.
\]

The first condition implies that all the \( q \)-tuples \( k_1 - 1, k_2 - k_1, \ldots, k_d - k_{d-1} \) and \( n + 1 - k_d \) are in \( \mathbb{N}_0^q \). Hence the string

\[
w = \tau^{k_1 - 1} \xi_1 \tau^{k_2 - k_1} \xi_2 \cdots \tau^{k_d - k_{d-1}} \xi_d \tau^{n+1-k_d}
\]

is defined. The two conditions together combined with Lemma 4.3 show that in fact \( m \) is in \( \mathcal{L}_n \). Hence \( m(w) = m \) proves the claimed surjectivity.

Second, we establish that \( m_n \) is injective. Consider any two words \( w, w' \in \mathcal{L}_n \) with \( m(w) = m(w') \).

We will show \( w = w' \).

Write \( w \) and \( w' \) as in Lemma 4.3:

\[
w = \tau^{k_1} \xi_1 \tau^{k_2} \xi_2 \cdots \tau^{k_d} \xi_d \tau^{k_{d+1}}, \quad w' = \tau^{k'_1} \xi_1' \tau^{k'_2} \xi_2' \cdots \tau^{k'_{d+1}} \xi_d' \tau^{k'_{d+1}}.
\]

Since \( m(w) \) has degree \( dq \) and \( m(w') \) has degree \( d'q \), we conclude \( d = d' \). Evaluating \( m \) we obtain

\[
\prod_{l=1}^d \left( \prod_{j=1}^q y_{j,i(l,j),k(l,j)} \right) = \prod_{l=1}^d \left( \prod_{j=1}^q y_{j,i(l,j),k'(l,j)} \right),
\]

(4-1)

where \( f(l,j) = k(1,j) + \cdots + k(l,j) + 1 \) and \( f'(l,j) = k'_(1,j) + \cdots + k'_(l,j) + 1 \). Fix any \( j \in [q] \). Comparing the third indices of the variables whose first index equals \( j \) and using that every index is nonnegative, we get
for each \( l \in [d] \)

\[ k_{(1,j)} + \cdots + k_{(l,j)} = k'_{(1,j)} + \cdots + k'_{(l,j)}. \]

It follows that \( k_l = k'_l \) for each \( l \in [d] \). Since \( w \) and \( w' \) are in \( L_n \), we have \( k_{d+1} = n - (k_1 + k_2 + \cdots + k_d) \) and an analogous equation for \( k'_{d+1} \), which gives \( k_{d+1} = k'_{d+1} \).

It remains to show \( i_l = i'_l \) for every \( l \in [d] \). Fix any \( j \in [q] \). If for some \( l \) there is only one variable of the form \( y_{j, \mu, f_{(l,j)}} \) with \( \mu \in [c_j] \) that divides \( m(w) \), this implies \( i_{(l,j)} = i'_{(l,j)} = \mu \), as desired. Otherwise, there is a maximal interval of consecutive indices \( k_{(l,j)} \) that are equal to zero; that is, there are integers \( a, b \) such that \( 1 \leq a \leq b \leq d \) and

- \( k_{(l,j)} = 0 \) if \( a \leq l \leq b \),
- \( k_{(a-1,j)} > 0 \), unless \( a = 1 \), and
- \( k_{(b+1,j)} > 0 \), unless \( b = d \).

Thus, the number of variables of the form \( y_{j, \mu, f_{(l,j)}} \) that divide \( m(w) \) is \( b - a + 2 \) if \( a \geq 2 \) and \( b - a + 1 \) if \( a = 1 \). Considering these variables, Lemma 4.3 gives

\[ i_{(a-1,j)} \leq i_{(a,j)} \leq \cdots \leq i_{(b,j)} \quad \text{and} \quad i'_{(a-1,j)} \leq i'_{(a,j)} \leq \cdots \leq i'_{(b,j)}, \]

where \( i_{(a-1,j)} \) and \( i'_{(a-1,j)} \) are omitted if \( a = 1 \). Using (4-1), it now follows that \( i_{(l,j)} = i'_{(l,j)} \) whenever \( a - 1 \leq l \leq b \), unless \( a = 1 \). If \( a = 1 \), the latter equality is true whenever \( a \leq l \leq b \).

Applying the latter argument to any interval of consecutive zero indices \( k_{(l,j)} \), we conclude \( i_{(l,j)} = i'_{(l,j)} \) for every \( l \in [d] \). This completes the argument. \( \square \)

Our next goal is to show that \( L \) is a regular language. By [10, Theorems 3.4 and 3.7], this is equivalent to proving that \( L \) is recognizable by a finite automaton. Recall that a deterministic finite automaton on an alphabet \( \Sigma \) is a 5-tuple \( A = (P, \Sigma, \delta, P_0, F) \) consisting of a finite set \( P \) of states, an initial state \( P_0 \in P \), a set \( F \subset P \) of accepting states and a transition map \( \delta : D \to P \), where \( D \) is some subset of \( P \times \Sigma \). We refer to \( A \) simply as a finite automaton because we will consider only deterministic automata.

The automaton \( A \) recognizes or accepts a word \( w = a_1 a_2 \cdots a_s \in \Sigma^* \) if there is a sequence of states \( r_0, r_1, \ldots, r_s \) satisfying \( r_0 = P_0, r_s \in F \) and

\[ r_{j+1} = \delta(r_j, a_{j+1}) \quad \text{whenever} \quad 0 \leq j < s. \]

In words, the automaton starts in state \( P_0 \) and transitions from state \( r_j \) to a state \( r_{j+1} \) based on the input \( a_{j+1} \). The word \( w \) is accepted if \( r_s \) is an accepting state. If \( \delta(p, a) \) is not defined the machine halts.

The automaton \( A \) recognizes a formal language \( L \subset \Sigma^* \) if \( L \) is precisely the set of words in \( \Sigma^* \) that are accepted by \( A \).

Returning to the formal language \( L \) specified in Definition 4.2, we are ready to show:

**Proposition 4.7.** The language \( L \) is recognized by a finite automaton.

**Proof.** We need some further notation. We say that a sequence \( C \) of \( l \geq 0 \) integers \( j_1, j_2, \ldots, j_l \) is an increasing chain in \( [q] \) if \( 1 \leq j_1 < j_2 < \cdots < j_l \leq q \). Define \( \text{max}(C) \) as the largest element \( j_l \) of \( C \). We put \( \text{max}(\emptyset) = 0 \). We denote the set of increasing chains in \( [q] \) by \( C \). Thus, the cardinality of \( C \) is \( 2^q \). We
write $j \in C$ if $j$ occurs in the chain $C$. For any $k \in \mathbb{N}_0$, we define the sequence of indices $j$ with $k_j > 0$ as its support $\text{Supp}(k)$. It is an element of $C$. For example, one has $\text{Supp}(7, 0, 1, 5, 0) = (1, 3, 4)$.

Now we define an automaton $A$ as follows: Let

$$P = \{p_j, p_i, p_{i,C,k} \mid 0 \leq j \leq q, \ i \in [e], \ C \in \mathcal{C}, \ k \in C\}$$

be the set of states, where $p_0$ is the initial state of $A$. Let

$$F = \{p_j, p_i, p_{i,C,k} \mid 0 \leq j \leq q, \ i \in [e], \ C \in \mathcal{C}, \ k = \max(C)\}$$

be the set of accepting states. Furthermore, define transitions

$$\begin{align*}
\delta(p_j, \tau_{j'}) &= p_{j'} & \text{if } j = 0 < j' \leq q \text{ or } 1 \leq j \leq j' \leq q, \\
\delta(p_j, \zeta_i) &= p_i & \text{if } 0 \leq j \leq q, \ i \in [e], \\
\delta(p_{i,C}, \tau_j) &= p_{i,C,j} & \text{if } i \in [e], \ C \in \mathcal{C}, \ j \in C, \\
\delta(p_{i,k}, \zeta_i) &= p_i & \text{if } i, i' \in [e], \ i \leq i', \\
\delta(p_{i,C,k}, \tau_k) &= p_{i,C,k} & \text{if } i \in [e], \ C \in \mathcal{C}, \ j \in C, \ k \text{ directly follows } j \text{ in } C \text{ or } k = j, \\
\delta(p_{i,C,k}, \zeta_i) &= p_i & \text{if } i, i' \in [e], \ j = \max(C), \ i_k \leq i_k' \text{ whenever } k \notin C.
\end{align*}$$

(4-2) (4-3) (4-4) (4-5) (4-6) (4-7)

If an element of $P \times \Sigma$ does not satisfy any of the above six conditions then it is not in the domain of $\delta$.

We claim that $A$ recognizes $L$. Indeed, let $w \in \Sigma^*$ be a word with exactly $d \geq 0$ $\zeta$-letters. We show by induction on $d$ that $w$ is recognized by $A$ if $w \in L$, but any word in $\Sigma^* \setminus L$ is not accepted by $A$. It turns out that $w \in L$ is accepted

- at a state $p_j$ for some $0 \leq j \leq q$ if $d = 0$,
- at a state $p_i$ for some $i \in [e]$ if $d \geq 1$ and $w$ ends with a $\zeta$-letter, and
- at a state $p_{i,C,j}$ for some $i \in [e], \ C \in \mathcal{C}, \ j = \max(C)$ if $d \geq 1$ and $w$ ends with a $\tau$-letter.

In particular, this explains the set of accepting states.

Consider any word $w \in \Sigma^*$ with exactly $d \geq 0$ $\zeta$-letters. Assume $d = 0$, that is, $w = \tau_{l_1} \tau_{l_2} \cdots \tau_{l_i}$. By transition rule (4-2), $A$ transitions from state $p_0$ to any state $p_j$ with $j \in [q]$ using input $\tau_j$. From any $p_j$ with $j \in [q]$ the automaton can transition to any state $p_{j'}$ with $j \leq j' \leq q$ by using input $\tau_{j'}$. Thus, $w$ is accepted by $A$ if and only if $l_1 \leq l_2 \leq \cdots \leq l_i$, that is, $w \in L$ (see Lemma 4.3).

Assume now that $d \geq 1$. We proceed in several steps.

(I) Assume $d = 1$ and $w$ ends with a $\zeta$-letter, that is,

$$w = \tau_{l_1} \tau_{l_2} \cdots \tau_{l_i} \zeta_i$$

for some $i \geq 0$. The argument for $d = 0$ shows that $\tau_{l_1} \tau_{l_2} \cdots \tau_{l_i}$ is accepted if and only if it can be written as some $\tau^k$. Processing input $\tau^k$, the automaton arrives at state $p_j$ with $j = \max(\text{Supp}(k))$. Using input $\zeta_i$, it then transitions to $p_i \in F$ by rule (4-3). Hence $w$ is accepted if and only if $w \in L$.

(II) Let $d \geq 1$ and assume $w$ ends with a $\tau$-letter, that is, $w$ can be written as

$$w = \zeta_i \tau_{l_1} \tau_{l_2} \cdots \tau_{l_i},$$
with \( t \geq 1 \). Furthermore assume that \( w' \zeta_i \) is accepted by \( \mathcal{A} \) in state \( p_i \). We show that \( w \) is accepted by \( \mathcal{A} \) if and only if \( w = w' \zeta_i \tau^k \) for some \( k \in \mathbb{N}_0 \). If \( w \) is recognized, it is accepted in state \( p_{i,C,\text{max}(C)} \), where \( C = \text{Supp}(k) \).

Indeed, let \( C \in \mathcal{C} \) be the chain corresponding to the set \( \{l_1, \ldots, l_i\} \). Processing input \( \tau_l \), rule (4-3) yields that \( \mathcal{A} \) transitions to state \( p_{i,C,l_i} \). If \( t = 1 \), then \( l_1 = \text{max}(C) \) and \( w \) is accepted in \( p_{i,C,l_i} \in F \), as claimed. If \( t \geq 2 \), rule (4-6) shows that \( \mathcal{A} \) can transition from \( p_{i,C,l_1} \) using input \( \tau_l \) precisely if \( l_2 \geq l_1 \). If transition is possible, \( \mathcal{A} \) gets to state \( p_{i,C,l_2} \). Hence rule (4-6) guarantees that \( \tau_l \tau_l \cdots \tau_l = \tau^k \) for some nonzero \( k \in \mathbb{N}_0 \). In this case \( w = w' \zeta_i \tau^k \) is accepted by \( \mathcal{A} \) in state \( p_{i,C,\text{max}(C)} \), where \( C = \text{Supp}(k) \).

(III) Assume now \( w \in \Sigma^* \) ends with a \( \zeta \)-letter; that is, \( w \) is of the form

\[
w = w' \tau_l \tau_l \cdots \tau_l \zeta_i,
\]

where \( w' \in \mathcal{L} \) is either empty or ends with a \( \zeta \)-letter and \( t \geq 0 \). We show by induction on \( d \geq 1 \) that \( w \) is recognized by \( \mathcal{A} \) if and only if \( w \in \mathcal{L} \). In this case, \( w \) is accepted in a state \( p_i \).

Indeed, if \( d = 1 \), i.e., \( w' \) is the empty word, this has been shown in step (I). If \( d \geq 2 \) write \( w' = w'' \zeta_i' \). If \( w' \) is not accepted by \( \mathcal{A} \), then neither is \( w \). Furthermore, the induction hypothesis gives \( w' \notin \mathcal{L} \), which implies \( w \notin \mathcal{L} \).

If \( w' = w'' \zeta_i' \) is recognized by \( \mathcal{A} \) the induction hypothesis yields \( w' \in \mathcal{L} \) and \( w' \) is accepted in state \( p_i \).

Step (II) shows that \( w'' \zeta_i \tau_l \tau_l \cdots \tau_l \) is accepted by \( \mathcal{A} \) if and only if it can be written as \( w'' \zeta_i \tau^k \) for some \( k \in \mathbb{N}_0 \), and so

\[
w = w'' \zeta_i \tau^k \zeta_i.
\]

We consider two cases.

Case 1: Suppose \( k \) is zero, i.e., \( \text{Supp}(k) = \emptyset \). Thus, \( \mathcal{A} \) accepts \( w'' \zeta_i \in \mathcal{L} \) in state \( p_i \). Using input \( \zeta_i \), rule (4-5) shows that \( \mathcal{A} \) does not halt in \( p_i \) if and only if \( i' \leq i \). By Lemma 4.3, this is equivalent to \( w = w'' \zeta_i \zeta_i \in \mathcal{L} \). Furthermore, if \( w \) is in \( \mathcal{L} \) it is accepted in state \( p_i \), as claimed.

Case 2: Suppose \( \text{Supp}(k) \neq \emptyset \). Set \( C = \text{Supp}(k) \). By step (II), \( w'' \zeta_i \tau^k \) is accepted in state \( p_{i',C,j} \), where \( j = \text{max}(C) \). Hence rule (4-7) gives that input \( \zeta_i \) can be processed if and only if \( i'_l \leq i_l \) whenever \( l \notin C \). By Lemma 4.3, this is equivalent to \( w = w'' \zeta_i \tau^k \zeta_i \in \mathcal{L} \). Moreover, if \( w \) is recognized it is accepted in state \( p_i \), as claimed.

(IV) By steps (I) and (III) it remains to consider the case where \( w \) ends with a \( \tau \)-letter, i.e., \( w = w' \zeta_i \tau_l \tau_l \cdots \tau_l \) with \( t \geq 1 \). By step (III), \( w' \zeta_i \) is recognized by \( \mathcal{A} \) if and only if \( w' \zeta_i \in \mathcal{L} \). Furthermore, if \( w' \zeta_i \in \mathcal{L} \) then it is accepted in state \( p_i \). Hence, the assumption in step (II) is satisfied and we conclude that \( w \) is accepted if and only if \( w = w' \zeta_i \tau^k \). The latter is equivalent to \( w' \zeta_i \tau^k \in \mathcal{L} \) because \( w' \zeta_i \) is in \( \mathcal{L} \). This completes the argument.

Remark 4.8. Any finite automaton \( \mathcal{A} = (P, \Sigma, \delta, p_0, F) \) can be represented by a labeled directed graph whose vertex set is the set of states \( P \). Accepting states are indicated by double circles. There is an edge from vertex \( p \) to vertex \( p' \) if there is a transition \( \delta(p, a) = p' \). In that case, the edge is labeled by all \( a \in \Sigma \) such that \( \delta(p, a) = p' \).
Figure 1. The automaton for $c = (1, 1, 1)$ and $T = [3]$.

We illustrate the automata constructed in Proposition 4.7 using such a graphical representation.

Example 4.9. Let $A$ be the automaton constructed in Proposition 4.7 if $q = 3$ and $c = (1, 1, 1)$. Note the only element in $[c]$ is $1 = (1, 1, 1)$. To simplify notation, we write $\zeta$ for $\zeta_{1,1,1}$ and $p_1$ for $p_{(1,1,1)}$. We denote the nonempty increasing chains in the interval $[3]$ by $C_1 = \{1\}$, $C_2 = \{2\}$, $C_3 = \{3\}$, $C_4 = \{1, 2\}$, $C_5 = \{1, 3\}$, $C_6 = \{2, 3\}$, $C_7 = \{1, 2, 3\}$ and write $p_{i,j}$ instead of $p_{1,C_{i,j}}$. Using this notation, the constructed automaton $A$ is represented by the graph in Figure 1.

Remark 4.10. The automaton constructed in Proposition 4.7 is often not the smallest automaton that recognizes the language $L$. Using the minimization technique described in [10, Theorem 4.26], one can obtain an automaton with fewer states that also recognizes $L$. For example, if $c = (1, 1, 1)$, this produces an automaton with only four states, shown in Figure 2.

In order to relate a language $L$ on an alphabet $\Sigma$ to a Hilbert series we need a suitable weight function. Let $T = \mathbb{K}[s_1, \ldots, s_k]$ be a polynomial ring in $k$ variables and denote by $\text{Mon}(T)$ the set of monomials in $T$. A weight function is a monoid homomorphism $\rho : \Sigma^* \to \text{Mon}(T)$ such that $\rho(\varepsilon) = 1$ only if $\varepsilon$ is the empty word. The corresponding generating function is a formal power series in variables $s_1, \ldots, s_k$:

$$P_{L,\rho}(s_1, \ldots, s_k) = \sum_{w \in L} \rho(w).$$
We will use the following result; see, e.g., [9] or [16, Theorem 4.7.2].

**Theorem 4.11.** If \( \rho \) is any weight function on a regular language \( \mathcal{L} \) then \( P_{\mathcal{L}, \rho} \) is a rational function in \( \mathbb{Q}(s_1, \ldots, s_k) \).

We are ready to establish the ingredient of the proof of Theorem 3.5 whose proof we had postponed.

**Proof of Proposition 3.8.** Since \( I_n = \ker \Phi_n \) and \( \Phi_n \) is a homomorphism of degree \( q \), we get \( R_n/I_n \cong A_n \) and, for each \( d \in \mathbb{Z} \),

\[
\dim_{\mathbb{K}}[R_n/I_n]_d = \dim_{\mathbb{K}}[A_n]_{dq}.
\]

Recall that the algebra \( A_n \) is generated by monomials. Hence, every graded component has a \( \mathbb{K} \)-basis consisting of monomials. It follows that \( \dim_{\mathbb{K}}[A_n]_{dq} = \# \text{Mon}([A_n]_{dq}) \). Therefore we get for the equivariant Hilbert series of the filtration \( \mathcal{F} \)

\[
equivH_{\mathcal{F}}(s_1, \ldots, s_q, t) = \sum_{n \in \mathbb{N}^q} \sum_{d \geq 0} \# \text{Mon}([A_n]_{dq}) \cdot s^n t^d,
\]

where \( s^n = s_1^{n_1} \cdots s_q^{n_q} \) if \( n = (n_1, \ldots, n_q) \).

Consider now the language \( \mathcal{L} \) described in Definition 4.2. Define a weight function \( \rho : \Sigma^* \rightarrow \text{Mon}(\mathbb{K}[s_1, \ldots, s_q, t]) \) by \( \rho(\tau_j) = s_j \) and \( \rho(\zeta_i) = t \) for \( i \in [c] \). Thus, for \( w \in \mathcal{L} \), one obtains \( \rho(w) = s^n t^d \) if \( d \) is the number of \( \zeta \)-letters occurring in \( w \) and \( n_j \) is the number of appearances of \( \tau_j \) in \( w \). Hence Proposition 4.6 gives that the number of words \( w \in \mathcal{L}_n \) with \( \rho(w) = s^n t^d \) is precisely \( \# \text{Mon}([A_{n+1}]_{dq}) \). Since \( \mathcal{L} \) is the disjoint union of all \( \mathcal{L}_n \), it follows

\[
s_1 \cdots s_q \cdot \equivH_{\mathcal{F}}(s_1, \ldots, s_q, t) = \sum_{n \in \mathbb{N}^q} \sum_{w \in \mathcal{L}_n} \rho(w) = P_{\mathcal{L}, \rho}(s_1, \ldots, s_q, t). \tag{4-8}
\]

As the right-hand side is rational by Theorem 4.11, the claim follows. \( \square \)

**Remark 4.12.** The method of proof for Theorem 3.5 is rather general and can also be used in other situations. An easy generalization is obtained as follows. Fix \( (a_1, \ldots, a_q) \in \mathbb{N}^q \). For \( n \in \mathbb{N}^q \), consider the
homomorphism

\[ \tilde{\Phi}_n : R_n = \mathbb{k}[x_{i,k} \mid (i,k) \in [c] \times [n]] \rightarrow \mathbb{k}[y_{j,i,k} \mid j \in [q], i_j \in [c_j], k_j \in [n_j]] = S_n, \]

\[ x_{i,k} \mapsto \prod_{j=1}^{q} y_{j,i_j,k_j}, \]

and set

\[ \tilde{A}_n = \text{im} \Phi_n = \mathbb{k}\left[ \prod_{j=1}^{q} y_{j,i_j,k_j} \mid i_j \in [c_j], k_j \in [n_j] \right], \]

\[ \tilde{I}_n = \ker \tilde{\Phi}_n. \]

Then \( \mathcal{I} = \{ \tilde{I}_n \}_{n \in \mathbb{N}} \) also is an \( \mathcal{S}^d_{\infty} \)-invariant filtration whose equivariant Hilbert series is rational. Indeed, this follows using the language \( \mathcal{L} \) as above with the following modifications. In the definition of the map \( m \) change rule (b) to \( \mathcal{m}(\zeta_i w) = \prod_{j=1}^{q} y_{j,i_j,1} \mathcal{m}(w) \), but keep rules (a) and (c) to obtain a map \( \mathcal{m} : \Sigma^* \rightarrow \text{Mon}(S) \). It induces bijections \( \mathcal{L}_n \rightarrow \text{Mon}(\tilde{A}_{n+1}) \) as in Proposition 4.6. Observe that \( [R_n/\tilde{I}_n]_{d} = [A_n]_{da} \), where \( a = a_1 + \cdots + a_q \). Thus, using the same weight function \( \rho \) as above, we obtain

\[ s_1 \cdots s_q \cdot \text{equivH}_{\mathcal{I}}(s_1, \ldots, s_q, t) = P_{A,\rho}(s_1, \ldots, s_q, t). \]

A systematic study of substantial generalizations will be presented in [13].

5. Explicit formulas

We provide explicit formulas for the Hilbert series of hierarchical models considered in Theorem 3.5.

It is useful to begin by discussing Segre products more generally. To this end we temporarily use some new notation.

**Lemma 5.1.** Let \( A = \mathbb{k}[a_1, \ldots, a_s] \subset R \) and \( B = \mathbb{k}[b_1, \ldots, b_t] \subset S \) be subalgebras of polynomial rings \( R = \mathbb{k}[x_1, \ldots, x_m] \) and \( S = \mathbb{k}[y_1, \ldots, y_n] \) that are generated by monomials \( a_1, \ldots, a_s \) of degree \( d_1 \) and monomials \( b_1, \ldots, b_t \) of degree \( d_2 \), respectively. Let \( C \) be the subalgebra of \( \mathbb{k}[x_1, \ldots, x_m, y_1, \ldots, y_n] \) that is generated by all monomials \( a_i b_j \) with \( i \in [s] \) and \( j \in [t] \). Using the gradings induced from the corresponding polynomials rings one has, for all \( k \in \mathbb{Z} \),

\[ \dim_{\mathbb{k}}[C]_{k(d_1+d_2)} = \dim_{\mathbb{k}}[A]_{kd_1} \cdot \dim_{\mathbb{k}}[B]_{kd_2}. \]

**Proof.** This follows from the fact that the nontrivial degree components of the algebras \( A, B, C \) have \( \mathbb{k} \)-bases generated by monomials in the respective algebra generators of suitable degrees. \( \square \)

It is customary to consider the algebras occurring in Lemma 5.1 as standard graded algebras that are generated in degree 1 by redefining their grading. In the new gradings, the degree \( k \) elements of \( A \) are elements that have degree \( kd_1 \), considered as polynomials in \( R \), and similarly the degree-\( k \) elements of \( C \) have degree \( k(d_1+d_2) \) when considered as elements of \( \mathbb{k}[x_1, \ldots, x_m, y_1, \ldots, y_n] \). Using this new grading, the statement in the above lemma reads

\[ \dim_{\mathbb{k}}[C]_{d} = \dim_{\mathbb{k}}[A]_{d} \cdot \dim_{\mathbb{k}}[B]_{d}. \]

(5-1)

This justifies calling \( C \) the Segre product of the algebras \( A \) and \( B \). We denote it by \( A \bowtie B \).
Iterating the above construction we get the following consequence.

**Corollary 5.2.** Let $A_1, \ldots, A_k$ be subalgebras of polynomial rings and assume every $A_i$ is generated by finitely many monomials of degree $d_i$. Regrade such that every $A_i$ is an algebra that is generated in degree 1. Then one has

$$\dim_K[A_1 \boxtimes \cdots \boxtimes A_k]_d = \prod_{i=1}^{k} \dim_K[A_i]_{d}. \,$$

We need an elementary observation.

**Lemma 5.3.** Let $\omega \in \mathbb{C}$ be a primitive $k$-th root of unity. If

$$f(t) = \sum_{n=0}^{\infty} c_n t^n$$

is a formal power series in $t$ with complex coefficients, then

$$\sum_{n=0}^{\infty} c_{kn} t^{kn} = \frac{1}{k} [f(t) + f(\omega t) + \cdots + f(\omega^{k-1}t)].$$

**Proof.** Using geometric sums one gets, for every $n \in \mathbb{N}_0$,

$$\sum_{j=0}^{k-1} (\omega^j)^n = \begin{cases} k & \text{if } k \text{ divides } n, \\ 0 & \text{otherwise}. \end{cases}$$

The claim follows. \hfill \Box

**Proposition 5.4.** Fix any $q \in \mathbb{N}$ and let $\mathcal{F}$ be the $S_q \infty$-invariant filtration considered in Proposition 3.8. For $j \in [q]$, let $\omega_j$ be a $c_j$-th primitive root of unity. Then the equivariant Hilbert series of $\mathcal{F}$ is

$$\text{equivH}_{\mathcal{F}}(s_1, \ldots, s_q, t) = \frac{1}{c_1 \cdots c_q} \sum_{m_1 \in [c_1], \ldots, m_q \in [c_q]} \frac{\omega_1^{m_1} / s_1^{1/c_1} \cdots \omega_q^{m_q} / s_q^{1/c_q}}{1 - \omega_1^{m_1} / s_1^{1/c_1} \cdots (1 - \omega_q^{m_q} / s_q^{1/c_q}) - t}. \,$$

**Proof.** By definition of the map $\Phi_{\mathcal{M}_n}$, its image is isomorphic to the Segre product of polynomial rings of dimension $c_jn_j$ with $j = 1, \ldots, q$. Hence Corollary 5.2 gives for the equivariant Hilbert series

$$\text{equivH}_{\mathcal{F}}(s_1, \ldots, s_q, t) = \sum_{d \geq 0, n \in \mathbb{N}^q} \frac{1}{d} \left( c_1n_1 + d - 1 \right) \cdots \left( c_qn_q + d - 1 \right) s_1^{n_1} \cdots s_q^{n_q} t^d$$

$$= \sum_{d \geq 0} \left\{ \prod_{j=1}^{q} \left[ \sum_{n_j \in \mathbb{N}} \left( c_jn_j + d - 1 \right) s_j^{n_j} \right] \right\} t^d. \quad (5-2)$$

For any integer $d \geq 0$, one computes

$$\sum_{n \in \mathbb{N}} \binom{n+d-1}{d} s^n = s \sum_{n \in \mathbb{N}_0} \binom{d+n}{n} s^n = \frac{s}{(1-s)^{d+1}}.$$
Combined with Lemma 5.3 and using a \( c \)-th primitive root of unity \( \omega \in \mathbb{C} \), we obtain, for any integer \( c > 0 \),
\[
\sum_{n \in \mathbb{N}} \left( \frac{cn+d-1}{d} \right) s^n = \frac{1}{c} \sum_{m \in [c]} \frac{\omega^m s^{1/c}}{(1 - \omega^m s^{1/c})^{d+1}}.
\]
Applying the last formula to the inner sums in (5-2) we get
\[
equivH_{\rho}(s_1, \ldots, s_q, t) = \sum_{d \geq 0} \left\{ \prod_{j=1}^{q} \left[ \frac{1}{c_j} \frac{\omega_j^{m_j} s_j^{1/c_j}}{(1 - \omega_j^{m_j} s_j^{1/c_j})^{d+1}} \right] \right\} t^d
\]
\[
= \sum_{d \geq 0} \frac{1}{c_1 \cdots c_q} \left\{ \sum_{m_1 \in [c_1] \ldots m_q \in [c_q]} \frac{\omega_1^{m_1} s_1^{1/c_1} \cdots \omega_q^{m_q} s_q^{1/c_q}}{(1 - \omega_1^{m_1} s_1^{1/c_1})^{d+1} \cdots (1 - \omega_q^{m_q} s_q^{1/c_q})^{d+1}} \right\} t^d
\]
\[
= \frac{1}{c_1 \cdots c_q} \sum_{m_1 \in [c_1] \ldots m_q \in [c_q]} \frac{\omega_1^{m_1} s_1^{1/c_1} \cdots \omega_q^{m_q} s_q^{1/c_q}}{(1 - \omega_1^{m_1} s_1^{1/c_1}) \cdots (1 - \omega_q^{m_q} s_q^{1/c_q}) - t},
\]
as claimed. \( \square \)

By Theorem 3.5, the above formula for the equivariant Hilbert series can be rewritten as a rational function with rational coefficients.

**Example 5.5.** (i) Let \( c_1 = \cdots = c_q = 1 \). Then Proposition 5.4 gives
\[
equivH_{\rho}(s_1, s_2, \ldots, s_q, t) = \frac{s_1 \cdots s_q}{(1 - s_1) \cdots (1 - s_q) - t}.
\]
By the argument at the beginning of the proof of Lemma 3.7, this model has the same equivariant Hilbert series as the corresponding independence model (see Example 3.6).

(ii) Let \( q = c_1 = c_2 = 2 \). Then Proposition 5.4 yields
\[
4 \cdot \equivH_{\rho}(s_1, s_2, t) = \frac{\sqrt{s_1 s_2}}{(1 - \sqrt{s_1})(1 - \sqrt{s_2}) - t} - \frac{\sqrt{s_1 s_2}}{(1 - \sqrt{s_1})(1 + \sqrt{s_2}) - t} - \frac{\sqrt{s_1 s_2}}{(1 + \sqrt{s_1})(1 - \sqrt{s_2}) - t} + \frac{\sqrt{s_1 s_2}}{(1 + \sqrt{s_1})(1 + \sqrt{s_2}) - t},
\]
Now a straightforward computation gives
\[
equivH_{\rho}(s_1, s_2, t) = \frac{s_1 s_2(s_1 s_2 - s_1 - s_2 - t^2)}{f},
\]
where
\[
f = s_1 s_2(s_1 - 2)(s_2 - 2) + s_1(s_1 - 2) + s_2(s_2 - 2) - 2t^2(s_1 s_2 + s_1 + s_2) - 4t(s_1 s_2 - s_1 - s_2) + (1 - t)^4.
\]
There is an alternative method to determine the equivariant Hilbert series whose rationality is guaranteed by Proposition 3.8. It directly produces a rational function with rational coefficients. This approach applies to any equivariant Hilbert series that is equal to the generating function \( P_{L, \rho} \) determined by a weight function \( \rho \) on a regular language \( L \). Indeed, let \( A = (P, \Sigma, \delta, p_0, F) \) be a finite automaton that recognizes \( L \). Suppose \( P \) has \( N \) elements \( p_0, \ldots, p_{N-1} \). For every letter \( a \in \Sigma \) define a 0-1 matrix \( M_{A,a} \)
of size $N \times N$. Its entry at position $(i, j)$ is 1 precisely if there is a transition $\delta(p_j, a) = p_i$. Let $e_i \in \mathbb{K}^N$ be the canonical basis vector corresponding to state $p_i$. Let $\mathbf{u} = \sum_{p_{i-1} \in F} e_i \in \mathbb{K}^N$ be the sum of the basis vectors corresponding to the accepting states. Then, for any word $w = w_1 \cdots w_d$ with $w_i \in \Sigma$, one has

$$\mathbf{u}^T M_{A,w_d} \cdots A_{A,w_1} e_1 = \begin{cases} 1 & \text{if } A \text{ accepts } w, \\ 0 & \text{if } A \text{ rejects } w. \end{cases}$$

Let $\rho : \Sigma^* \to \text{Mon}(\mathbb{K}[s_1, \ldots, s_k])$ be a weight function. Thus, $\rho(w_1 w_2) = \rho(w_1) \cdot \rho(w_2)$ for any $w_1, w_2 \in \Sigma^*$. It follows (see, e.g., [16, Section 4.7])

$$P_{\mathcal{L},\rho}(s_1, \ldots, s_k) = \sum_{w \in \mathcal{L}} \rho(w)$$

$$\sum_{d \geq 0} \sum_{w_1, \ldots, w_d \in \Sigma} \mathbf{u}^T (\rho(w_1 \cdots w_d) M_{A,w_d} \cdots A_{A,w_1}) e_1$$

$$= \sum_{d \geq 0} \mathbf{u}^T \left( \sum_{a \in \Sigma} \rho(a) M_{A,a} \right)^d e_1 = \mathbf{u}^T \left( \text{id}_N - \sum_{a \in \Sigma} \rho(a) M_{A,a} \right)^{-1} e_1.$$

Thus, the generating function $P_{\mathcal{L},\rho}(s_1, \ldots, s_k)$ is rational with rational coefficients and can be explicitly computed from the automaton $A$ using linear algebra.

In the proof of Proposition 3.8, we showed (see (4-8)) that the equivariant Hilbert series of a considered filtration is, up to a degree shift, equal to a generating function. Hence, the above approach can be used to compute directly this Hilbert series as a rational function with rational coefficients. We implemented the resulting algorithm in Macaulay2 [6]. It is available at http://www.sites.google.com/view/aidamaraj/research.

**Example 5.6.** In Proposition 3.8, consider the case where $c = (1, 1, \ldots, 1) \in \mathbb{N}^q$. The automaton constructed in Proposition 4.7 can be reduced to one with only $q+1$ states (see Remark 4.10 if $q = 3$), shown in Figure 3.

![Figure 3](image-url)
Hence, listing $p_1$ as the last state, we obtain for the equivariant Hilbert series of the filtration $I$

\[
equivH_{\mathcal{I}}(s_1, \ldots, s_q, t) = s_1 s_2 \cdots s_q \cdot u^T \left( \id_{q+1} - \sum_{a \in \Sigma} \rho(a) M_{A_w} \right)^{-1} e_1
\]

\[
= s_1 s_2 \cdots s_q \begin{bmatrix}
1 & -s_1 & 1 - s_2 & 0 & \cdots & 0 & -s_1 \\
0 & 1 - s_2 & 0 & \cdots & 0 & 0 & -s_2 \\
-s_3 & 0 & 1 - s_3 & \cdots & 0 & 0 & -s_3 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
-s_{q-1} & -s_{q-1} & -s_{q-1} & \cdots & 1 - s_{q-1} & 0 & -s_{q-1} \\
-s_q & -s_q & -s_q & \cdots & -s_q & 1 - s_q & -s_q \\
-1 & -t & -1 & \cdots & -t & -t & 1 - t
\end{bmatrix}^{-1}
\]

\[
= \frac{s_1 \cdots s_q}{(1 - s_1) \cdots (1 - s_q) - t^q},
\]

where the first column of the inverse matrix can be determined using suitable minors. Of course, the result is the same as in Example 5.5.

References


Received 2019-11-04. Revised 2020-07-31. Accepted 2020-08-31.

**AIDA MARAJ:** aida.maraj@mis.mpg.de  
Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany

**UWE NAGEL:** uwe.nagel@uky.edu  
Department of Mathematics, University of Kentucky, Lexington, KY, United States
1-WASSERSTEIN DISTANCE ON THE STANDARD SIMPLEX

ANDREW FROHMADER AND HANS VOLKMER

Wasserstein distances provide a metric on a space of probability measures. We consider the space $\Omega$ of all probability measures on the finite set $\chi = \{1, \ldots, n\}$, where $n$ is a positive integer. The 1-Wasserstein distance, $W_1(\mu, \nu)$, is a function from $\Omega \times \Omega$ to $[0, \infty)$. This paper derives closed-form expressions for the first and second moments of $W_1$ on $\Omega \times \Omega$ assuming a uniform distribution on $\Omega \times \Omega$.

1. Introduction

Wasserstein distances provide a natural metric on a space of probability measures. Intuitively, they measure the minimum amount of work required to transform one distribution into another. They have been applied in numerous fields including computer vision [8], statistics [6], dynamical systems [10], and many others.

Let $(\chi, d)$ be a Polish space and $\mu$ and $\nu$ be two probability measures on $\chi$. The $p$-Wasserstein distance between $\mu$ and $\nu$ is defined by

$$W_p(\mu, \nu) = \left( \inf_{\pi \in \Pi(\mu, \nu)} \int_{\chi \times \chi} d(x, y)^p \, d\pi(x, y) \right)^{1/p},$$

where $\Pi(\mu, \nu)$ is the set of probability measures $\pi$ on $\chi \times \chi$ such that $\pi(A \times \chi) = \mu(A)$ and $\pi(\chi \times B) = \nu(B)$ for all measurable $A, B \subseteq \chi$. The measure $\pi$ with marginals $\mu$ and $\nu$ on the two components of $\chi \times \chi$ is called a coupling of $\mu$ and $\nu$.

In general, the Wasserstein distances do not admit closed-form expressions. One important exception is the case where $\chi = \mathbb{R}$, $d(x, y) = |x - y|$, and $p = 1$. In this case, we have the explicit formula [5]

$$W_1(\mu, \nu) = \int_\mathbb{R} |F_\mu(t) - F_\nu(t)| \, dt,$$

where $F_\mu(t)$ and $F_\nu(t)$ are the cumulative distribution functions of $\mu$ and $\nu$ respectively.

Here we specialize further, considering only $\chi = \{1, \ldots, n\}$, where $n$ is a positive integer. Denote this finite set by $[n]$ and define the ground metric $d(x, y) = |x - y|$ on $[n]$. Unless stated otherwise, we will assume this metric on $[n]$ without comment. Let $\mu$ and $\nu$ be two probability measures on $[n]$. Then $W_1(\mu, \nu)$ reduces to the $l_1$-distance between finite vectors [9; 1]:

$$W_1(\mu, \nu) = \sum_{i=1}^n |F_\mu(i) - F_\nu(i)| = \|F_\mu - F_\nu\|_1.$$

MSC2010: 28A33, 60B05.

Keywords: Wasserstein, earth mover, probability simplex, simplex, standard simplex, moments, expected value, variance.
There are numerous applications of Wasserstein distances on finite sets. In [1] Bourn and Willenbring, use \( W_1 \), which they refer to as the *earth mover’s distance*, or EMD, following the computer science convention, to compare grade distributions — distributions on the finite set \{A, A−, B+, . . . , F\}. In that context, Bourn and Willenbring considered the space of all probability distributions on \([n]\), denoted by \( \mathcal{P}_n \)— the *standard simplex* or *probability simplex*. This space embeds naturally in \( \mathbb{R}^n \), inherits Lebesgue measure, and has finite volume. A uniform probability measure is obtained by normalizing Lebesgue measure on \( \mathcal{P}_n \) so that total mass of \( \mathcal{P}_n \) is 1. Similarly, \( \mathcal{P}_n \times \mathcal{P}_n = \Omega_n \) with normalized Lebesgue measure be our probability space. Define the random variable \( X_n(\mu, \nu) = W_1(\mu, \nu) \) on \( \Omega_n \). What can we say about \( X_n \)?

The contribution of [1] to this question is the derivation of a recurrence \( M_{p,q} \), such that \( \mathbb{E}(X_n) = M_{n,n} \). The recurrence is
\[
M_{p,q} = \frac{(p - 1)M_{p-1,q} + (q - 1)M_{p,q-1} + |p - q|}{p + q - 1},
\]
with \( M_{p,q} = 0 \) if either \( p \) or \( q \) is not positive. In this paper, we find a closed form for the first and second moments of \( X_n \). Additionally, we “solve” the recurrence provided by [1] to verify consistency of results. This provides an interesting link between the discrete combinatorial approach of [1] and our calculus-based approach.

The uniform distribution on the probability simplex is a special case of the Dirichlet distribution. In future work, we hope to expand our results to the entire family of Dirichlet distributions. We deal primarily with the ground metric \( |x - y| \) on \([n]\) given by the natural embedding into \( \mathbb{Z} \). Another direction is to extend these results to a larger class of metrics on \([n]\) we consider this in Section 2.

This paper provides an avenue for deriving closed forms for the moments of \( X_n \). These same techniques may be useful for related questions. In particular, one could attempt to repeat our approach for an indexed family of statistical models. We have in mind a family \( \{\mathcal{V}_n\} \) where \( \mathcal{V}_n \) is an algebraic variety inside \( \mathcal{P}_n \). More generally, consider two indexed families of statistical models \( \mathcal{V}_n \) and \( \mathcal{W}_n \). What can be said about \( W_1(\mu, \nu) \), with \( \mu \in \mathcal{V}_n \) and \( \nu \in \mathcal{W}_n \)? These questions also relate to the recent work [3; 4] of Çelik et al., where the problem of determining the minimal Wasserstein distance from a given distribution \( \mu \) to a fixed statistical model \( \mathcal{V} \subset \mathcal{P}_n \) is taken up from the perspective of linear programming. In this context, an understanding of \( W_1 \) on the space \( \mathcal{P}_n \times \mathcal{P}_n \) or \( \mathcal{P}_n \times \mathcal{V} \) can provide information on the quality of the minimum distance estimator.

The remainder of this paper is organized as follows. Section 2 presents our results and provides some discussion. Section 3 proves the closed form for the first moment. Section 4 proves the closed form for the second moment. Section 5 solves the recurrence (3) of Bourn and Willenbring.

### 2. Main result and discussion

The main results of this paper are the following two theorems.

**Theorem 1.** The first moment of \( X_n \) is given by
\[
\mathbb{E}(X_n) = \frac{2^{2n-3}(n-1)}{(2n-1)!} (n-1)!^2.
\]
Table 1 presents approximate numeric values for the first and second moments, and the variance, for $X_n$ and $\tilde{X}_n$.

Table 1. Approximate numeric values for the first and second moments, and the variance, for $X_n$ and $\tilde{X}_n$.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\mathbb{E}(X_n)$</th>
<th>$\mathbb{E}(X_n^2)$</th>
<th>$\text{Var}(X_n)$</th>
<th>$\mathbb{E}(\tilde{X}_n)$</th>
<th>$\mathbb{E}(\tilde{X}_n^2)$</th>
<th>$\text{Var}(\tilde{X}_n)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.3333</td>
<td>0.1667</td>
<td>0.05556</td>
<td>0.3333</td>
<td>0.1667</td>
<td>0.05556</td>
</tr>
<tr>
<td>3</td>
<td>0.5333</td>
<td>0.3778</td>
<td>0.09333</td>
<td>0.2667</td>
<td>0.09444</td>
<td>0.02333</td>
</tr>
<tr>
<td>4</td>
<td>0.6857</td>
<td>0.6000</td>
<td>0.1298</td>
<td>0.2286</td>
<td>0.06667</td>
<td>0.01442</td>
</tr>
<tr>
<td>5</td>
<td>0.8127</td>
<td>0.8267</td>
<td>0.1662</td>
<td>0.2032</td>
<td>0.05167</td>
<td>0.01039</td>
</tr>
<tr>
<td>6</td>
<td>0.9235</td>
<td>1.0560</td>
<td>0.2027</td>
<td>0.1847</td>
<td>0.04222</td>
<td>0.008107</td>
</tr>
<tr>
<td>7</td>
<td>1.0230</td>
<td>1.2860</td>
<td>0.2392</td>
<td>0.1705</td>
<td>0.03571</td>
<td>0.006645</td>
</tr>
<tr>
<td>8</td>
<td>1.1140</td>
<td>1.5176</td>
<td>0.2759</td>
<td>0.1591</td>
<td>0.03095</td>
<td>0.005630</td>
</tr>
<tr>
<td>9</td>
<td>1.1980</td>
<td>1.7480</td>
<td>0.3126</td>
<td>0.1498</td>
<td>0.02731</td>
<td>0.004884</td>
</tr>
<tr>
<td>10</td>
<td>1.2770</td>
<td>1.9800</td>
<td>0.3493</td>
<td>0.1419</td>
<td>0.02444</td>
<td>0.004313</td>
</tr>
</tbody>
</table>

Theorem 2. The second moment of $X_n$ is given by

$$\mathbb{E}(X_n^2) = \frac{(n-1)(7n-4)}{30n}.$$

2.1. Changing the ground metric. Closed forms for $W_1$ are not available for an arbitrary metric on $[n]$. However, by using (1) we can obtain a closed form for any metric induced by an embedding of $[n]$ into $\mathbb{R}$. Here we generalize our result by considering the metric induced by the simple embedding $k \rightarrow \alpha k$, where $\alpha \in \mathbb{R}_{>0}$. More general metrics are left open for future work.

Let $d_\alpha(x, y)$ be the metric induced by this embedding and $X_{\alpha,n}$ be the corresponding random variable on $\Omega_n$. Then $d_\alpha(x, y) = |\alpha x - \alpha y| = \alpha \cdot d(x, y)$ and we easily find the closed forms for this family of ground metrics to be

$$\mathbb{E}(X_{\alpha,n}) = \alpha \frac{2^{2n-3}(n-1)}{(2n-1)!}(n-1)!^2, \quad \mathbb{E}(X_{\alpha,n}^2) = \alpha^2 \frac{(n-1)(7n-4)}{30n}.$$

Note, our original random variable $X_n$ is equal to $X_{1,n}$. In [1] Bourn and Willenbring define the unit normalized EMD on $\mathcal{P}_n$ as the EMD scaled by $\frac{1}{n-1}$ and extend their results to this case. In our notation, this just amounts to taking $\alpha = \frac{1}{n-1}$. We refer to this random variable as $\tilde{X}_n$. These two choices of ground metric have a simple interpretation in terms of embeddings into $\mathbb{R}$. Taking $\alpha = 1$ places the finite set $[n]$ on the first $n$ positive integers of $\mathbb{R}$, while $\alpha = \frac{1}{n-1}$ partitions the unit interval $[\frac{1}{n}, 1 + \frac{1}{n}]$ with $n$ equally spaced points.

2.2. Asymptotic behavior. We briefly consider the asymptotic behavior of the first and second moments. Table 1 presents approximate numeric values for the first and second moments, and the variance, for $X_n$ and $\tilde{X}_n$. Figure 1 plots $\mathbb{E}(\tilde{X}_n)$ and $\mathbb{E}(\tilde{X}_n^2)$ for $n = 2$ to $n = 50$. These graphs suggest $\mathbb{E}(\tilde{X}_n)$ and $\mathbb{E}(\tilde{X}_n^2)$ converge to 0 as $n \rightarrow \infty$. This is in fact true and follows from the more general case considered below.

We consider the behavior of $\mathbb{E}(X_{\alpha,n})$ as $n \rightarrow \infty$. As discussed above,

$$\mathbb{E}(X_{\alpha,n}) = \alpha \frac{2^{2n-3}(n-1)}{(2n-1)!}(n-1)!^2.$$
We can rewrite this in the form
\[ \mathbb{E}(X_{\alpha,n}) = \alpha 4^{n-1} \frac{(n-1)}{n} \frac{1}{\binom{2n}{n}}. \]

It is known that [2]
\[ \binom{2n}{n} \sim \frac{4^n}{\sqrt{\pi n}}. \]
The symbol \( \sim \) means the quotient of the left and right sides converges to 1 as \( n \to \infty \). Therefore,
\[ \mathbb{E}(X_{\alpha,n}) \sim \alpha \frac{\sqrt{\pi n}}{4}. \]

We also have
\[ \mathbb{E}(X_{\alpha,n}^2) = \alpha^2 \frac{(n-1)(7n-4)}{30n}. \]

Therefore, for the variance we obtain
\[ \text{Var}(X_n) = \mathbb{E}(X_n^2) - \mathbb{E}(X_n)^2 \sim \alpha^2 \left( \frac{7}{30} - \frac{\pi}{16} \right)n. \]
2.3. Probability measures for $X_n$. It would be nice to have explicit probability measures for $X_n$. Toward this end, we computed the probability density function for $X_2$.

$$p_{X_2}(t) = 2 - 2t, \quad 0 \leq t \leq 1,$$

and for $X_3$,

$$p_{X_3}(t) = \begin{cases} \frac{19}{3}t^3 - 14t^2 + 8t & \text{if } 0 \leq t \leq 1, \\ -\frac{1}{3}t^3 + 2t^2 - 4t + \frac{8}{3} & \text{if } 1 < t \leq 2. \end{cases}$$

Figure 2 displays $p_{X_3}(t)$. Further work could be done to determine probability measures for $n > 3$ although the number of cases required to derive these functions with our current approach appears to grow rapidly.

3. Proof of Theorem 1

In what follows, we are led to the evaluation of certain sums for which [7] is a general reference. We begin with the probability space $(\Omega_n, B(\Omega_n), \phi)$, where $B(\Omega_n)$ denotes the Borel $\sigma$-algebra on $\Omega_n$ and $\phi$ is the Lebesgue measure normalized such that $\phi(\Omega_n) = 1$. Our random variable is $X_n(\mu, \nu) = W_1(\mu, \nu)$. It is easy to write the first moment of $X_n$ as an integral over $\Omega_n$:

$$\mathbb{E}(X_n) = \int_{\Omega_n} X_n(\mu, \nu) \, d\phi.$$
If we set \( F_1(s, t) = 0 \), then the functions \( F_n \) are determined recursively by

\[
F_n(s, t) = \int_0^t \int_y^s \left( F_{n-1}(x, y) + |x - y| V_{n-1}(x, y) \right) dx dy.
\]

Here \( x = x_{n-1} \) and \( y = y_{n-1} \) are the last components of \( C_n(t) \) and \( C_n(s) \) respectively. Fubini’s theorem is relied on to justify evaluating \( F_n(s, t) \) as an iterated integral.

This recursion formula is awkward to use because of the appearance of the absolute value \( |x - y| \). We remove this absolute value as follows. The function \( F_n(s, t) \) is symmetric in \( (s, t) \). Therefore, it is sufficient to consider \( s \geq t \), which we will always assume. Introduce the integral operator

\[
(I f)(s, t) = 2 \int_0^t \int_y^s f(x, y) \, dx \, dy
\]

Then, for \( 0 \leq t \leq s \),

\[
F_n(s, t) = I (F_{n-1}(s, t) + (s - y) V_{n-1}(s, t)). \tag{4}
\]

This is a recursion formula for \( F_n(s, t) \) in the region \( 0 \leq t \leq s \). We note that for \( f(x, y) = x^i y^j \), \( i, j = 0, 1, 2, \ldots \),

\[
I(f)(s, t) = \frac{s^{i+1} t^{j+1}}{(i+1)(j+1)} + \frac{(i-j)t^{i+j+2}}{(i+1)(j+1)(i+j+2)}. \tag{5}
\]

It follows from (4) and (5) that \( F_n(s, t) \) is a polynomial in \( s, t \) of the form

\[
F_n(s, t) = \sum_{k=0}^{n} f_{n,k} t^{n+k-1}s^{n-k}. \tag{6}
\]

For example, we get

\[
F_2(s, t) = \frac{1}{2}ts^2 - \frac{1}{3}t^2 s + \frac{1}{3}t^3,
\]

\[
F_3(s, t) = \frac{1}{4}t^2 s^3 - \frac{1}{4}t^3 s^2 + \frac{1}{12}t^4 s + \frac{1}{20}t^5,
\]

\[
F_4(s, t) = \frac{1}{24}t^3 s^4 - \frac{1}{24}t^4 s^3 + \frac{1}{120}t^5 s^2 + \frac{1}{420}t^6 s + \frac{1}{210}t^7,
\]

\[
F_5(s, t) = \frac{1}{288}t^4 s^5 - \frac{1}{288}t^5 s^4 + \frac{1}{1660}t^6 s^3 + \frac{1}{1080}t^7 s^2 + \frac{1}{3360}t^8 s + \frac{1}{418144}t^9.
\]

Using (4) and (5), we find

\[
f_{n,0} = \frac{f_{n-1,0}}{n(n-1)} + \frac{1}{n(n-1)(n-2)!^2} \text{ for } n \geq 2.
\]

We set

\[
\tilde{f}_{n,k} = f_{n,k}(n+k-1)! (n-k)!.
\]

Then

\[
\tilde{f}_{n,0} - \tilde{f}_{n-1,0} = n - 1.
\]

Since \( f_{1,0} = 0 \) we find \( \tilde{f}_{n,0} = \frac{1}{2} n(n-1) \) so

\[
f_{n,0} = \frac{n(n-1)}{2(n-1)! n!} = \frac{1}{2(n-1)! (n-2)!}.
\]
Similarly, we obtain
\[ f_{n,1} = -f_{n,0} = -\frac{1}{2(n-1)! (n-2)!}. \] (8)

Now (4) and (5) give
\[ f_{n,k} = \frac{f_{n-1,k}}{(n+k-1)(n-k)} \quad \text{for } 2 \leq k < n. \]

Therefore,
\[ \tilde{f}_{n,k} = \tilde{f}_{n-1,k} \quad \text{for } 2 \leq k < n \]
and
\[ f_{n,k} = \frac{f_{k,k}(2k-1)!}{(n+k-1)(n-k)!} \quad \text{for } 2 \leq k \leq n. \] (9)

Using again (4) and (5) we obtain
\[ f_{n,n} = \sum_{k=0}^{n-1} \frac{f_{n-1,k}(1-2k)}{(n+k-1)(n-k)(2n-1)} + \frac{2}{n! (n-2)! (2n-1)}. \] (10)

In (10) we substitute (7), (8), (9) and find
\[ f_{n,n} = \sum_{k=2}^{n-1} \frac{f_{k,k}(2k-1)! (1-2k)}{(n+k-1)(n-k)! (2n-1)} + \frac{1}{(n-1)! (n-2)! (2n-1)}. \] (11)

This is a recursion formula for the “diagonal sequence” \( f_{k,k} \). With the help of a computer we guess the solution
\[ f_{k,k} = \frac{(k-1)k}{(2k-1)!}. \] (12)

We prove (12) by induction on \( k \). The formula is true for \( k = 1 \). Suppose (12) holds for all \( k \) less than a given \( n \geq 2 \). Then we substitute \( f_{k,k} \) given by (12) on the right-hand side of (11). Since (see the Appendix)
\[ \sum_{k=2}^{n} \frac{k(k-1)(2k-1)}{(n+k-1)! (n-k)!} = 2A_{n,3} - 3A_{n,2} + A_{n,1} = \frac{1}{(n-1)! (n-2)!}, \]
we obtain (12) for \( k = n \). This completes the inductive proof of (12).

Now (9), (12) give
\[ f_{n,k} = \frac{(k-1)k}{(n+k-1)! (n-k)!} \quad \text{for } 2 \leq k \leq n. \] (13)

In particular, using again the Appendix, we have
\[ F_{n}(1,1) = \sum_{k=0}^{n} f_{n,k} = \sum_{k=2}^{n} f_{n,k} = \sum_{k=2}^{n} \frac{(k-1)k}{(n+k-1)! (n-k)!} = A_{n,2} - A_{n,1} = \frac{2^{2n-3}(n-1)}{(2n-1)!}. \]

Thus,
\[ \mathbb{E}(X_{n}) = \mathbb{E}(Y_{n}) = \frac{F_{n}(1,1)}{V_{n}(1,1)} = \frac{2^{2n-3}(n-1)}{(2n-1)! (n-1)^2}. \]
4. Proof of Theorem 2

Let \( s, t \geq 0, \ x \in \mathcal{C}_n(t), \ y \in \mathcal{C}_n(s) \) and define

\[
\mathcal{G}_n(s, t) = \int_{\mathcal{C}_n(s)} \int_{\mathcal{C}_n(t)} \|x - y\|^2 \, dx \, dy.
\]

If we set \( \mathcal{G}_1(s, t) = 0 \) then the function \( \mathcal{G}_n \) is determined recursively by

\[
\mathcal{G}_n(s, t) = \int_0^t \int_0^s (\mathcal{G}_{n-1}(x, y) + 2|x - y|\mathcal{F}_{n-1}(x, y) + (x - y)^2\mathcal{V}_{n-1}(x, y)) \, dx \, dy.
\]

As in the proof of Theorem 1, \( x = x_{n-1}, \ y = y_{n-1} \) and Fubini’s theorem justifies evaluating \( \mathcal{G}_n(s, t) \) as an iterated integral.

The function \( \mathcal{G}_n(s, t) \) is symmetric in \( (s, t) \). Therefore, it is sufficient to consider \( s \geq t \). Then, for \( 0 \leq t \leq s \),

\[
\mathcal{G}_n(s, t) = I(\mathcal{G}_{n-1}(x, y) + 2(x - y)\mathcal{F}_{n-1}(x, y) + (x - y)^2\mathcal{V}_{n-1}(x, y)).
\]

This is a recursion formula for \( \mathcal{G}_n(s, t) \) in the region \( 0 \leq t \leq s \). It follows from (14) and (5) that \( \mathcal{G}_n(s, t) \) is a polynomial in \( s, t \) of the form

\[
\mathcal{G}_n(s, t) = \sum_{k=-1}^n g_{n,k}t^{n+k} s^{n-k}.
\]

For example,

\[
\mathcal{G}_2(s, t) = \frac{1}{2}ts^3 - \frac{1}{2}t^2s^2 + \frac{1}{2}t^3s,
\]

\[
\mathcal{G}_3(s, t) = \frac{1}{24}t^2s^4 - \frac{1}{2}t^3s^3 + \frac{1}{8}t^4s^2 - \frac{1}{15}t^5s + \frac{1}{180}11t^6,
\]

\[
\mathcal{G}_4(s, t) = \frac{1}{72}t^3s^5 - \frac{1}{8}t^4s^4 + \frac{1}{360}31t^5s^3 - \frac{1}{60}t^6s^2 - \frac{1}{180}7t^7s + \frac{1}{120}t^8,
\]

\[
\mathcal{G}_5(s, t) = \frac{1}{1728}13t^4s^6 - \frac{1}{72}t^5s^5 + \frac{1}{8640}77t^6s^4 - \frac{1}{1250}7t^7s^3 - \frac{1}{1008}11t^8s^2 + \frac{1}{2520}9t^9s + \frac{1}{50400}19t^{10}.
\]

First we determine \( g_{n,k} \) for \( k = -1 \). From (14) and (5) we obtain

\[
g_{n,-1} = \frac{g_{n-1,-1}}{(n-1)(n+1)} + \frac{2f_{n-1,0}}{(n-1)(n+1)} + \frac{1}{(n-2)!^2(n-1)(n+1)}.
\]

Using (7) this yields

\[
g_{n,-1} = \frac{g_{n-1,-1}}{(n-1)(n+1)} + \frac{1}{(n-1)!^2(n-3)!^2(n+1)!} + \frac{1}{(n-2)!^2(n-1)(n+1)!}.
\]

Let

\[
g_{n,k} = \frac{\tilde{g}_{n,k}}{(n-k)!^2(n+k)!}.
\]

Then

\[
\tilde{g}_{n,-1} - \tilde{g}_{n-1,-1} = (n-1)^2n.
\]

This gives

\[
\tilde{g}_{n,-1} = \tilde{g}_{1,-1} + \sum_{m=2}^n (m-1)^2m = \frac{1}{12}n(n-1)(3n-2)(n+1).
\]
Therefore, \[ g_{n,-1} = \frac{n(n-1)(3n-2)(n+1)}{12(n-1)! (n+1)!} = \frac{n(3n-2)}{12(n-2)! n!}. \] (16)

From (14) and (5) we obtain the recursion formula \[ g_{n,0} = \frac{g_{n-1,0}}{n^2} + \frac{2f_{n-1,1} - 2f_{n-1,0}}{n^2} - \frac{2}{n^2(n-2)!}. \]

This gives \( \tilde{g}_{n,0} - \tilde{g}_{n-1,0} = -2(n-1)^3 \)
and \[ g_{n,0} = -\frac{1}{2(n-2)!^2}. \] (17)

From (14) and (5) we obtain the recursion formula \[ g_{n,1} = \frac{g_{n-1,1}}{(n-1)(n+1)} + \frac{2f_{n-1,2} - 2f_{n-1,1}}{(n-1)(n+1)} + \frac{1}{(n-2)!^2(n-1)(n+1)}. \]

This gives \( \tilde{g}_{n,1} - \tilde{g}_{n-1,1} = 4(n-2) + n(n-1)^2 \)
and \[ g_{n,1} = \frac{2(n-1)(n-2)}{(n-1)! (n+1)!} + \frac{n(3n-2)(n+1)}{12(n-1)! (n+1)!}. \] (18)

Now let \( k \geq 2 \). Then (14) and (5) give \[ g_{n,k} = \frac{g_{n-1,k}}{(n+k)(n-k)} + \frac{2f_{n-1,k+1} - 2f_{n-1,k}}{(n+k)(n-k)}. \]

This leads to the recursion formula \( \tilde{g}_{n,k} - \tilde{g}_{n-1,k} = 4k(n-k^2-1) \) for \( 2 \leq k < n \).

This gives \[ \tilde{g}_{n,k} = \tilde{g}_{k,k} + 2k(k-n)(2k^2 - k - n + 1). \] (19)

From (14) and (5) we find \[ g_{n,n} = -\sum_{k=1}^{n-1} \frac{kg_{n-1,k}}{(n-k)(n+k)n} - \sum_{k=1}^{n-2} \frac{2kf_{n-1,k+1}}{(n-k)(n+k)n} + \sum_{k=2}^{n-1} \frac{2kf_{n-1,k}}{(n-k)(n+k)n}. \] (20)

It should be noted that the term \( x - y \) in (14) does not contribute to this formula because it is a symmetric polynomial. Using (13) and the Appendix we evaluate the second and third sums on the right-hand side of (20). We find

\[ -\sum_{k=1}^{n-2} \frac{2kf_{n-1,k+1}}{(n-k)(n+k)n} = \frac{4^{n-1}(n+1)}{(2n)!} - \frac{n(n-2)}{n!^2} - \frac{n+1}{(2n-1)!}, \]
\[ \sum_{k=2}^{n-1} \frac{2kf_{n-1,k}}{(n-k)(n+k)n} = \frac{4^{n-1}(n+1)}{(2n)!} + \frac{n(n-2)}{n!^2} - \frac{1}{2(2n-3)!}. \]
Therefore, (20) can be written as
\[ g_{n,n} = -\sum_{k=1}^{n-1} \frac{k g_{n-1,k}}{(n-k)(n+k)n} + \frac{2^{n-1}(n+1)}{(2n)!} \frac{2(n^2-n+1)}{(2n-1)!}. \]  \hspace{1cm} (21)

Using (16), (18) we find that
\[ -\sum_{k=1}^{n-1} \frac{k g_{n-1,k}}{(n-k)(n+k)n} = \frac{2(n-3)(n-2)}{n!^2(n+1)}. \]  \hspace{1cm} (22)

We also have
\[ -\sum_{k=2}^{n-1} \frac{2k^2(k-n+1)(2k^2-k-n+2)}{n(n-k)! (n+k)!} = \frac{4^{n-1}(n+1)(4n-3)}{(2n)!} + \frac{2(n^2-n+1)}{(2n-1)!} \frac{4n^3+n^2+9n-12}{(n+1)!^2}. \]  \hspace{1cm} (23)

We now substitute (19), (22), (23) in (21) and obtain
\[ g_{n,n} = -\sum_{k=2}^{n-1} \frac{k(2k)! g_{k,k}}{(n-k)!(n+k)! n} + \frac{4^{n-1}(n+1)(4n-1)}{(2n)!} - \frac{4n-1}{n(n-1)!^2}. \]  \hspace{1cm} (24)

This is a recursion formula for the sequence \( g_{n,n} \). With the help of a computer we guess
\[ g_{k,k} = \frac{(k-2)(4k-1)}{30(2k-3)!} \]  \hspace{1cm} for \( k \geq 2 \). \hspace{1cm} (25)

We prove (25) by induction on \( k \). This equation is true for \( k = 2 \). Suppose it is true for all \( k \) less than a given \( n \geq 3 \). Then we use (25) in (24). Using the Appendix, we evaluate the sum and obtain the correct expression for \( g_{n,n} \). The inductive proof is complete.

From (19) we get, for \( 2 \leq k \leq n \),
\[ g_{n,k} = \frac{k(2k-1)(2k-2)(2k-4)}{15(n+k)!(n-k)!} + \frac{2k(k-n)(2k^2-k-n+1)}{(n+k)!(n-k)!}. \]  \hspace{1cm} (26)

In particular, we have
\[ G_n(1,1) = \sum_{k=-1}^{n} g_{n,k} = \frac{n(n-1)(7n-4)}{30n!^2}. \]

Thus,
\[ E(X_n^2) = E(Y_n^2) = \frac{G_n(1,1)}{V_n(1,1)} = \frac{(n-1)(7n-4)}{30n}. \]

5. Solution to the recurrence of Bourn and Willenbring

In [1] a double sequence \( \mathcal{M}_{p,q} \), \( p, q \in \mathbb{N}_0 \), is defined as the solution of the recursion (3). Here we “solve” this recursion.

In order to simplify the recursion we set
\[ \ell_{p,q} = \frac{(p+q-1)!}{(p-1)!(q-1)!} \mathcal{M}_{p,q}, \quad p, q \in \mathbb{N}, \]
and $\mathcal{L}_{p,0} = \mathcal{L}_{0,q} = 0$. Then (3) becomes

$$
\mathcal{L}_{p,q} = \mathcal{L}_{p-1,q} + \mathcal{L}_{p,q-1} + |p-q| \binom{p+q-2}{p-1}, \quad p, q \in \mathbb{N}.
$$

(27)

Note that $\mathcal{L}_{p,q}$ is a nonnegative integer for all $p, q$. The matrix $\mathcal{L}_{p,q}, p, q = 0, \ldots, 6$, is

$$
\begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 3 & 6 & 10 & 15 \\
0 & 1 & 2 & 8 & 22 & 47 & 86 \\
0 & 3 & 8 & 16 & 48 & 125 & 274 \\
0 & 6 & 22 & 48 & 96 & 256 & 642 \\
0 & 10 & 47 & 125 & 256 & 512 & 1280 \\
0 & 15 & 86 & 274 & 642 & 1280 & 2560
\end{pmatrix}.
$$

Consider the more general recursion

$$
\mathcal{K}_{p,q} = \mathcal{K}_{p-1,q} + \mathcal{K}_{p,q-1} + a_{p,q}, \quad p, q \in \mathbb{N},
$$

(28)

with initial condition $\mathcal{K}_{p,0} = \mathcal{K}_{0,q} = 0$, where $\{a_{p,q}\}$ is a given double sequence. As a special case consider $a_{p,q} = 1$ if $p = i, q = j$ for some given $i, j \in \mathbb{N}$, and $a_{p,q} = 0$ otherwise. If $p \geq i, q \geq j$ the corresponding solution is

$$
\mathcal{K}_{p,q} = \binom{p+q-i-j}{p-i}.
$$

Therefore, by superposition, the solution of (27) is

$$
\mathcal{L}_{p,q} = \sum_{i=1}^{p} \sum_{j=1}^{q} |i-j| \binom{i+j-2}{i-1} \binom{p+q-i-j}{p-i}.
$$

We rewrite this in the slightly nicer form

$$
\mathcal{L}_{p+1,q+1} = \sum_{i=0}^{p} \sum_{j=0}^{q} |i-j| \binom{i+j}{i} \binom{p+q-i-j}{p-i}.
$$

(29)

In a sense we found the solution of (27) and so also of (3). However, the question arises whether we can express the double sum in (29) (more) explicitly. We evaluate the double sum in (29) when $p = q$. There is some evidence (from the computer) that this is not possible when $p \neq q$. We could handle some special case like small $p$ or small $|p-q|$.

We now evaluate the double sum in (29) when $p = q$. First, we introduce $i+j = m$ as a new summation index. Then

$$
\mathcal{L}_{p+1,p+1} = \sum_{m=0}^{2p} \sum_{i=0}^{p} |m-2i| \binom{m}{i} \binom{2p-m}{p-i},
$$

where we use the standard convention that $\binom{n}{k} = 0$ unless $0 \leq k \leq n$. We remove the absolute value by writing

$$
\mathcal{L}_{p+1,p+1} = 2 \sum_{m=0}^{2p} \sum_{i=0}^{\lfloor m/2 \rfloor} (m-2i) \binom{m}{i} \binom{2p-m}{p-i}.
$$

(30)
The inside sum can be evaluated as a telescoping sum as follows. Consider first an even \( m = 2r \). Then
\[
\sum_{i=0}^{r} 2(r-i)\binom{2r}{i}\binom{2p-2r}{p-i} = \sum_{j=0}^{r} 2j\binom{2r}{r-j}\binom{2s}{s+j},
\]
where we set \( i = r-j \) and \( s = p-r \). Note that \( \binom{2r}{r-j} = \binom{2r}{r+j} \). Now
\[
2j\binom{2r}{r-j}\binom{2s}{s+j} = d_j - d_{j+1}, \quad j = 0, 1, \ldots, r,
\]
where
\[
d_j = \frac{(r+j)(s+j)}{r+s}\binom{2r}{r-j}\binom{2s}{s+j}.
\]
Note that \( d_{r+1} = 0 \). Therefore,
\[
\sum_{i=0}^{r} 2(r-i)\binom{2r}{i}\binom{2p-2r}{p-i} = d_0 - d_{r+1} = d_0 = \frac{rs}{r+s}\binom{2r}{r}\binom{2s}{s}.
\]
If \( m = 2r + 1 \) is odd, we use
\[
(2j + 1)\binom{2r+1}{r-j}\binom{2s-1}{s+j} = e_j - e_{j+1},
\]
where
\[
e_j = \frac{(r+1+j)(s+j)}{r+s}\binom{2r+1}{r-j}\binom{2s-1}{s+j}.
\]
Therefore,
\[
\sum_{i=0}^{r} (2r+1-2i)\binom{2r+1}{i}\binom{2p-2r-1}{p-i} = e_0 = \frac{(r+1)s}{r+s}\binom{2r+1}{r}\binom{2s-1}{s-1}.
\]
Substituting these results in (30) we obtain
\[
\mathcal{L}_{p+1, p+1} = \sum_{r=0}^{p} \frac{2r(p-r)}{p} \binom{2r}{r}\binom{2p-2r}{p-r} + \sum_{r=0}^{p-1} \frac{2(r+1)(p-r)}{p} \binom{2r+1}{r}\binom{2p-2r-1}{p-r-1}.
\]
The sums appearing in (31) are easy to evaluate. We note that
\[
(1 - 4x)^{-\frac{1}{2}} = \sum_{n=0}^{\infty} \binom{2n}{n} x^n \quad \text{for } |x| < \frac{1}{4},
\]
so
\[
2x(1 - 4x)^{-3/2} = \sum_{n=0}^{\infty} n\binom{2n}{n} x^n.
\]
If we compare coefficients in
\[
\frac{2x}{(1 - 4x)^{3/2}} = \frac{4x^2}{(1 - 4x)^3} = \sum_{n=0}^{\infty} 2^{n-3} n(n-1)x^n,
\]
we find the first sum in (31). Using \( \binom{2r+1}{r} = \frac{1}{2} \binom{2r+2}{r+1} \) we also find the second sum. Finally, we obtain
\[
\mathcal{L}_{p+1, p+1} = \frac{2}{p} 2^{2p-3} p(p-1) + \frac{2}{p} 2^{2p-3} p(p+1).
\]
Therefore,
\[ L_{p+1, p+1} = 2^{2p-1} p, \quad M_{p+1, p+1} = 2^{2p-1} p \frac{p!^2}{(2p + 1)!}. \]

Since [1] showed that \( E(X_n) = M(p, p) \), this result agrees with Theorem 1.

### Appendix: Binomial sums

For \( p \in \mathbb{N}_0, \ n \in \mathbb{N} \) we define
\[
A_{n,p} = \sum_{k=1}^{n} \frac{k^p}{(n+k-1)! (n-k)!}, \quad B_{n,p} = \sum_{k=1}^{n} \frac{k^p}{(n+k)! (n-k)!}.
\]

These sums can be calculated recursively from
\[
A_{n,0} = \frac{4^{n-1}}{(2n-1)!}, \quad B_{n,0} = \frac{2^{2n-1}}{(2n)!} - \frac{1}{2n!^2},
\]
and
\[
A_{n,p} = nA_{n,p-1} - B_{n-1,p-1}, \quad B_{n,p} = A_{n,p-1} - nB_{n,p-1}.
\]

For example, we get
\[
A_{n,1} = \frac{2^{2n-3}}{(2n-1)!} + \frac{1}{2(2n-1)!^2}, \quad B_{n,1} = \frac{1}{2n! (n-1)!},
\]
\[
A_{n,2} = \frac{n2^{2n-3}}{(2n-1)!} + \frac{1}{2(2n-1)!^2}, \quad B_{n,2} = \frac{2^{2n-3}}{(2n-1)!},
\]
\[
A_{n,3} = \frac{4^{n-2}(3n-1)}{(2n-1)!} + \frac{1}{2n! (n-1)!}, \quad B_{n,3} = \frac{1}{2(n-1)!^2}.
\]

### Acknowledgement

Frohmader would like to thank Alexander Heaton, Jeb Willenbring, and Rebecca Bourn for helpful discussions.

### References


Received 2019-12-12. Revised 2020-09-08. Accepted 2020-09-29.

**ANDREW FROHMADER**: frohmad4@uwm.edu

*Department of Mathematical Sciences, University of Wisconsin, Milwaukee, WI, United States*

**HANS VOLKMER**: volkmer@uwm.edu

*Department of Mathematical Sciences, University of Wisconsin, Milwaukee, WI, United States*
ASSOCIATION AND SIMPSON CONVERSION IN $2 \times 2 \times 2$ CONTINGENCY TABLES

SVANTE LINUSSON AND MATTHEW T. STAMPS

We study a generalisation of Simpson reversal (also known as Simpson’s paradox or the Yule–Simpson effect) to $2 \times 2 \times 2$ contingency tables and characterise the cases for which it can and cannot occur with two combinatorial-geometric lemmas. We also present a conjecture based on some computational experiments on the expected likelihood of such events.

1. Introduction

Simpson reversal is a phenomenon in statistics in which a common trend among different groups disappears, or even reverses, when the groups are combined. More precisely, if $A_1$, $A_2$, and $B$ are events such that $A_1$ and $A_2$ are positively associated given $B$ and also given its complement $\overline{B}$, i.e.,

$$P(A_1 \cap A_2 \mid B) > P(A_1 \mid B)P(A_2 \mid B) \quad \text{and} \quad P(A_1 \cap A_2 \mid \overline{B}) > P(A_1 \mid \overline{B})P(A_2 \mid \overline{B}),$$

then it is possible for $A_1$ and $A_2$ to be independent to each other or negatively associated marginally, independent of $B$, i.e.,

$$P(A_1 \cap A_2) \leq P(A_1)P(A_2).$$

Real-world examples of the paradox are well-documented, for instance in [2; 10], and it has also been observed and studied in several applications from biology, such as [6; 11].

In this paper, we consider a generalisation of Simpson reversal to trios of events with respect to a fourth. The practicality of such a generalisation can be illustrated with the following scenario that is motivated by a recent paper in biology [4]. When measuring the association between two bacterial DNA loci in a sample, it is important to be aware of Simpson reversal since it is possible that the measured association might actually be the opposite if measured on two subsets of the bacteria separately, which could lead to a misinterpretation of the data. In the same manner, it is important to understand the possible misinterpretations that could occur when measuring association on three loci, sometimes called the fitness landscape [1]. This paper is a first step in understanding the statistical pitfalls in such investigations.

The generalisation of Simpson reversal to trios is significantly more complex than Simpson reversal for pairs since a trio of random events can satisfy a combination of mutual, marginal, and conditional associations with respect to one another, whereas a pair of events can be positively or negatively associated or independent. We ask if $A_1$, $A_2$, $A_3$, and $B$ are events such that $A_1$, $A_2$, and $A_3$ satisfy a common


Keywords: Simpson’s paradox, correlation reversal, association, triangulations.
set of mutual, marginal, and conditional associations given $B$ and also given its complement, what sets of associations can the $A_1$, $A_2$, and $A_3$ satisfy independent of $B$? We show that while there are many ways in which a set of associations among $A_1$, $A_2$, and $A_3$ can change (we call such instances Simpson conversions), it is not possible for every set of associations to convert into every other set of associations. Our main result, Theorem 5.7, characterises the Simpson conversions for trios of events. The proof extends a well-known geometric interpretation of Simpson reversal in terms of triangulations of the square to a geometric interpretation of Simpson conversion to triangulations of the cube. Our characterisation consists of two parts: First, we establish some combinatorial-geometric lemmas (involving triangulations of the cube) to preclude certain instances of Simpson conversions. Then, we verify experimentally instances of all the other Simpson conversions.

The remainder of this paper is structured as follows: In Section 2, we review a well-known example of Simpson reversal and several geometric interpretations in the literature. In Section 3, we propose the generalisation of Simpson reversal for $2 \times 2 \times 2$ contingency tables, which we call Simpson conversion, that involves an arrangement of hyperplanes in $\mathbb{R}^8$ and the set of triangulations of the 3-dimensional cube. We describe the relationship between the linear forms defining said hyperplane arrangement and the triangulations in Section 4. Section 5 contains the main results of the paper, in which we characterise the 3-dimensional analogue of Simpson's reversals for $2 \times 2 \times 2$ contingency tables. We conclude the paper with a conjecture on the frequency of Simpson conversion and some observations based on computational experiments in Section 6.

### 2. Simpson reversal in two dimensions

Here we review a well-known example and several geometric interpretations of Simpson reversal in two dimensions.

**Example 2.1.** A concrete example of Simpson reversal is illustrated by the voting results for (the Senate version) of the Civil Rights Act of 1964 in the United States House of Representatives. The votes are listed in Table 1 broken down according to political party (Democrats and Republicans) and region of the country (Northern, Southern, and all states), as presented in [5]. From these tables, one can observe that a higher percentage of Democrats voted in favour of the bill in both the Northern and the Southern states (95% and 9% of the Democrats compared to 85% and 0% of the Republicans, respectively), but a higher percentage of Republicans voted in favour of the bill overall (80% of the Republicans compared to 63% of the Democrats). This instance of Simpson reversal can be explained by the fact that the relationship

<table>
<thead>
<tr>
<th>Northern states</th>
<th>Southern states</th>
<th>all states</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democrats</td>
<td>Democrats</td>
<td>Democrats</td>
</tr>
<tr>
<td>144 8</td>
<td>8 83</td>
<td>152 91</td>
</tr>
<tr>
<td>Republicans</td>
<td>Republicans</td>
<td>Republicans</td>
</tr>
<tr>
<td>137 24</td>
<td>0 11</td>
<td>137 35</td>
</tr>
</tbody>
</table>

**Table 1.** House of Representatives voting results for the Civil Rights Act of 1964 according to political party among Northern states (left), Southern states (middle), or all states (right).
between party and vote on the bill was significantly affected by the regions corresponding to the voters. Indeed, most of the Southern representatives at that time were Democrats, and the vast majority of negative votes came from that region.

While the existence of a Simpson reversal is surprising — even paradoxical — at first glance, there are several geometric interpretations that illustrate why the phenomenon is not only possible, but a relatively common occurrence. A well-known geometric interpretation of Simpson reversal is as follows: Suppose $v_1, v_2, w_1,$ and $w_2$ are vectors in $\mathbb{R}^2_+$ based at the origin such that the slope of $v_i$ is greater than the slope of $w_i$ for $i = 1, 2$. Then it is possible that the slope of $v_1 + v_2$ is less than the slope of $w_1 + w_2$ as shown in Figure 1. In Example 2.1 it corresponds to $v_1 = (8, 144), v_2 = (83, 8), w_1 = (24, 137), w_2 = (11, 0)$.

Another interpretation of Simpson reversal can be seen geometrically via triangulations of a square. A real-valued function $f : P_0 \to \mathbb{R}$ on the vertices $P_0$ of a convex polygon $P$ induces a unique triangulation on $P$ by taking the convex hull of the set $P_f = \{(p_1, p_2, f(p_1, p_2)) \mid (p_1, p_2) \in P_0\} \subseteq \mathbb{R}^3$ and projecting its upper envelope onto $P$, provided no four points in $P_f$ are coplanar. This process is nicely explained in [3] and illustrated for the case where $P$ is a square in Figure 2.

Note that a function $f : \{0, 1\}^2 \to \mathbb{R}$ induces the triangulation of the square with the diagonal edge between the points $(0, 0)$ and $(1, 1)$ if $f(0, 0) + f(1, 1) > f(0, 1) + f(1, 0)$ and between the points $(0, 1)$ and $(1, 0)$ if the inequality is reversed. Since a $2 \times 2$ contingency table with values $F_{00}, F_{10}, F_{01}, F_{11}$ is positively associated if $F_{00} \cdot F_{11} > F_{01} \cdot F_{10}$ and negatively associated if the inequality is reversed, the sign of the association of the contingency table is encoded by the triangulation of the square induced by the function $f : \{0, 1\}^2 \to \mathbb{R}$ given by $f(x, y) = \ln(F_{xy})$. Using upper case letters to denote the entries of a table and lower case letters to denote the corresponding function (as defined above), a pair of contingency tables $F$ and $G$ exhibit a Simpson reversal when their corresponding functions $f$ and $g$ induce the same triangulation of the square, while the function $h : \{0, 1\}^2 \to \mathbb{R}$ given by $h(x, y) = \ln(F_{xy} + G_{xy})$ corresponding to the table $F + G$ induces the other triangulation of the square.
This can be seen geometrically via triangulations of a square: Any generic real-valued vector in the function on the unit square in triangulation to every generic 2 \times 2 contingency tables corresponding to each chamber of congress over each region into a 2 \times 2 \times 2 contingency table. Understanding how

To see how such an occurrence is possible, observe that while the association of a contingency table \( F \) is determined by the linear form \( w := f_{00} + f_{11} - f_{01} - f_{10}, \) where \( f_{xy} = \ln F_{xy}, \) that decomposes \( \mathbb{R}^d \) into convex cones (half-spaces), the corresponding regions in the space of upper case letters, where the addition of contingency tables is done, are nonconvex. Thus, the sum of two contingency tables belonging to the region of one association can belong to the other association. It is interesting to consider what these nonconvex regions look like, but that is not the focus of this paper.

**Example 2.2.** To illustrate the ideas above with the data in Example 2.1, let \( N, S, \) and \( A = N + S \) denote the contingency tables in Table 1 corresponding to Northern, Southern, and all states, respectively, and label the upper left, upper right, lower left, and lower right entries of each table by \((0, 0), (1, 0), (0, 1),\) and \((1, 1),\) respectively. We can see that \( N \) and \( S \) are both positively associated, while their sum \( A \) is negatively associated since

\[
N_{00} \cdot N_{11} = 144 \cdot 24 > 137 \cdot 8 = N_{01} \cdot N_{10},
\]
\[
S_{00} \cdot S_{11} = 8 \cdot 11 > 83 \cdot 0 = S_{01} \cdot S_{10},
\]
\[
A_{00} \cdot A_{11} = 152 \cdot 35 < 91 \cdot 137 = N_{01} \cdot N_{10}.
\]

With these interpretations, the existence of Simpson reversals is not so surprising. In fact, from the perspective of causality, see [8; 9], it is a rather natural notion.

**3. Associations in three dimensions**

The instance of Simpson reversal in the House of Representatives vote on the Civil Rights Bill of 1964 presented in Example 2.1 (taken from [5]) can also be observed in the Senate vote on the same bill, the results of which are listed in Table 2 appended to the data in Table 1.

With this, one might be interested in studying the various relationships between party, vote on the bill, and chamber of congress simultaneously by combining the two \( 2 \times 2 \) contingency tables corresponding to each chamber of congress over each region into a \( 2 \times 2 \times 2 \) contingency table. Understanding how
Association and Simpson Conversion in $2 \times 2 \times 2$ Contingency Tables

<table>
<thead>
<tr>
<th></th>
<th>Northern states</th>
<th>yes</th>
<th>no</th>
<th>Southern states</th>
<th>yes</th>
<th>no</th>
<th>all states</th>
<th>yes</th>
<th>no</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>House</strong></td>
<td>Democrats</td>
<td>144</td>
<td>8</td>
<td>Democrats</td>
<td>8</td>
<td>83</td>
<td>Democrats</td>
<td>152</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>Republicans</td>
<td>137</td>
<td>24</td>
<td>Republicans</td>
<td>0</td>
<td>11</td>
<td>Republicans</td>
<td>137</td>
<td>35</td>
</tr>
<tr>
<td><strong>Senate</strong></td>
<td>Democrats</td>
<td>45</td>
<td>1</td>
<td>Democrats</td>
<td>1</td>
<td>20</td>
<td>Democrats</td>
<td>46</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>Republicans</td>
<td>27</td>
<td>5</td>
<td>Republicans</td>
<td>0</td>
<td>1</td>
<td>Republicans</td>
<td>27</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 2. Voting results for the Civil Rights Act of 1964 according to legislative chamber and political party among Northern states, Southern states, and all states.

Observed associations change when the data from different regions are combined becomes much more complicated than in the 2-dimensional case since there are many more notions of association for a 3-way contingency table. Indeed, a $2 \times 2 \times 2$ contingency table can exhibit mutual (all variables dependent on each other), marginal (two variables are dependent ignoring the third), and conditional (two variables are dependent given the third) associations, and these three types of associations are distinct from one another in the sense that dependencies of one type do not necessarily imply dependencies of the other types. More precisely, if $A_1$, $A_2$, and $A_3$ are random events, then there are eight distinct relations arising from mutual dependencies of the form

$$P(X \cap Y \cap Z) \neq P(X)P(Y)P(Z),$$

three distinct relations arising from marginal dependencies of the form

$$P(X \cap Y) \neq P(X)P(Y),$$

and six distinct relations arising from conditional dependencies of the form

$$P(X \cap Y \mid Z) \neq P(X \mid Z)P(Y \mid Z),$$

where $X$, $Y$, and $Z$ are distinct events chosen from $A_1$ or $\bar{A}_1$, $A_2$ or $\bar{A}_2$, and $A_3$ or $\bar{A}_3$. Note that there are only three and six distinct relations from marginal and conditional dependencies, respectively, since

$$P(X \cap Y) \neq P(X)P(Y) \iff P(X \cap \bar{Y}) \neq P(X)P(\bar{Y}).$$

**Example 3.1.** For the data in Table 2, voting “yes” ($Y$) and being a Democrat ($D$) are positively marginally associated, independent of the legislative chamber, for both Northern ($N$) and Southern ($S$) states, but negatively marginally associated for all states, since

$$P(Y \cap D \mid N) = \frac{189}{391} > \frac{353}{391} \cdot \frac{198}{391} = P(Y \mid N) \cdot P(D \mid N),$$

$$P(Y \cap D \mid S) = \frac{9}{124} > \frac{9}{124} \cdot \frac{112}{124} = P(Y \mid S) \cdot P(D \mid S),$$

$$P(Y \cap D) = \frac{198}{515} < \frac{362}{515} \cdot \frac{310}{515} = P(Y) \cdot P(D).$$

Therefore, the marginal association between voting “yes” and being a Democrat, independent of the legislative chamber, in Table 2 exhibits a Simpson reversal. We leave it to the reader to check that all of the other marginal and mutual, but only two of the conditional, associations in Table 2 exhibit Simpson reversals.
For a $2 \times 2 \times 2$ contingency table with values $F_{xyz}$, these 17 dependencies give linear relations on the variables $f_{xyz} = \ln F_{xyz}$, which have a natural correspondence with the set of linear relations arising from the 74 triangulations of the 3-dimensional cube. Just as a generic $2 \times 2$ contingency table induces a triangulation of the square, a generic $2 \times 2 \times 2$ contingency table induces a triangulation of the cube (into tetrahedra) via a projection of the upper envelope of the convex hull of the points $(x, y, z, f_{xyz})$ in $\mathbb{R}^4$. In [1] it is shown that the contingency tables that induce each triangulation of the cube are determined by 20 linear relations, which we list in Appendix A. The conditional associations correspond to the forms labelled $a$ through $f$; the marginal associations correspond to the sums $g + h, i + j,$ and $k + l$; and the mutual associations correspond to the forms labelled $m$ through $t$. The triangulations of the cube are listed in Appendix B, organised into six types according to symmetry — every triangulation of a given type can be obtained from any other triangulation of that type via a rotation or reflection. Triangulations of type I consist of five tetrahedra, while all other triangulations consist of six. We have used the same notation as in [1] for easy comparison, and we give a detailed explanation of how to interpret the figures in Appendix B in Example 4.1.

4. The correspondence between linear forms and triangulations

In this section, we describe the algebrogeometric correspondence between the linear forms in Appendix A and the triangulations of the 3-dimensional cube in Appendix B. The positive quadrant of $\mathbb{R}^8$ is divided into 74 regions by the 20 hyperplanes associated to the linear forms, where each region corresponds to a unique triangulation of the cube. The signs of the forms $a$ through $f$ correspond to the face diagonals on the six squares that make up the surface of the cube; the signs of the forms $g$ through $l$ correspond to flipping the interior diagonal of the cube within the six rectangles formed by opposite pairs of edges in the cube (these six rectangles each slice the cube into two triangular prisms); and finally the forms $m$ through $t$ correspond to whether or not the interior diagonal is present when passing between the triangulations of types I and II.

Example 4.1. Consider the triangulation labelled 3 in Appendix B shown in Figure 3. This triangulation consists of six tetrahedra, namely

\[
\begin{align*}
\{000, 001, 010, 100\}, & \quad \{011, 001, 010, 100\}, & \quad \{011, 001, 111, 100\}, \\
\{101, 001, 111, 100\}, & \quad \{011, 010, 111, 100\}, & \quad \{110, 010, 111, 100\}, 
\end{align*}
\]

where the notation $xyz$ means $(x, y, z)$. The sequence of letters $b, d, -e, -t$ below the cube indicates this triangulation is induced by the contingency tables $F$ satisfying the relations

\[
\begin{align*}
0 < b & := f_{001} + f_{111} - f_{011} - f_{101}, \\
0 < d & := f_{010} + f_{111} - f_{110} - f_{011}, \\
0 > e & := f_{000} + f_{011} - f_{010} - f_{001}, \\
0 > t & := f_{010} + f_{001} + f_{111} - f_{100} - 2f_{011}, 
\end{align*}
\]
Figure 3. Triangulation 3 from Appendix B.

from Appendix A or, equivalently,

\[
\frac{F_{001}F_{111}}{F_{011}F_{101}} > 1, \quad \frac{F_{010}F_{111}}{F_{110}F_{011}} > 1, \quad \frac{F_{000}F_{011}}{F_{010}F_{001}} < 1, \quad \frac{F_{010}F_{001}F_{111}}{F_{100}F_{011}^2} < 1.
\]

The values on the other 16 linear forms follow from these four.

Notice that the vertices 001, 100, 010, and 111 in triangulation 3 are each incident with three face (blue) diagonals, and the remaining four vertices are not incident with any face diagonals. Such vertices will play an important role in the next section, so we will give them a name.

**Definition 4.2.** For a given triangulation of the cube:

- A vertex is called *full* and marked with a filled circle in Example 4.1 and Appendix B if it is incident with all three face diagonals belonging to the square faces that contain it.

- A vertex is called *empty* and marked with an empty circle in Example 4.1 and Appendix B if it is not incident with any of the three face diagonals belonging to the square faces that contain it.

One can see in Appendix B that triangulations of types I and II have four full and four empty vertices, triangulations of type III have two full and two empty vertices, triangulations of type IV have no full and two empty vertices, triangulations of type V have one full and one empty vertex, and triangulations of type VI have two full and no empty vertices.

We conclude this section with some additional observations on the relationship between associations and triangulations of the cube: In the triangulations of types I and II, the face diagonals in each pair of opposite faces have opposite directions, which means that the association of any two variables in the corresponding contingency tables is dependent on the value of the third. On the contrary, the face diagonals in each pair of opposite faces in the triangulations of types IV and VI have the same direction, which means that the association of any two variables in the corresponding contingency tables is not dependent on the value of the third. Triangulations of type III have exactly one pair of opposite faces whose face diagonals have the same direction, while triangulations of type V have exactly one pair of opposite faces whose face diagonals have different directions.
5. Simpson conversion in three dimensions

This is the main section where we consider the question: if a pair of $2 \times 2 \times 2$ contingency tables induce the same triangulation of the cube, is it possible that their componentwise sum induces a different triangulation? Just as in the case of $2 \times 2$ contingency tables, it is not difficult to find an example that answers this question in the affirmative. Unlike the case of $2 \times 2$ contingency tables, however, there are many different ways in which these instances can arise: instead of reversing from one (or two) triangulations to the other, it is possible to convert from one triangulation of a cube to several of the other 73. We call a pair of $2 \times 2 \times 2$ contingency tables that induce the same triangulation $A$ of the cube whose sum induces a different triangulation $B$ of the cube a Simpson conversion from $A$ to $B$. Our main theorem is a characterisation of the pairs of triangulations, $A$ and $B$, for which there exists a Simpson conversion from $A$ to $B$. We proceed with some essential lemmas for the characterisation.

5A. Setup and essential lemmas. For each $x \in \{0, 1\}$, we define

$$\bar{x} = \begin{cases} 1, & x = 0, \\ 0, & x = 1. \end{cases}$$

We denote by $F_v$ the value of the function $F : \{0, 1\}^k \rightarrow \mathbb{R}_+$ on input $v \in \{0, 1\}^k$, i.e., the entry in the corresponding contingency table. We start with a lemma giving relations that follow if we have Simpson reversal happening in two dimensions.

**Lemma 5.1.** Let $F, G : \{0, 1\}^2 \rightarrow \mathbb{R}_+$ and $(x, y) \in \{0, 1\}^2$. If $F_{\bar{x}y} F_{\bar{y}x} < F_{xy} F_{\bar{y}x}$, $G_{\bar{x}y} G_{\bar{y}x} < G_{xy} G_{\bar{y}x}$, and

$$(F_{\bar{x}y} + G_{\bar{x}y})(F_{xy} + G_{xy}) > (F_{xy} + G_{xy})(F_{\bar{x}y} + G_{\bar{x}y}),$$

then one of the following pairs of inequalities must hold:

$$F_{\bar{x}y} G_{xy} > F_{xy} G_{\bar{x}y} \text{ and } F_{\bar{x}y} G_{xy} < F_{xy} G_{\bar{y}x} \quad \text{or} \quad F_{\bar{x}y} G_{xy} < F_{xy} G_{\bar{y}x} \text{ and } F_{\bar{x}y} G_{xy} > F_{xy} G_{\bar{y}x}.$$ 

*Proof.* It suffices to prove this for the case where $x = y = 0$. Suppose $F_{10} F_{01} < F_{00} F_{11}$, $G_{10} G_{01} < G_{00} G_{11}$, and $(F_{10} + G_{10})(F_{01} + G_{01}) > (F_{00} + G_{00})(F_{11} + G_{11})$. Expanding the products in the third inequality and multiplying each side by $F_0 G_{00}$, we get

$$F_{00} F_{10} F_{00} G_{00} + F_{00} F_{10} G_{00} G_{01} + F_{00} F_{01} G_{00} G_{10} + F_{00} G_{00} G_{10} G_{01} > F_{00} F_{00} F_{11} G_{00} + F_{00} F_{00} G_{00} G_{11} + F_{00} F_{11} G_{00} G_{00} + F_{00} G_{00} G_{00} G_{11}.$$ 

Applying the first and second inequalities to the first and fourth summands, we get

$$F_{00} F_{10} G_{00} G_{01} + F_{00} F_{01} G_{00} G_{10} > F_{00} F_{00} G_{00} G_{11} + F_{00} F_{11} G_{00} G_{00},$$

and hence $(F_{10} G_{00} - F_{00} G_{10})(F_{00} G_{01} - F_{01} G_{00}) > 0$. The desired result follows immediately. \qed

The next lemma says that if two $2 \times 2 \times 2$ contingency tables, $F$ and $G$, both have a full vertex at $xyz$, then the sum $F + G$ cannot have an empty vertex at $xyz$. That is, if all three face diagonals are incident to vertex $xyz$ in both $F$ and $G$, then they cannot all flip in the sum $F + G$. 
Lemma 5.2. Let $F, G : \{0, 1\}^3 \to \mathbb{R}_+$ and $(x, y, z) \in \{0, 1\}^3$. If each of the six inequalities

$$
F_{x\bar{y}z}F_{\bar{x}y\bar{z}} < F_{x\bar{y}y}F_{\bar{x}\bar{y}\bar{z}}, \quad G_{\bar{x}y\bar{z}}G_{x\bar{y}z} < G_{x\bar{y}y}G_{\bar{x}\bar{y}\bar{z}},
$$

$$
F_{x\bar{y}z}F_{\bar{x}y\bar{z}} < F_{x\bar{y}y}F_{\bar{x}\bar{y}\bar{z}}, \quad G_{\bar{x}y\bar{z}}G_{x\bar{y}z} < G_{x\bar{y}y}G_{\bar{x}\bar{y}\bar{z}},
$$

$$
F_{x\bar{y}z}F_{\bar{x}y\bar{z}} < F_{x\bar{y}y}F_{\bar{x}\bar{y}\bar{z}}, \quad G_{\bar{x}y\bar{z}}G_{x\bar{y}z} < G_{x\bar{y}y}G_{\bar{x}\bar{y}\bar{z}}
$$

hold, then it is not possible for all three of the following three inequalities to hold:

$$(F_{x\bar{y}y} + G_{\bar{x}y\bar{z}})(F_{x\bar{y}y} + G_{x\bar{y}z}) > (F_{x\bar{y}z} + G_{\bar{x}y\bar{z}})(F_{x\bar{y}z} + G_{\bar{x}\bar{y}\bar{z}}),$$

$$(F_{x\bar{y}y} + G_{\bar{x}y\bar{z}})(F_{x\bar{y}y} + G_{x\bar{y}z}) > (F_{x\bar{y}z} + G_{\bar{x}y\bar{z}})(F_{x\bar{y}z} + G_{\bar{x}\bar{y}\bar{z}}),$$

$$(F_{x\bar{y}z} + G_{\bar{x}y\bar{z}})(F_{x\bar{y}z} + G_{x\bar{y}y}) > (F_{x\bar{y}y} + G_{\bar{x}y\bar{z}})(F_{x\bar{y}y} + G_{\bar{x}\bar{y}\bar{z}}).$$

Proof. If all nine inequalities hold, then by Lemma 5.1 we get the three logical conclusions

$$F_{\bar{x}y\bar{z}}G_{x\bar{y}y} > F_{x\bar{y}z}G_{\bar{x}y\bar{z}} \quad \text{and} \quad F_{x\bar{y}y}G_{x\bar{y}z} < F_{x\bar{y}z}G_{\bar{x}y\bar{z}} \quad \text{or} \quad F_{\bar{x}y\bar{z}}G_{x\bar{y}y} < F_{x\bar{y}z}G_{\bar{x}y\bar{z}} \quad \text{and} \quad F_{x\bar{y}y}G_{x\bar{y}z} > F_{x\bar{y}z}G_{\bar{x}y\bar{z}}, \quad (1)$$

$$F_{\bar{x}y\bar{z}}G_{x\bar{y}y} > F_{x\bar{y}z}G_{\bar{x}y\bar{z}} \quad \text{and} \quad F_{x\bar{y}y}G_{x\bar{y}z} < F_{x\bar{y}z}G_{\bar{x}y\bar{z}} \quad \text{or} \quad F_{\bar{x}y\bar{z}}G_{x\bar{y}y} < F_{x\bar{y}z}G_{\bar{x}y\bar{z}} \quad \text{and} \quad F_{x\bar{y}y}G_{x\bar{y}z} > F_{x\bar{y}z}G_{\bar{x}y\bar{z}}, \quad (2)$$

$$F_{x\bar{y}z}G_{x\bar{y}y} > F_{x\bar{y}z}G_{\bar{x}y\bar{z}} \quad \text{and} \quad F_{x\bar{y}y}G_{x\bar{y}z} < F_{x\bar{y}z}G_{\bar{x}y\bar{z}} \quad \text{or} \quad F_{x\bar{y}z}G_{x\bar{y}y} < F_{x\bar{y}z}G_{\bar{x}y\bar{z}} \quad \text{and} \quad F_{x\bar{y}y}G_{x\bar{y}z} > F_{x\bar{y}z}G_{\bar{x}y\bar{z}}. \quad (3)$$

Note that if the first conjunction in (1) is true, then the first conjunction of (3) cannot hold, so the second conjunction of (3) must be true. This in turn implies that the second conjunction of (2) is true, which is a direct contradiction with the first conjunction of (1). We get a similar contradiction if we assume the second conjunction of (1) to be true. □

From Lemma 5.2 we can directly draw the following conclusions.

Corollary 5.3. There are no Simpson conversions from a triangulation of type I or II to a triangulation of type IV. □

Corollary 5.4. There are no Simpson conversions from a triangulation of type VI to a triangulation of type I or II. □

There is a parity argument in the proof of Lemma 5.2 that relies on the assumption that there are three face diagonals emanating from the vertex $xyz$. This can be extended to the case where there is exactly one face diagonal emanating from $xyz$. In particular, if a table $F$ has exactly one diagonal incident with vertex $xyz$ and another table $G$ also has only that same face diagonal incident with vertex $xyz$, then the sum $F + G$ cannot have all three face diagonals on the sides incident with $xyz$ different from those in $F$ and $G$. This is the content of the next lemma.
Among the 167 symmetry classes of ordered pairs, 55 satisfy the hypotheses of Lemma 5.2. This is very similar to the proof of Lemma 5.2.

**Lemma 5.5.** Let \( F, G : \{0, 1\}^3 \to \mathbb{R}_+ \) and \((x, y, z) \in \{0, 1\}^3\). If each of the six inequalities
\[
F_{\bar{x}y\bar{z}}F_{\bar{x}y\bar{z}} < F_{xy\bar{z}}F_{\bar{x}y\bar{z}}, \quad G_{\bar{x}y\bar{z}}G_{x\bar{y}\bar{z}} < G_{xy\bar{z}}G_{\bar{x}y\bar{z}},
\]
\[
F_{\bar{x}y\bar{z}}F_{xy\bar{z}} > F_{xy\bar{z}}F_{\bar{x}y\bar{z}}, \quad G_{\bar{x}y\bar{z}}G_{xy\bar{z}} > G_{xy\bar{z}}G_{\bar{x}y\bar{z}},
\]
\[
F_{\bar{x}y\bar{z}}F_{xy\bar{z}} > F_{\bar{x}y\bar{z}}F_{xy\bar{z}}, \quad G_{\bar{x}y\bar{z}}G_{xy\bar{z}} > G_{xy\bar{z}}G_{\bar{x}y\bar{z}}
\]
hold, then it is not possible for all three of the following three inequalities to hold:
\[
(F_{\bar{x}y\bar{z}} + G_{\bar{x}y\bar{z}})(F_{\bar{x}y\bar{z}} + G_{\bar{x}y\bar{z}}) > (F_{xy\bar{z}} + G_{xy\bar{z}})(F_{\bar{x}y\bar{z}} + G_{\bar{x}y\bar{z}}),
\]
\[
(F_{\bar{x}y\bar{z}} + G_{\bar{x}y\bar{z}})(F_{\bar{x}y\bar{z}} + G_{\bar{x}y\bar{z}}) < (F_{xy\bar{z}} + G_{xy\bar{z}})(F_{\bar{x}y\bar{z}} + G_{\bar{x}y\bar{z}}),
\]
\[
(F_{x\bar{y}\bar{z}} + G_{x\bar{y}\bar{z}})(F_{xy\bar{z}} + G_{xy\bar{z}}) < (F_{xy\bar{z}} + G_{xy\bar{z}})(F_{x\bar{y}\bar{z}} + G_{x\bar{y}\bar{z}}).
\]

**Proof.** This is very similar to the proof of Lemma 5.2. \(\square\)

**5B. Infeasible 3-dimensional Simpson conversions.** We are now ready to describe the pairs of triangulations \(A\) and \(B\) for which there are no Simpson conversions from \(A\) to \(B\). For brevity, we will only list examples of such pairs up to symmetry. There are 5476 ordered pairs of triangulations, but only 167 up to symmetry.\(^1\) Among the 167 symmetry classes of ordered pairs, 55 satisfy the hypotheses of Lemmas 5.2 and 5.5 and therefore cannot give rise to Simpson conversions. Representatives from each of those symmetry classes of ordered pairs are listed in Table 3.

**Example 5.6.** To see how the content of Table 3 can be used in a practical setting, suppose we measure the presence of a certain mutation at three places in a genome from several individuals and record the data. We can determine which triangulation the data corresponds to by computing the linear forms in Appendix A. If we find that \(e < 0\), \(f < 0\), \(j < 0\), and \(l < 0\), then the association corresponds to triangulation 35 in Appendix B, which means we can conclude (directly from Table 3) that the set of individuals cannot be subdivided into two sets, both of which have association corresponding to triangulation 1 or triangulation 3. To get the full set of infeasible subdivisions we would also have to look at triangulations 36 to 46, which are all symmetric to 35. For instance, because triangulation 43 is a

\(^1\)We calculated the number of distinct ordered pairs of triangulations of the cube up to symmetry by generating the list of all \(74^2 = 5476\) ordered pairs and partitioning it into equivalence classes with respect to the relation that two pairs, \((A, B)\) and \((A', B')\), of triangulations of the cube are equivalent if there exists a symmetry of the cube that simultaneously maps \(A\) to \(A'\) and \(B\) to \(B'\).
90-degree rotation about the vertical axis of triangulation 35, the set of individuals cannot be subdivided into two sets, each having association corresponding to triangulation 34, which is a 90-degree rotation (in the opposite direction) of triangulation 11.

The easiest way to conduct this type of inference, however, is to identify empty vertices and apply Lemma 5.2. For instance, the empty vertex at 111 in triangulation 50 implies that the corresponding table cannot be the sum of any two subtables corresponding to a triangulation with a full vertex at 111 (and similarly for vertex 000). One may also apply Lemma 5.5 in a similar manner.

5C. Feasible 3-dimensional Simpson conversions. We searched for explicit instances of Simpson conversions for each of the remaining 112 symmetry classes by sampling $2 \times 2 \times 2$ contingency tables uniformly from the probability simplex in $\mathbb{R}^8$ using a method proposed in [7, Section 2]. For each representative pair of triangulations, $A$ and $B$, we randomly generated $2 \times 2 \times 2$ contingency tables until we found two that induce triangulation $A$: if the sum of those tables induced triangulation $B$, we recorded the instance of Simpson conversion; otherwise, we repeated the process. Following that approach, we found explicit instances for all 112 cases. The feasible Simpson conversions (up to symmetry) are listed in Table 4.

Remarkably, Lemmas 5.2 and 5.5 characterise all the cases for which no Simpson conversions can occur, as can be seen in Tables 3 and 4. We have thus established the following result.

Theorem 5.7. Let $A$ and $B$ be triangulations of the cube. There exists a Simpson conversion from $A$ to $B$ if and only if there is no vertex of the cube that is incident to an odd number of face diagonals in $A$ and the opposite set of face diagonals in $B$. □

The Python script used to generate the data in Tables 3 and 4 is available at http://www.mattstamps.com/simpson/supplementary.zip.
6. Additional computations and future work

We conclude with a conjecture and some observations from additional computational experiments, along with some open-ended problems/questions that merit further exploration.

6A. Frequency of Simpson conversion. In addition to there being many possible ways in which Simpson conversions can occur, Simpson conversions appear to occur somewhat frequently. For the 2-dimensional case, [7] experimentally verified — and presented a proof by Hadjicostas — that the probability of a Simpson reversal occurring is $\frac{1}{60}$. More precisely, they showed that if 

$$\begin{bmatrix}
p_{ijk} | i, j, k = 0, 1, & p_{ijk} \geq 0, & \text{and} \sum p_{ijk} = 1
\end{bmatrix}$$

is a random $2 \times 2 \times 2$ table (i.e., pairs of $2 \times 2$ tables) from the uniform distribution on the probability simplex in $\mathbb{R}^8$, then the probability that the $2 \times 2$ subtables

$$[p_{ij0} | i, j = 0, 1] \quad \text{and} \quad [p_{ij1} | i, j = 0, 1]$$

both exhibit positive (or negative) associations while their sum

$$[p_{ij0} + p_{ij1} | i, j = 0, 1]$$

exhibits a negative (resp. positive) association is $\frac{1}{60}$.

To estimate the analogous probability for the 3-dimensional case, we sampled $2 \times 2 \times 2 \times 2$ tables (i.e., pairs of $2 \times 2 \times 2$ tables) of the form

$$\begin{bmatrix}
p_{ijkl} | i, j, k, \ell = 0, 1, & p_{ijkl} \geq 0, & \text{and} \sum p_{ijkl} = 1
\end{bmatrix}$$

uniformly from the probability simplex in $\mathbb{R}^{16}$ and calculated the proportion of tables for which the $2 \times 2 \times 2$ subtables

$$[p_{ijkl0} | i, j, k = 0, 1] \quad \text{and} \quad [p_{ijkl1} | i, j, k = 0, 1]$$

induce the same triangulation of the cube (or, equivalently, satisfy identical sets of mutual, marginal, and conditional associations), while their sum

$$[p_{ijkl0} + p_{ijkl1} | i, j, k = 0, 1]$$

induces a different triangulation of the cube.

**Conjecture 6.1.** The probability that a Simpson conversion occurs in the context above is $\frac{1}{430}$. 

For this estimation, we used the same Dirichlet distribution method as [7, Section 2] for uniformly sampling the probability simplex in $\mathbb{R}^n$, namely by generating the entries of each table independently according to the Gamma$(1, 1)$ distribution and normalising. We generated five million $2 \times 2 \times 2 \times 2$ contingency tables and recorded the percentage of tables that decomposed into two $2 \times 2 \times 2$ contingency tables that induced a common triangulation of the cube, the percentage of those tables that exhibited Simpson conversions, and the percentage of those tables that did not. We repeated this experiment one hundred times, which resulted in 95% confidence intervals $1.888 \pm 0.012\%$, $0.223 \pm 0.004\%$, and
1.664 ± 0.012%, respectively, which led us to conjecture that the exact values are $\frac{17}{900}$, $\frac{2}{900}$, and $\frac{15}{900}$. The Python script for the experiment is available at the website listed in Section 5C.

6B. Generalised Simpson conversions. We also considered the following, more general, version of the main question from Section 5: for which triangulations $A$, $B$, and $C$ with $A \neq B$ is it possible for the sum of a contingency table that induces triangulation $A$ and a contingency table that induces triangulation $B$ to induce triangulation $C$? Just as before, it is not difficult to find examples of such triples (for instance, in Table 2, the contingency tables for the Northern, Southern, and all states induce triangulations 19, 30, and 35, respectively), but Lemmas 5.2 and 5.5 still imply that it is not possible to find such examples for every triple of triangulations. The 199,874 triples of triangulations can be partitioned into 4,655 equivalence classes based on the symmetries of the cube.\(^2\) Of those 4,655 symmetry classes, 351 cannot occur because of Lemmas 5.2 and 5.5. Using the same random search technique as the one described in Section 5C, we have found specific instances for 4,287 the remaining 4,304 cases, leaving 17 unaccounted-for triples. Since it is not clear whether these remaining triples are infeasible (for reasons other than Lemmas 5.2 and 5.5) or simply very rare, we are continuing to search for specific instances and maintaining an up-to-date spreadsheet of known instances with the supplementary documents at http://www.mattstamps.com/simpson/feasible-triples.csv.

6C. Generalisation to higher dimensions and further exploration. The geometric approach for studying associations among 3-way contingency tables presented in this paper raises a number of questions that merit further exploration. For instance, what do the 74 nonconvex regions in the 8-dimensional probability simplex that correspond to the triangulations of the cube (described in Section 4) look like? How does their geometry relate to that of the regions cut out by the independence hypersurfaces corresponding to the various 3-way (conditional, marginal, and mutual) associations? Are there formulas for the volumes of these regions that could shed light on the probability of a particular Simpson conversion occurring? We thank the anonymous referees for suggesting that we state these questions explicitly.

There is also the question of how the methods in this paper might be extended to higher dimensions. While a straightforward modification of the technique described in Sections 2 and 3 allows one to map a triangulation of the $k$-dimensional cube to each binary $k$-way contingency table, it is not clear what correspondence should exist between multiway associations and triangulations of high-dimensional cubes. One could investigate the geometry of the log-linear hypersurfaces in $\mathbb{R}^{2^k}$ that separate the regions corresponding to different triangulations of the $k$-dimensional cube induced by binary $k$-way contingency tables independently of statistical implications, but the triangulations of high-dimensional cubes are not well understood in general. For instance, the number of different triangulations of a cube grows very rapidly with respect to dimension, even though we only need to consider regular triangulations. It would thus be difficult to produce complete lists of feasible Simpson conversions in dimensions higher than 3. It could, however, be interesting to find generalisations of Lemmas 5.2 and 5.5.

\(^2\)We calculated the number of distinct triples of triangulations of the cube up to symmetry by generating the list of all $\binom{74}{2} \cdot 74 = 199,874$ triples and partitioning it into equivalence classes with respect to the relation that a pair of triples, $(A, B, C)$ and $(A', B', C')$, of triangulations are equivalent if there exists a symmetry of the cube that simultaneously maps $A$ to $A'$, $B$ to $B'$, and $C$ to $C'$. 
Appendix A. Linear forms

The following list of linear forms, taken from [1], is used to determine the triangulation of the cube induced by each $2 \times 2 \times 2$ contingency table.

\[
\begin{align*}
a &:= f_{000} + f_{110} - f_{010} - f_{100}, \\
b &:= f_{001} + f_{111} - f_{011} - f_{101}, \\
c &:= f_{000} + f_{101} - f_{001} - f_{100}, \\
d &:= f_{010} + f_{111} - f_{011} - f_{101}, \\
e &:= f_{000} + f_{111} - f_{001} - f_{110}, \\
f &:= f_{100} + f_{111} - f_{101} - f_{110}, \\
g &:= f_{000} + f_{111} - f_{011} - f_{100}, \\
 &:= f_{001} + f_{101} + f_{010} - f_{111}, \\
i &:= f_{000} + f_{111} - f_{010} - f_{101}, \\
 &:= f_{010} + f_{111} - f_{001} - f_{110}, \\
o &:= f_{010} + f_{101} + f_{111} - f_{110}, \\
p &:= f_{011} + f_{110} + f_{010} - f_{110}, \\
q &:= f_{011} + f_{100} + f_{011} - f_{101}, \\
r &:= f_{110} + f_{011} + f_{000} - f_{101}, \\
s &:= f_{101} + f_{110} + f_{000} - f_{011}, \\
t &:= f_{010} + f_{001} + f_{111} - f_{100}.
\end{align*}
\]

Appendix B: Triangulations

The following chart comprises the 74 triangulations of the 3-dimensional cube. The numbering is taken from the paper [1] for convenience, but we have chosen to list them in a slightly different order to enhance some similarities.\(^3\) The diagonals on the surface of the cube are blue and the interior diagonals are red.

---

\(^3\)We discovered two errors in [1, Table 5.1]: in the row beginning with 57/5, the 43 should be replaced with 42 and, in the row beginning with 63/5, the 42 should be replaced with 43.
ASSOCIATION AND SIMPSON CONVERSION IN $2 \times 2 \times 2$ CONTINGENCY TABLES
Acknowledgements

We are grateful to Lior Pachter for asking this question and Soo Go for several helpful suggestions that increased the efficiency and speed of our computations. We would also like to thank the anonymous referees for the many comments and suggestions that helped us improve the overall quality of this article. This work was supported in part by grants 2014-4780 and 2018-05218 from the Swedish Science Council, National Science Foundation grant 1159206, and Yale-NUS College grant R-607-265-246-121.
References


Received 2020-05-22. Revised 2021-01-27. Accepted 2021-02-22.

SVANTE LINUSSON: linusson@math.kth.se
Department of Mathematics, KTH Royal Institute of Technology, Stockholm, Sweden

MATTHEW T. STAMPS: matthewstamps@gmail.com
Ottawa, ON, Canada
LOGARITHMIC VORONOI CELLS

YULIA ALEXANDR AND ALEXANDER HEATON

We study Voronoi cells in the statistical setting by considering preimages of the maximum likelihood estimator that tessellate an open probability simplex. In general, logarithmic Voronoi cells are convex sets. However, for certain algebraic models, namely finite models, models with ML degree 1, linear models, and log-linear (or toric) models, we show that logarithmic Voronoi cells are polytopes. As a corollary, the algebraic moment map has polytopes for both its fibers and its image, when restricted to the simplex. We also compute nonpolytopal logarithmic Voronoi cells using numerical algebraic geometry. Finally, we determine logarithmic Voronoi polytopes for the finite model consisting of all empirical distributions of a fixed sample size. These polytopes are dual to the logarithmic root polytopes of Lie type A, and we characterize their faces.

1. Introduction

For any subset $X \subset \mathbb{R}^n$, the Voronoi cell of a point $p \in X$ consists of all points of $\mathbb{R}^n$ which are closer to $p$ than to any other point of $X$ in the Euclidean metric. In this article we discuss the analogous logarithmic Voronoi cells which find application in statistics. A discrete statistical model is a subset of the probability simplex $M \subset \Delta_{n-1}$, since probabilities are positive and sum to 1. The maximum likelihood estimator $\Phi$ (MLE) sends an empirical distribution $u \in \Delta_{n-1}$ of observed data to the point in the model which best explains the data. This means $p = \Phi(u)$ maximizes the log-likelihood function $\ell_u(p) := \sum_{i=1}^n u_i \log(p_i)$ restricted to $M$. Note that $\ell_u$ is strictly concave on $\Delta_{n-1}$ and takes its maximum value at $u$. Usually, $u \notin M$, and we must find the point $\Phi(u) \in M$ which is closest in the log-likelihood sense. For $p \in M$ we define the logarithmic Voronoi cell

$$\text{log Vor}_M(p) := \{ u \in \Delta_{n-1} : \Phi(u) = p \}.$$ 

Information geometry [6] considers the MLE in the context of the Kullback–Leibler divergence of probability distributions, sending data to the nearest point with respect to a Riemannian metric on $\Delta_{n-1}$. Algebraic statistics [14] considers the case where $M$ can be described as either the image or kernel of algebraic maps. Recent work in metric algebraic geometry [10; 12; 13; 24] concerns the properties of real algebraic varieties that depend on a distance metric. Logarithmic Voronoi cells are natural objects of interest in all three subjects.

Keywords: Voronoi cell, logarithmic Voronoi cell, MLE, numerical algebraic geometry, statistics, probability simplex, root polytopes, algebraic moment map, ML degree, statistical model.
As an example, consider flipping a biased coin three times. There are four possible outcomes, three heads (hhh), two heads (hht,htth,thh), one head (htt,tht,tth), and zero heads (ttt). Parametrically, the twisted cubic is given by
\[ t \mapsto p(t) = (t^3, 3t^2(1-t), 3t(1-t)^2, (1-t)^3) \in M. \]

For this model’s many lives, see [30]. We compute logarithmic Voronoi cells \( \log \text{Vor}_M(p(t)) \) with parameter values
\[ t \in \left\{ \frac{1}{25}, \frac{2}{25}, \ldots, \frac{24}{25} \right\} \]
which live inside the simplex \( \Delta_3 \subset \mathbb{R}^4 \), and whose orthogonal projections into 3-space are shown in Figure 1. In this case, the logarithmic Voronoi cells are polytopes, and we get both triangles and quadrilaterals, depending on the point \( p(t) \in M \). The fact that these polytopes are equal to the logarithmic Voronoi cells will follow from Theorem 11 below.

After giving the basic definitions in Section 2, Section 3 describes the relationship between logarithmic Voronoi cells and logarithmic polytopes in the context of algebraic statistics. In particular, we show that ML degree 1 implies that the logarithmic Voronoi cells are polytopes, and we give counterexamples to the converse statement. We also consider both linear models and log-linear (toric) models, showing that both families of statistical models have the property that logarithmic Voronoi cells are polytopes. These include the twisted cubic of Figure 1, decomposable graphical models [29], Bayesian networks [18], staged tree models [11; 35], multinomial distributions, phylogenetic models, hidden Markov models, and many others arising in applications [33]. Corollary 12 states that both the image and fibers of the algebraic moment map are polytopes. In Section 4 we show how to compute a (not necessarily polytopal) logarithmic Voronoi cell using numerical algebraic geometry. By calculating \( \Phi(u) \) for 60,000 points \( u \) with respect to a model of ML degree 39, we demonstrate that logarithmic Voronoi cells can be reliably computed using numerical methods. Finally, in Section 5 we discuss the historical motivation of Georgy Voronoi and adapt it to the statistical setting by analyzing a model with finitely many points, namely all possible empirical distributions on \( n \) states with \( d \) trials. We call the polytopes that arise logarithmic root polytopes of type \( A_{n-1} \), show they are dual to the logarithmic Voronoi cells in Theorem 19, and characterize their faces in Theorem 17.
2. Preliminaries

We work with the open probability simplex $\Delta_{n-1} \subset \mathbb{R}^n$ defined by

$$\Delta_{n-1} \coloneqq \{ u \in \mathbb{R}^n : \sum_{i=1}^n u_i = 1, \; u_i > 0 \text{ for all } i \in [n] \}.$$ 

A statistical model $\mathcal{M}$ is a subset of the probability simplex. When $\mathcal{M}$ is defined as the intersection of $\Delta_{n-1}$ with an algebraic variety or the image of a rational map, we say that $\mathcal{M}$ is an algebraic statistical model [33; 38]. For any point $u \in \Delta_{n-1}$, the log-likelihood function $\ell_u : \mathbb{R}_{>0}^n \to \mathbb{R}$ is defined by $\ell_u(p) = \sum_{i=1}^n u_i \log(p_i)$. For any model $\mathcal{M} \subset \Delta_{n-1}$, we define the relation $\Phi \subset \Delta_{n-1} \times \mathcal{M}$ by

$$(u, p) \in \Phi \iff p = \arg\max_{q \in \mathcal{M}} \{ \ell_u(q) : q \in \mathcal{M} \}.$$ 

If $(u, p) \in \Phi$, then we also write $\Phi(u) = p$. We write $\Delta_{n-1}^\mathcal{M}$ for the set of $u \in \Delta_{n-1}$ such that $\Phi(u)$ exists. Describing the set $\Delta_{n-1}^\mathcal{M}$ and how it extends to the boundary of $\Delta_{n-1}$ is an active area of research, especially with respect to zeros in the data [17; 22]. MLE existence is also connected to polystable and stable orbits in invariant theory [4]. For the important family of log-linear (toric) models, [16] shows that positive data $u \in \Delta_{n-1}$ guarantees existence, and in general the MLE exists exactly when the observed margins belong to the relative interior of a certain polytope. See also [38, Theorem 8.2.1].

Finally, we note that for models with more complicated geometry, $\Phi(u)$ cannot always be computed by finding critical points of $\ell_u$ restricted to manifold points of $\mathcal{M}$. The present article takes the first step of computing logarithmic Voronoi cells for models where critical points of $\ell_u$ succeed in finding the MLE, as well as some interesting finite models. More complicated examples outside the scope of the present article include models of nonnegative rank-$r$ matrices, which were studied in [28].

Whenever $p \in \mathcal{M} \subset \mathbb{R}^n$ admits a tangent space at the point $p$, we denote by $N_p \mathcal{M}$ its orthogonal complement with respect to the Euclidean inner product on $\mathbb{R}^n$. We are also interested in the log-normal space at the point $p \in \mathcal{M}$, defined by

$$\log N_p \mathcal{M} \coloneqq \{ u \in \mathbb{R}^n : \nabla \ell_u(p) \in N_p \mathcal{M} \}.$$ 

Here, $\nabla \ell_u(p)$ is the vector whose entries are given by the partial derivatives of $\ell_u$ with respect to each of the variables $p_1, \ldots, p_n$. We are interested in the log-normal space since, in many cases, it will contain the logarithmic Voronoi cell. In Section 3 we will see several different situations where the logarithmic Voronoi cell is equal to the intersection of the log-normal space with the probability simplex.

**Lemma 1.** The log-normal space $\log N_p \mathcal{M}$ is a linear subspace of $\mathbb{R}^n$.

**Proof.** The normal space $N_p \mathcal{M}$ is a linear subspace. Arrange a basis of the normal space as the rows of a matrix. Adjoin another row with entries $u_i/p_i$, the partial derivatives of $\ell_u(p)$ with respect to each $p_i$. The maximal minors of the resulting matrix are linear equations in the variables $u_i$ and therefore cut out a linear space of such $u \in \mathbb{R}^n$. This space is the log-normal space at $p$. \hfill \Box

By Lemma 1, the intersection of the log-normal space at a point $p \in \mathcal{M}$ with the closed probability simplex $\overline{\Delta}_{n-1}$ is a polytope $\log \text{Poly}_\mathcal{M}(p)$, which we call its log-normal polytope. In what follows, when
we say that a logarithmic Voronoi cell equals its log-normal polytope, we mean that they are equal as sets, excepting the points in the boundary of the simplex.

3. Logarithmic Voronoi cells and polytopes

In this section we will prove several basic results on logarithmic Voronoi cells. The main focus is to explore criteria which ensure that the logarithmic Voronoi cells are polytopes, rather than more general convex sets, as in Proposition 4 below. The section ends with a list of five open questions we find interesting for future research.

Proposition 2. Let $M$ be any finite statistical model. Then the logarithmic Voronoi cells $\text{log Vor}_M(p)$ are polytopes for each $p \in M$.

Proof. Fix $p \in M$. The set of all points $u \in \Delta_{n-1}$ such that $\ell_u(p) \geq \ell_u(q)$ for all $q \in M$ is the logarithmic Voronoi cell of $p$. For any $q \in M$, $\ell_u(p) \geq \ell_u(q)$ becomes the condition that

$$\sum_{i=1}^{n} u_i \log \left( \frac{p_i}{q_i} \right) \geq 0.$$ 

But this is linear in $u$ and so defines a closed half-space. Since there are finitely many points in $M$, we see that the logarithmic Voronoi cell is an intersection of finitely many closed half-spaces (including those defining $\Delta_{n-1}$). Therefore it is a polytope. \qed

For infinite models, the logarithmic Voronoi cells are, in general, not polytopes. However, if the model is smooth at $p$, the logarithmic Voronoi cell will be contained in the log-normal polytope. Figure 2 shows a logarithmic Voronoi cell for $p \in M \subset \Delta_5 \subset \mathbb{R}^6$ which is not a polytope, but is contained in a polytope. In this case it is the hexagon given by $\text{log Poly}(p) = \text{log } N_{p,M} \cap \Delta_5$. Since the log-normal space is 2-dimensional, by choosing an orthonormal basis agreeing with this subspace we can visualize the logarithmic Voronoi cell, despite it living in $\mathbb{R}^6$. We discuss this example in detail in Section 4. For more on finite models, see Section 5.

**Figure 2.** Logarithmic Voronoi cell (green) inside its log-normal polytope (pink) for a given point (yellow) in the model from Example 15. The 60,000 green and pink dots were computed using numerical algebraic geometry, as explained in Section 4.
Lemma 3. Let $\Phi(u) = p$ for some $p \in \mathcal{M} \subset \Delta_{n-1}$ such that $U \cap \mathcal{M}$ is a manifold for some $p$-neighborhood $U$ in $\mathbb{R}^n$. Then $u$ lies in the logarithmic normal space $\log N_p \mathcal{M}$ and

$$\log \text{Vor}_{\mathcal{M}}(p) \subset \log \text{Poly}_{\mathcal{M}}(p).$$

Proof. Note that $\ell_u(x) := \sum u_i \log(x_i)$ is a smooth function on any neighborhood of $p \in \mathcal{M}$ contained in $\Delta_{n-1}$. Consider the gradient $\nabla \ell_u(p)$. Note that $\mathbb{R}^n = T_p \mathcal{M} \oplus N_p \mathcal{M}$ and if $\nabla \ell_u(p)$ had any nonzero tangential component then there would exist some $q \in \mathcal{M}$ such that $\ell_u(q) > \ell_u(p)$, contradicting the fact that $\Phi(u) = p$. \hfill \Box

Proposition 4. Logarithmic Voronoi cells are convex sets.

Proof. As in the proof of Proposition 2, the logarithmic Voronoi cell of $p$ is defined by the inequalities $\sum_{i \in [n]} u_i \log(p_i/q_i) \geq 0$ for every $q \in \mathcal{M}$, each linear in $u$. Hence, the logarithmic Voronoi cell of $p$ is an intersection of (possibly infinitely many) closed half-spaces, and the result follows. \hfill \Box

For an algebraic statistical model $\mathcal{M}$, the **ML degree** is the number of complex critical points of $\ell_u$ on the Zariski closure of $\mathcal{M}$ for generic data $u \in \Delta_{n-1}$ [38, p. 140]. The following theorem concerns algebraic models with ML degree 1. These were characterized in [25] and studied further in [15]. They include, for example, Bayesian networks and decomposable graphical models.

Theorem 5. Let $\mathcal{M}$ be any algebraic model with ML degree 1 which is smooth on $\Delta_{n-1}$. Then the logarithmic Voronoi cell at every $p \in \mathcal{M}$ equals its log-normal polytope on $\Delta_{n-1}$.

Proof. We will show that $\log \text{Vor}_{\mathcal{M}}(p) = \log N_p \mathcal{M} \cap \Delta_{n-1}^\mathcal{M}$. Let $u \in \Delta_{n-1}$ be an element of $\log \text{Vor}_{\mathcal{M}}(p)$. Then $\Phi(u) = p$ and since $\mathcal{M}$ is smooth, $u \in \log N_p \mathcal{M} \cap \Delta_{n-1}^\mathcal{M}$ by Lemma 3. For the reverse direction, let $u \in \log N_p \mathcal{M} \cap \Delta_{n-1}^\mathcal{M}$. Recall that $\Phi(u)$ is the argmax of $\ell_u(q)$ over all points $q \in \mathcal{M}$. Since $\Phi(u)$ exists and $\mathcal{M}$ is smooth, this argmax must be among the critical points of $\ell_u$ restricted to $\mathcal{M}$, which include $p$. But since the ML degree is 1, there is only one complex critical point, and hence $\Phi(u) = p$. Therefore $u$ is in the logarithmic Voronoi cell of $p$, and the result follows. \hfill \Box

Example 6. Consider $\mathcal{M} = V(f)$ for $f : \mathbb{C}^4 \rightarrow \mathbb{C}^2$ given by the polynomial system

$$f(x) = \begin{bmatrix} x_1 x_4 - x_2 x_3 \\ x_1 + x_2 + x_3 + x_4 - 1 \end{bmatrix} : \mathbb{C}^4 \rightarrow \mathbb{C}^2.$$ 

A parametrization of this model is given by

$$(p_1, p_2) \mapsto (p_1 p_2, p_1(1-p_2), (1 - p_1)p_2, (1 - p_1)(1 - p_2)).$$

This is the **independence model** on two binary random variables, and also the Segre embedding of $\mathbb{P}^1 \times \mathbb{P}^1$. The points of this 2-dimensional model live in the 3-dimensional hyperplane $\sum x_i = 1$ inside $\mathbb{R}^4$, so we can choose a basis agreeing with this hyperplane to plot them.

For each $x \in \mathcal{M}$, we construct an $(m+1) \times n$ matrix $A(x)$ by augmenting the row $\nabla \ell_u$ to the Jacobian matrix $df$:

$$A(x) = \begin{bmatrix} x_4 & -x_3 & -x_2 & x_1 \\ 1 & 1 & 1 & 1 \\ u_1/x_1 & u_2/x_2 & u_3/x_3 & u_4/x_4 \end{bmatrix}.$$
Since our model has codimension 2, the $3 \times 3$ minors of $A(x)$ give linear equations describing the log-normal space:

$$
\begin{align*}
&u_2 - u_3 - \frac{u_1 x_2}{x_1} + \frac{u_1 x_3}{x_1} + \frac{u_2 x_4}{x_2} - \frac{u_3 x_4}{x_3}, \\
&u_1 - u_4 - \frac{u_2 x_1}{x_2} + \frac{u_1 x_3}{x_1} - \frac{u_4 x_3}{x_4} + \frac{u_2 x_4}{x_2}, \\
&u_1 - u_4 + \frac{u_1 x_2}{x_1} + \frac{u_3 x_1}{x_3} + \frac{u_4 x_2}{x_4} + \frac{u_3 x_4}{x_3}, \\
&u_2 - u_3 + \frac{u_2 x_1}{x_2} - \frac{u_3 x_1}{x_3} - \frac{u_4 x_2}{x_4} + \frac{u_4 x_3}{x_4}.
\end{align*}
$$

Restricting this space to its intersection with the simplex $u_1 + u_2 + u_3 + u_4 - 1 = 0$ to compute the log-normal polytope, we find that the polytopes are line segments. We plot them for various points on the model in Figure 3. Since $M$ has ML degree 1, Theorem 5 tells us that log-Voronoi cells equal log-normal polytopes, so they are also line segments.

Example 7 below shows that the ML degree 1 condition in Theorem 5 is sufficient, but not necessary, for the equality of logarithmic Voronoi cells with the interior of their log-normal polytopes. Consider the independence model of two identically distributed binary random variables. The natural parametrization in a statistical context leads to the Hardy–Weinberg curve defined by $x_2^2 - 4x_1x_3$, which has ML degree 1 [26]. A similar-looking model, which has been called the cousin of the Hardy–Weinberg curve [23], is defined by the polynomial $f = x_2^2 - x_1x_3$. It turns out that the ML degree of this model is 2 [23, p. 394]. It was demonstrated in [23] that ML degree is extremely sensitive to scaling of the coordinates, so the difference between the ML degrees of the Hardy–Weinberg curve and its cousin is not surprising. The effect of scaling on the ML degree of toric varieties has been studied in [3].

**Example 7.** The algebraic model defined by the polynomial $f = x_2^2 - x_1x_3$ has ML degree 2, yet the logarithmic Voronoi cells are equal to their log-normal polytopes. Although this follows from the later Theorem 11, we will first prove it more explicitly. Calculate the Jacobian matrix of Lemma 1 by taking the gradients of $f = x_2^2 - x_1x_3$ and $g = x_1 + x_2 + x_3 - 1$, augmenting this matrix with an additional row of the $u_i/x_i$. Consider the equation of the plane given by the determinant of this matrix. Note that $M$ is a curve in $\Delta_2$, so the log-normal space at each point is
defined by the vanishing of the determinant at that point. This plane has normal vector given by

\[
\begin{pmatrix}
2x_1x_2^2 - x_1^2x_3 + x_1x_2x_3 + x_1x_3^2 \\
-2x_1x_2^2 - 2x_1x_2x_3 - 2x_2^2x_3 \\
x_2^2x_3 + x_1x_2x_3 + 2x_2^2x_3 - x_1x_3^2
\end{pmatrix}
\]

where \((x_1, x_2, x_3)\) is any point in the common zero set of \(f\) and \(g\). Consider the cross-product of this vector with the all-ones vector, which will give us the direction vector of the log-normal polytope at \((x_1, x_2, x_3)\). Computing and simplifying each coordinate in the quotient ring

\[
\mathbb{Q}[x_1, x_2, x_3]/(x_1 + x_2 + x_3 - 1, -x_2^2 + x_1x_3)\mathbb{Q}[x_1, x_2, x_3],
\]

we find that this cross product is given by

\[
\begin{pmatrix}
-(x_2 + x_3 - 1)x_3 \\
2(x_2 + x_3 - 1)x_3 \\
-(x_2 + x_3 - 1)x_3
\end{pmatrix}
= x_1x_3 \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}.
\]

This means that regardless of the point on the curve, the log-normal polytopes will be line segments whose direction vector is \((-1, 2, -1)\). We claim that for any distinct \(p, q \in M\) the corresponding line segments are disjoint. Consider the tangent space at some point \(x\) in the intersection of \(\Delta_2\) and the common zero set of \(f\) and \(g\). Applying Gaussian elimination to the \(2 \times 3\) Jacobian matrix, it can be shown that if \(2x_2 + x_3 \neq 0\) then all tangent vectors are multiples of

\[
\begin{pmatrix}
x_3 - x_1 \\
2x_2 + x_3 - 1 \\
x_1 - x_3
\end{pmatrix},
\]

while if \(2x_2 + x_3 = 0\) then all tangent vectors are multiples of \((-1, 1, 0)\). In neither case is it possible that a tangent vector is parallel to \((1, -2, 1)\). For \((-1, 1, 0)\) this is obvious, but for \((1)\), a contradiction can be derived by showing that if the vector is parallel to \((1, -2, 1)\) the first and the last coordinates in \((1)\) are equal, forcing \(x_1 + 4x_2 + x_3 = 0\). But on \(\Delta_2\) all coordinates are positive. Thus no line parallel to \((1, -2, 1)\) meets the model in two distinct points. We conclude the log-normal polytopes are disjoint, and the result follows from Lemma 8 below.

**Lemma 8.** Let \(M\) be any model smooth on \(\Delta_{n-1}\). If all log-normal polytopes for each point \(p \in M\) are disjoint, then the logarithmic Voronoi cells equal log-normal polytopes on \(\Delta^M_{n-1}\).

**Proof.** We will show that \(\log \text{Vor}_M(p) = \log N_p, M \cap \Delta^M_{n-1}\). The \(\subset\) direction follows from Lemma 3. For the reverse direction, let \(u \in \log N_p, M \cap \Delta^M_{n-1}\). Recall that \(\Phi(u)\) is the argmax of \(\ell_u(q)\) over all points \(q \in M\). Since \(\Phi(u)\) exists and \(M\) is smooth, this argmax must be among the critical points of \(\ell_u\) restricted to \(M\), which include \(p\). If \(\Phi(u)\) were not equal to \(p\) then \(u\) would be in the intersection of \(\Delta_{n-1}\) with the log-normal space to the point \(\Phi(u) \in M\). But the log-normal polytopes were assumed to be disjoint by the hypothesis. Therefore \(\Phi(u) = p\), which means that \(u \in \log \text{Vor}_M(p)\), and the result follows. \(\square\)

Let \(f_1(\theta), \ldots, f_r(\theta)\) be nonzero linear polynomials in \(\theta\) such that \(\sum_{i=1}^r f_i(\theta) = 1\). Let \(\Theta\) be the set such that \(f_i(\theta) > 0\) for all \(\theta \in \Theta\) and suppose that \(\dim \Theta = d\). The model \(M = f(\Theta) \subseteq \Delta_{r-1}\) is called a
Figure 4. Nonlinear boundary arising from two disjoint linear models.

discrete linear model [38, p. 152]. Linear models appear in [33, Section 1.2]. An example is DiaNA’s model in Example 1.1 of [33].

Theorem 9. Let $\mathcal{M}$ be a linear model. Then the logarithmic Voronoi cells are equal to their log-normal polytopes.

Proof. We will show that $\text{log Vor}_\mathcal{M}(p) = \log N_p \mathcal{M} \cap \Delta_{n-1}$. The “⊂” direction follows from Lemma 3 since an affine linear subspace intersected with $\Delta_{n-1}$ is smooth. For the reverse direction, let $u \in \log N_p \mathcal{M} \cap \Delta_{n-1}$. We must show $\Phi(u) = p$. Since $\ell_u$ is strictly concave on $\Delta_{n-1}$, it is strictly concave when restricted to any convex subset, such as the affine-linear subspace $\mathcal{M}$. Therefore there is only one critical point. Since $\mathcal{M}$ is smooth, $u$ must be in the log-normal space of $\Phi(u)$, and so $\Phi(u)$ must be $p$. □

Example 10. From the above results, one might hope that finite unions of linear models would admit logarithmic Voronoi cells which are polytopes. However, this is not the case. The log-normal spaces from two disjoint linear models can meet in such a way that the boundary created on a logarithmic Voronoi cell is nonlinear. For an explicit example with two linear models, $\mathcal{M} := \mathcal{M}_1 \cup \mathcal{M}_2 \subset \Delta_3 \subset \mathbb{R}^4$, see Figure 4. Here

\[
\mathcal{M}_1 = \{(1-s)p_1 + sp_2 : s \in [0, 1] \subset \mathbb{R}\}, \quad p_1 = \left(\frac{1}{5}, \frac{1}{2}, \frac{3}{5}, 0\right), \quad q_1 = \left(\frac{1}{13}, \frac{9}{13}, \frac{3}{13}, 0\right),
\]
\[
\mathcal{M}_2 = \{(1-t)q_1 + tq_2 : t \in [0, 1] \subset \mathbb{R}\}, \quad p_2 = \left(\frac{1}{7}, \frac{3}{7}, 0, \frac{2}{7}\right), \quad q_2 = \left(\frac{4}{13}, \frac{4}{13}, 0, \frac{5}{13}\right).
\]

We sampled 3000 points from the log-normal polytope of a given point

$p = \left(\frac{19}{105}, \frac{29}{105}, \frac{2}{5}, \frac{4}{7}\right) \in \mathcal{M}_1$

and colored them blue or red depending on if their MLE was $p$ or if their MLE was located on $\mathcal{M}_2$. Therefore, the blue convex set is the logarithmic Voronoi cell of $p \in \mathcal{M}$. Computations for this example can be found at https://mathrepo.mis.mpg.de/logarithmicVoronoi.

Next we consider log-linear, or toric, models. These include many important families of statistical models, such as undirected graphical models [19], independence models [38], and others as mentioned in the Introduction. For an $m \times n$ integer matrix $A$ with $\mathbf{1} \in \text{rowspan}(A)$, the corresponding log-linear model $\mathcal{M}_A$ is defined to be the set of all points $p \in \Delta_{n-1}$ such that $\log(p) \in \text{rowspan}(A)$ [38, p. 122].
**Theorem 11.** Let $A \in \mathbb{Z}^{m \times n}$ be an integer matrix such that $1 \in \text{rowspan}(A)$. Let $\mathcal{M}$ be the associated log-linear (toric) model. Then for any point $p \in \mathcal{M}$, the log-Voronoi cell of $p$ is equal to the log-normal polytope at $p$.

**Proof.** We will show that $\log \text{Vor}_\mathcal{M}(p) = \log N_p \mathcal{M} \cap \Delta_{n-1}$. The forward direction follows from Lemma 3, since these models are smooth off the coordinate hyperplanes (see [38, p. 150] and [3]). For the reverse direction, let $u \in \log N_p \mathcal{M}$. Although the log-likelihood function can have many complex critical points, it is strictly concave on log-linear models $\mathcal{M}$ for positive $u$, in particular for $u \in \Delta_{n-1}$. This means that there is exactly one critical point in the positive orthant, and it is the unique solution $p \in \mathcal{M}$ to the linear system $Ap = Au$ [14, Proposition 2.1.5]. This is known as Birch’s theorem. It follows that $\Phi(u) = p$, as desired. \hfill \square

As a corollary, the polytopes shown in Figures 1 and 3 are logarithmic Voronoi cells. Following [31], define the map sending a point in projective space to a convex combination of the columns $a_i$ of $A$, so that the image is a polytope, namely

$$\phi_A : \mathbb{P}^{n-1}_\mathbb{C} \to \mathbb{R}^m, \quad z \mapsto \frac{1}{\sum_{i=1}^{n} |z_i|} \sum_{i=1}^{n} |z_i| a_i.$$ 

This restricts to what [31, p. 120] calls the *algebraic moment map* $\phi_A|_{\mathcal{M}_A} = \mu_A : \mathcal{M}_A \to \mathbb{R}^m$, where $\mathcal{M}_A$ is the projective toric variety associated to $A$. The maximum likelihood estimator, then, is the map $\mu_A^{-1} \circ \phi_A$ restricted to $\Delta_{n-1}$, identified as a subset of $\mathbb{P}^{n-1}_\mathbb{C}$ by extending scalars and using the quotient map defining projective space. The fact [31, Corollary 8.24] that there is a unique preimage, allowing the definition of $\mu_A^{-1}$, played a crucial role in Theorem 11. Thus we have the following

**Corollary 12.** For toric models, the logarithmic Voronoi cells are the preimages $\phi_A^{-1}(\mu_A(p))$ intersected with $\Delta_{n-1}$. Thus, $\phi_A|_{\Delta_{n-1}}$ is a map whose image is a polytope and whose fibers are also polytopes.

For the Segre embedding of Example 6, the image is a square and the fibers are line segments, depicted in Figure 5, which adjoins our Figure 3 with [31, Figure 2, p. 121]. For more on the algebraic moment map, see [37].
Open questions. We point out several open questions that we find interesting:

(1) When $\mathcal{M}$ does not equal its log-normal polytope, an interesting open question is how to describe the boundary of the logarithmic Voronoi cells. The algebraic boundaries of Euclidean Voronoi cells were computed in [10].

(2) Initial attempts to investigate the boundary raised questions of transcendentality. An attempt to mimic the computations in [10] leads to an ideal generated by logarithmic functions, rather than polynomials, when applied to logarithmic rather than Euclidean Voronoi cells. This raises an important question of whether the boundaries of logarithmic Voronoi cells are algebraic or transcendental. If algebraic, how could one compute them? If transcendental, can this be proved, as was done in [2] or [21]?

(3) In addition, while we only considered smooth models in this paper, one could describe logarithmic Voronoi cells in the nonsmooth case. In particular, what can be said about the Voronoi cells of points in the singular locus? This is relevant for the important families of mixture models and secant varieties, as in Example 15 of Section 4.

(4) For matrices and tensors of fixed nonnegative rank, the geometry is more complicated. It would be interesting to study logarithmic Voronoi cells in this context, possibly in relation to the basins of attraction of the EM algorithm [28].

(5) Finally, we have focused on the discrete case, but continuous distributions could also be investigated. One promising case is linear Gaussian covariance models [5], since their maximum likelihood estimation is an algebraic optimization problem over a spectrahedral cone.

4. Computing Voronoi with numerical algebraic geometry

In case the logarithmic Voronoi cell is not a polytope, techniques from numerical nonlinear algebra can still be used to compute it effectively. In this section we demonstrate these methods. In particular, we take some care to explain how to set up the randomization that must be used in case the algebraic variety is defined by more polynomials than its codimension. The relevant equation (2) is explained below, and then used in Theorem 13. Numerical algebraic geometry [7; 36] can be used to efficiently find all isolated solutions of a square system of polynomial equations (square means equal number of equations and variables). The system of equations used in Theorem 13 formulates our problem specifically to take advantage of these tools.

Let $f$ be the $1 \times m$ row vector whose entries are the polynomials $f_1, \ldots, f_m$ in the variables $x_1, \ldots, x_n$. We assume that the first polynomial defines the simplex, i.e., $f_1 = \sum_{i=1}^{n} x_i - 1$. Let the algebraic set defined by $f_1, \ldots, f_m$ have codimension $c$. Let $df$ denote the $m \times n$ Jacobian matrix whose rows are the gradients of $f_1, \ldots, f_m$. Let $A$ be a $c \times m$ matrix whose entries are chosen randomly from independent normal distributions. Let $B$ be a similarly chosen random $(m-c) \times (n+c)$ matrix. Let $[\lambda \ -1]$ be the row vector of length $c + 1$ whose first $c$ entries are variables $\lambda_1, \ldots, \lambda_c$ and whose last entry is $-1$ and let $I_{n+c}$ be the identity matrix of size $n + c$. We are interested in the following vector equation whose
components give \( n + c \) polynomial equations in \( n + c \) unknowns:

\[
\begin{bmatrix}
\lambda - 1 \\
\nabla \ell_u \\
\end{bmatrix}
\begin{bmatrix}
A \cdot df \\
f \\
B
\end{bmatrix}
\begin{bmatrix}
I_{n+c} \\
\end{bmatrix}
= \begin{bmatrix} 0 & \cdots & 0 \end{bmatrix}.
\]  

(2)

**Theorem 13.** Let \( \mathcal{M} \) be the intersection of \( \Delta_{n-1} \) and an irreducible algebraic model given by the polynomial map \( f : \mathbb{R}^n \to \mathbb{R}^m \). Let \( u \in \Delta_{n-1} \) be fixed and generic. With probability 1, all points \( p \in \mathcal{M} \) such that \( u \in \log N_{p,\mathcal{M}} \) are among the finitely many isolated solutions to the square system of equations given in (2).

**Proof.** As a consequence of [26, Theorem 1.6], if \( u \in \Delta_{n-1} \) is generic, then with probability 1 there will be finitely many critical points of \( \ell_u \) restricted to \( \mathcal{M} \). If the algebraic set defined by \( f \) has codimension \( c \) then the dimension of the row space of \( df \) will be equal to \( c \) and the rows will span \( N_x \mathcal{M} \) for any generic \( x \in \mathcal{M} \) [34, p. 93]. With probability 1, multiplying by the random matrix \( A \) will result in a \( c \times n \) matrix of full row rank, whose rows also span \( N_x \mathcal{M} \). Appending the row \( \nabla \ell_u \) and multiplying the resulting matrix by the row vector \( [\lambda - 1] \) produces \( n \) polynomials which evaluate to zero whenever \( \nabla \ell_u \) is in the normal space \( N_x \mathcal{M} \). Appending the polynomials \( f_1, \ldots, f_m \) gives a \( 1 \times (n+m) \) row vector of polynomials evaluating to zero whenever \( x \in \mathcal{M} \) and \( \nabla \ell_u \) lies in the normal space \( N_x \mathcal{M} \). However, this system of equations is overdetermined. Applying Bertini’s theorem [7, Theorem 9.3] or [36, Theorem A.8.7] we can take random linear combinations of these polynomials using \( I_{n+c} \) and \( B \), and with probability 1; the isolated solutions of the resulting square system of polynomials will contain all isolated solutions of the original system of equations. The result follows. \( \square \)

**Remark 14.** If we are interested in computing the logarithmic Voronoi cell of a specific point \( p \in \mathcal{M} \), then we can generate a generic point \( u_0 \in \log N_{p,\mathcal{M}} \) by taking a random linear combination of the gradients of \( f_1, \ldots, f_m \). Using this point \( u_0 \) we can formulate our system of equations (2), one of whose solutions we already know, namely \( p \). Using monodromy, we can quickly find many other solutions \( p' \) by perturbing our parametrized system of equations through a loop in parameter space. For more details, see [1]. This is especially useful in the case where the ML degree is known a priori, since we can stop our monodromy search after finding ML-degree-many solutions. This process yields an optimal start system for homotopy continuation, allowing us to almost immediately compute solutions for other data points since we need only follow the ML-degree-many solution paths via homotopy continuation.

In the next example, we utilize the formulation in Theorem 13 to numerically compute a logarithmic Voronoi cell in a larger example of statistical interest, a mixture of two binomial distributions, also known as a secant variety.

**Example 15.** Bob has three biased coins, one in each pocket, and one in his hand. He flips the coin in his hand, and depending on the outcome, chooses either the coin in his left or right pocket, which he then flips five times, recording the total number of heads in the last five flips. To estimate the biases of Bob’s coins, Alice treats this situation as a 3-dimensional statistical model \( \mathcal{M} \subset \Delta_5 \subset \mathbb{R}^6 \). Using implicitization
We easily find each \( p \in \mathcal{M} \):

\[
f(x) = \begin{bmatrix}
20x_1x_3x_5 - 10x_1x_2^2 - 8x_2^2x_5 + 4x_2x_3x_4 - x_3^3 \\
100x_1x_2x_6 - 20x_1x_4x_5 - 40x_1^2x_6 + 4x_2x_3x_4 + 2x_4^2x_7 - x_5^2x_4 \\
100x_1x_4x_6 - 40x_1^2x_6 - 20x_2x_3x_6 + 4x_2x_4x_5 + 2x_3^2x_7 - x_5^2x_4 \\
20x_2x_4x_6 - 8x_2^2x_7 - 10x_2^2x_5 + 4x_3x_4x_5 - x_4^3 \\
x_1 + x_2 + x_3 + x_4 + x_5 + x_6 - 1
\end{bmatrix}.
\]

For a concrete example, consider the point which arises by setting the biases of the coins to \( b_1 = \frac{7}{11} \), \( b_2 = \frac{3}{7} \), \( b_3 = \frac{2}{7} \). Explicitly this point \( p \in \mathcal{M} \) is

\[
p = \left( \frac{518}{9375}, \frac{124}{625}, \frac{192}{625}, \frac{168}{625}, \frac{86}{625}, \frac{307}{9375} \right).
\]

The log-normal space \( \log N_p \mathcal{M} \) is 3-dimensional, becoming a 2-dimensional polytope when intersected with \( \Delta_5 \subset \mathbb{R}^6 \). This intersection is the log-normal polytope, in this case, a hexagon. In fact, this hexagon is the (2-dimensional) convex hull of the six vertices

\[
(0, 0, 0, 0, 0), (0, \frac{124}{375}, \frac{375}{86}, \frac{77}{375}, \frac{1376}{9375}), (\frac{259}{1875}, 0, \frac{518}{76875}, 0, 0, 0), (\frac{8288}{76875}, 0, \frac{3176}{1125}, 0, \frac{1376}{5125}, \frac{307}{76875}),
\]

By choosing an orthonormal basis agreeing with \( \log N_p \mathcal{M} \) we can plot this hexagon, though it lives in \( \mathbb{R}^6 \). Figure 2 shows the log-normal polytope and our numerical approximation of the logarithmic Voronoi cell (which is not a polytope) surrounding the point \( p \). By rejection sampling, we computed 60,000 points \( u_1, u_2, \ldots, u_{60000} \in \log N_p \mathcal{M} \cap \Delta_{n-1} \) in the log-normal polytope. By a result in [23], we know that the ML degree of this model is 39. Using the formulation presented in Theorem 13, we successfully computed all 39 critical points for each \( \ell_{u_i} \), \( i \in \{1, 2, \ldots, 60000\} \), restricted to \( \mathcal{M} \).

By rotating the log-normal polytope and choosing the maximum. If \( p = \Phi(u_i) \) then \( u_i \in \log \text{Vor}_\mathcal{M}(p) \) and we color that point green in Figure 2, while if \( p \neq \Phi(u_i) \) we color the point pink. The repeated computations of each set of 39 critical points were accomplished using the software HomotopyContinuation.jl [8], which can efficiently compute the isolated solutions to systems of polynomial equations using homotopy continuation [7; 36]. A full description of the Julia code needed to compute this example can be found online at [1].

5. All empirical distributions of fixed sample size

Consider running experiments with sample size \( d \) and choosing the model defined by

\[
\mathcal{M} := \frac{\mathbb{Z}^n \cap d \cdot \Delta_{n-1}}{d}.
\]

Philosophically, \( \mathcal{M} \) is the chaotic universe model. Adopting this model is to abandon the idea that experiments tell us about some simpler underlying truth, since the experimental data will always lie exactly on the model. In this section we investigate the Euclidean and logarithmic Voronoi cells for \( p \in \mathcal{M}_{n,d} \). Our motivation to study this model is historical, since Georgy Voronoi was interested in lattices and the partitions of Euclidean space they induce from the closest-point map. These became Voronoi...
cells. It led us to study the logarithmic Voronoi cells coming from maximum likelihood estimation for a lattice intersected with the probability simplex. In doing so, we found interesting connections with root polytopes of type $A$ and were able to generalize [9, Theorem 1] to our context, finding a complete combinatorial description of the face structure of the logarithmic Voronoi cells for $\mathcal{M}_{n,d}$. We give more historical context in the end of the section.

For convenience we work with the scaled set $d \cdot \Delta_{n-1}$ since all polytopes considered will be combinatorially equivalent to those we could define in $\Delta_{n-1}$. Then we define $\mathcal{M}_{n,d}$ as the set of all nonnegative integer vectors summing to $d$. Thus $(p_1, p_2, \ldots, p_n) = p \in \mathcal{M}_{n,d}$ has all coordinates $p_i \in \mathbb{N}$. These vectors can be used to create a projective toric variety, the $d$-th Veronese embedding of $\mathbb{P}^{n-1}$ into $\mathbb{P}^{N-1}$ [31, Chapter 8], but instead we treat them as the model itself. By Proposition 2, the logarithmic Voronoi cells for $p \in \mathcal{M}_{n,d}$ are polytopes. For any $p \in \mathcal{M}_{n,d}$ such that all coordinates $p_i$ are greater than 1, we will provide a full characterization of the faces of the corresponding logarithmic root polytopes in Theorem 17. Theorem 19 shows that these logarithmic root polytopes are dual to the logarithmic Voronoi cells. These are the main results of the section. Again using orthogonal projection from $\mathbb{R}^4$, Figure 6 shows all the logarithmic Voronoi cells for interior points of $\mathcal{M}_{4,9}$ and $\mathcal{M}_{4,10}$.

The Euclidean Voronoi cells for $p \in \mathcal{M}_{n,d}$ are the duals of root polytopes of type $A_{n-1}$; i.e., the facets are defined by inequalities whose normal vectors are $\{e_i - e_j : i \neq j\}$. Root polytope often refers to the convex hull of the origin and the positive roots $\{e_i - e_j : i < j\}$. These were studied in [20] in terms of their relationship to certain hypergeometric functions. However, we define root polytopes to be the convex hull of all roots, as studied in [9]. We also note that these polytopes are Young orbit polytopes for the partition $(n-1, 1)$ and find application in combinatorial optimization [32].

Denote the $(n-1)$-dimensional root polytope by $P_n \subset \mathbb{R}^n$, so that the Euclidean Voronoi cells of $p \in \mathcal{M}_{n,d}$ are the dual $P_n^*$. The volume of $P_n$ is equal to $(n/(n-1)!)C_{n-1}$, where $C_{n-1}$ is a Catalan number. Every nontrivial face of $P_n$ is a Cartesian product of two simplices and corresponds to a pair of nonempty, disjoint subsets $I, J \subset [n]$. Every $m$-dimensional face of $P_n$ is the convex hull of the vectors $\{e_i - e_j : i \in I, j \in J\}$, with $|I| + |J| = m + 2$, so there is a bijection between nontrivial faces and the set of ordered partitions of subsets of $[n]$ with two blocks [9, Theorem 1]. This result is related to the face

\begin{figure}[h]
\centering
\includegraphics[width=0.8\textwidth]{voronoi_cells.png}
\caption{Logarithmic Voronoi cells (rhombic dodecahedra) of interior points for $n = 4$, $d = 9$ (on the left) and $d = 10$ (on the right).}
\end{figure}
description of $\Pi_{n-1}$, the permutahedron, since $P_n$ is a \textit{generalized permutahedron} and can be obtained by collapsing certain faces of $\Pi_{n-1}$.

In the logarithmic setting, analogous polytopes $\log P_n(p)$ exist, playing the same role as the root polytopes in the Euclidean case. However, their details are more complicated. The correct modifications motivate the following definition.

**Definition 16.** The \textit{logarithmic root polytope} for $p \in \mathcal{M}_{n,d}$ is defined as the convex hull of the $2\binom{n}{2}$ vertices $v_{ij}$ for $i \neq j \in [n]$ given by the formulas

$$v_{ij} := \frac{1}{b_j p_j - a_i p_i} \left[ a_i e_i - b_j e_j - \frac{(a_i - b_j)}{n} \mathbf{1} \right],$$

where

$$a_i := \log \left( \frac{p_i + 1}{p_i} \right) \quad \text{and} \quad b_j := \log \left( \frac{p_j}{p_j - 1} \right)$$

and where $\mathbf{1} := \sum_{k \in [n]} e_k$. Note that $a_i, b_j > 0$ are always positive real numbers and all vectors $v_{ij}$ are orthogonal to $\mathbf{1}$. We denote the polytope by $\log P_n(p)$.

The statement and proof of the following Theorem 17 was inspired by and closely follows [9, Theorem 1]. However, significant details needed to be modified. For example, the linear functional

$$g = (1, 0, -1, 1, -1, 0, -1)$$

is replaced by

$$\begin{bmatrix}
-a_1 a_4 b_3 b_5 p_1 - a_1 a_4 b_3 b_7 p_1 - a_1 a_4 b_5 b_7 p_1 + a_1 b_3 b_5 b_7 p_1 + a_1 b_3 b_3 b_7 p_3 + a_4 b_3 b_7 p_3 + a_4 b_3 b_3 b_7 p_5 + a_4 b_3 b_5 b_7 p_7 \\
0 \\
a_1 a_1 b_3 b_7 p_1 - a_1 a_4 b_3 b_5 p_3 - a_1 a_4 b_3 b_7 p_3 - a_1 b_3 b_5 b_7 p_3 + a_1 a_4 b_5 b_7 p_4 + a_1 a_4 b_3 b_5 p_4 + a_1 b_3 b_5 b_7 p_4 + a_1 a_4 b_3 b_7 p_7 \\
a_1 a_3 b_3 b_7 p_1 + a_1 b_3 b_3 b_7 p_3 - a_1 a_4 b_3 b_5 p_4 - a_1 a_4 b_5 b_7 p_4 + a_1 a_4 b_3 b_5 p_4 + a_1 b_3 b_5 b_7 p_4 + a_1 a_4 b_3 b_7 p_7 \\
a_1 a_3 b_3 b_7 p_1 + a_1 a_4 b_3 b_5 p_3 + a_1 a_4 b_3 b_7 p_3 - a_1 a_4 b_5 b_7 p_4 - a_1 a_4 b_3 b_5 p_4 + a_1 b_3 b_5 b_7 p_4 + a_1 a_4 b_3 b_7 p_7 \\
0 \\
0 \\
a_1 a_4 b_3 b_5 p_1 + a_1 a_4 b_3 b_5 p_3 + a_1 a_4 b_3 b_5 p_4 + a_1 a_4 b_3 b_5 p_5 - a_1 a_4 b_3 b_7 p_7 - a_1 a_4 b_3 b_7 p_7 - a_1 b_3 b_5 b_7 p_7 - a_1 b_3 b_5 b_7 p_7 - a_1 b_3 b_5 b_7 p_7
\end{bmatrix}.$$

This linear functional plays the same role for the logarithmic root polytope of $(p_1, p_2, \ldots, p_7) \in \mathcal{M}_{7,d}$ as $g$ plays for the usual root polytope in the proof of [9, Theorem 1].

**Theorem 17.** For $m \in \{0, 1, \ldots, n - 2\}$, every $m$-dimensional face of the logarithmic root polytope for $p \in \mathcal{M}_{n,d}$ is given by the convex hull of the vertices $v_{ij}$ for $i \in I, j \in J$, where $I, J$ are disjoint nonempty subsets of $[n]$ such that $|I| + |J| = m + 2$. Thus there is a bijection between nontrivial faces and the set of ordered partitions of subsets of $[n]$ with two blocks, where the dimension of the face corresponding to $(I, J)$ is $|I| + |J| - 2$.

**Proof.** Each face of a polytope can be described as the subset of the polytope maximizing a linear functional. Recall that we have fixed some $p \in \mathcal{M}_{n,d}$ with all $p_k > 1$ and that

$$a_i := \log \left( \frac{p_i + 1}{p_i} \right) \quad \text{and} \quad b_j := \log \left( \frac{p_j}{p_j - 1} \right).$$
In our formula (3) we use a shorthand for writing square-free monomials in the $a_1, a_2, \ldots, a_n$ and the $b_1, b_2, \ldots, b_n$. For example if $I = \{1, 2, 4\}$ then $a^I = a_1a_2a_4$, while if $J = \{3, 5\}$ then $b^J = b_3b_5$. For a pair of disjoint nonempty subsets $I, J$ of $[n]$ we define the linear functional $g_{IJ} = (g_1, g_2, \ldots, g_n) \in (\mathbb{R}^n)^*$ by the formulas

\[
\begin{align*}
\text{if } \ell \in I, & \quad g_\ell = \sum_{i \in I \setminus \ell} a^I \setminus i b^I (a_i p_i - a_\ell p_\ell) + \sum_{j \in J} a^I \setminus J b^J (b_j p_j - a_\ell p_\ell), \\
\text{if } \ell \in J, & \quad g_\ell = \sum_{i \in I} a^I b^J (a_i p_i - a_\ell p_\ell) + \sum_{j \in J \setminus \ell} a^I \setminus J b^J (b_j p_j - b_\ell p_\ell), \\
\text{else,} & \quad g_\ell = 0.
\end{align*}
\]

Then the convex hull of the vectors $\{v_{ij} : i \in I, j \in J\}$ is the face maximizing $g_{IJ}$. To see this, first note that $g_{IJ} \cdot 1 = 0$. Because of this fact we can ignore the component of $v_{ij}$ in the 1 direction. Recall that

\[
v_{ij} := \frac{1}{b_j p_j - a_i p_i} \left[ a_i e_i - b_j e_j - \frac{(a_i - b_j)}{n} 1 \right],
\]

so that to evaluate $g_{IJ}$ on $v_{ij}$ it is enough to evaluate on

\[
\frac{1}{b_j p_j - a_i p_i} [a_i e_i - b_j e_j].
\]

Recalling that the $a_i$ and $b_j$ are always positive and that $p_k > 1$, it can be seen that $g_{IJ}$ takes equal value on every vertex $v_{rs}$ for $r \in I, s \in J$, and strictly less on every other vertex. We omit the details of the admittedly lengthy calculation, but note that the common maximum value attained on all vertices $v_{rs}$ for $r \in I, s \in J$ is equal to

\[
\sum_{i \in I} a^I b^I + \sum_{j \in J} a^I b^J.
\]

Conversely, given an arbitrary linear functional $f = (f_1, f_2, \ldots, f_n)$ determining a nontrivial face $F$, collect the indices where its components are nonnegative in a set $I$ and the indices where its components are negative in a set $J$. Then $(I, J)$ is a partition of $[n]$ and we refer to the same formulas (3) as above in order to define the sets $(I', J')$ as follows. If $I \neq \emptyset$ and $J \neq \emptyset$ then let

\[
I' := \{ i : f_i / g_i = \max(f_\ell / g_\ell : \ell \in I) \}, \\
J' := \{ j : f_j / g_j = \max(f_\ell / g_\ell : \ell \in J) \}.
\]

If $I = \emptyset$ then let

\[
I' := \{ i : f_i / g_i = \min(f_\ell / g_\ell : \ell \in J) \}, \\
J' := \{ j : f_j / g_j = \max(f_\ell / g_\ell : \ell \in J) \},
\]

while if $J = \emptyset$ then let

\[
I' := \{ i : f_i / g_i = \max(f_\ell / g_\ell : \ell \in I) \}, \\
J' := \{ j : f_j / g_j = \min(f_\ell / g_\ell : \ell \in I) \}.
\]

Note that the face $F$ is the convex hull of the vectors $\{v_{ij} : i \in I', j \in J'\}$ and hence $(I', J')$ are determined independently of the choice of linear functional which maximizes the given face.
Now we show that the dimension of the face corresponding to disjoint nonempty sets $I, J$ of $[n]$ is $|I| + |J| - 2$. Let $I = \{i_1, \ldots, i_{|I|}\}$ and $J = \{j_1, \ldots, j_{|J|}\}$. Then

$$X = \{v_{i_\ell, j_\ell} : \ell = 1, \ldots, |J|\} \cup \{v_{i_\ell, j_\ell} : \ell = 2, \ldots, |I|\}$$

is a maximal linearly independent subset of $|I| + |J| - 1$ of the vectors $v_{ij}$, $i \in I$, $j \in J$. In addition, for any $i \in I, j \in J$ either $v_{ij} \in X$ or we can write it as an affine combination (coefficients sum to 1) of vectors in $X$, namely

$$v_{i, j} = \left(\frac{b_{ji}p_{ji} - a_{ij}p_i}{b_jp_j - a_{ij}p_i}\right)v_{i, j} - \left(\frac{b_{ji}p_{ji} - a_{ij}p_i}{b_jp_j - a_{ij}p_i}\right)v_{i_1, j} + \left(\frac{b_{j}p_{j} - a_{ij}p_i}{b_jp_j - a_{ij}p_i}\right)v_{i, j}.$$

Hence, $X$ is an affine basis of the face corresponding to $I, J$, whose dimension is $|X| - 1$, which is $|I| + |J| - 2$ as desired. 

**Example 18.** Let $n = 6$, $I = \{1, 4\}$, $J = \{2, 3, 5\}$ and $p = (2, 15, 3, 5, 9, 6)$. We implemented the formulas (3) in floating point arithmetic (due to the logarithms) and obtained (shown to only three digits)

$$g_{IJ} = (0.00415, -0.00200, -0.00398, 0.00474, -0.00291, -0.000).$$

We can evaluate this linear functional on the vertices $v_{ij}$ for $i \neq j$, where $i, j \in [6]$, and obtain the following values, which attain their maximum on $v_{12}, v_{13}, v_{15}, v_{42}, v_{43}, v_{45},$ as expected:

- $0.008135843945 v(1, 2) = (1.56, -0.559, -0.251, -0.251, -0.251, -0.251),$
- $0.008135843948 v(1, 3) = (1.00, 0.000, -1.00, 0.000, 0.000, 0.000),$
- $0.00205214856 v(1, 4) = (1.23, -0.0997, -0.0997, -0.832, -0.0997, -0.0997),$
- $0.008135843948 v(1, 5) = (1.43, -0.192, -0.192, -0.192, -0.665, -0.192),$
- $0.00595301519 v(1, 6) = (1.30, -0.132, -0.132, -0.132, -0.132, -0.776),$
- $0.00719236292 v(2, 1) = (-1.41, 0.405, 0.250, 0.250, 0.250, 0.250),$
- $0.00598264733 v(2, 3) = (0.229, 0.488, -1.40, 0.229, 0.229, 0.229),$
- $0.00804489030 v(2, 4) = (0.179, 0.615, 0.179, -1.33, 0.179, 0.179),$
- $0.002322216671 v(2, 5) = (0.0963, 0.796, 0.0963, 0.0963, -1.18, 0.0963),$
- $0.001027169161 v(2, 6) = (0.156, 0.669, 0.156, 0.156, 0.156, -1.29),$
- $0.007691322875 v(3, 1) = (-1.20, 0.129, 0.679, 0.129, 0.129, 0.129),$
- $0.005863205380 v(3, 2) = (-0.213, -0.615, 1.47, -0.213, -0.213, -0.213),$
- $0.008723741580 v(3, 4) = (-0.0426, -0.0426, 1.10, -0.926, -0.0426, -0.0426),$
- $0.004075725208 v(3, 5) = (-0.144, -0.144, 1.32, -0.144, -0.742, -0.144),$
- $0.004962538041 v(3, 6) = (-0.0762, -0.0762, 1.17, -0.0762, -0.0762, -0.867),$
- $0.004242519680 v(4, 1) = (-1.28, 0.179, 0.179, 0.563, 0.179, 0.179),$
- $0.00813584394 v(4, 2) = (-0.153, -0.713, -0.153, 1.33, -0.153, -0.153),$
- $0.008135843947 v(4, 3) = (0.122, 0.122, -1.21, 0.720, 0.122, 0.122),$
- $0.008135843947 v(4, 5) = (-0.0723, -0.0723, -0.0723, 1.15, -0.865, -0.0723),$
We say that this system of inequalities is sufficient with all the half-spaces $H$.

Let $p$.

Proof. The vectors $v$ of inequalities which is of the form $\sum \text{polytope on the hyperplane}$ of these hyperplanes along the all-ones vector $q$ of zero since $p$ for all $n$.

From every point $q$ reachable from $p$ for all points $q \in M_n,d$ with $q \neq p$, where

$$H_q(u) := \sum_{i \in [n]} u_i \log \left( \frac{p_i}{q_i} \right).$$

We say that this system of inequalities is sufficient to define the logarithmic Voronoi cell. However, not all of these inequalities are necessary. Lemma 20 shows that a certain set of $2^n$ inequalities is sufficient for all $n \in \mathbb{Z}_{\geq 2}$. These are the inequalities $H_q(u) \geq 0$ for $q = p + e_i - e_j$ for $i \neq j$. We avoid logarithms of zero since $p_k > 1$ and we are away from the simplex boundary. In other words, we get one inequality from every point $q$ reachable from $p$ by moving along a root of type $A_{n-1}$.

These $H_q(u) \geq 0$ inequalities are linear, with constant term zero. However, projecting the normal vectors of these hyperplanes along the all-ones vector $1$ and viewing $p$ as the origin of a new coordinate system, we obtain inequalities with nonzero constant terms. These inequalities describe the same logarithmic Voronoi polytope on the hyperplane $\sum_k u_k = d$. Dividing each inequality by the constant terms we obtain a system of inequalities which is of the form $Au \leq 1$, following the notation of [42], where the rows of $A$ are exactly the vectors $v_{ij}$. By [42, Theorem 2.11], the dual polytope is given by the convex hull of these $v_{ij}$. □

Lemma 20. Let $p \in M_n,d$ with every entry $p_i$ greater than $1$. A sufficient system of inequalities defining the logarithmic Voronoi cell is given by the $2^n$ half-spaces $u \in \mathbb{R}^n$ such that $H_\delta(u) \geq 0$ for $\delta \in R := \{e_i - e_j : i \neq j, i, j \in [n]\}$ and the affine plane $\sum u_i = d$, where

$$H_{\delta}(u) := \sum_{i \in [n]} u_i \log \left( \frac{p_i}{p_i + \delta_i} \right).$$

Proof. We prove that the $2^n$ inequalities $H_\delta(u) \geq 0$ for $\delta \in R$ are sufficient. Fix $p \in M$ with all $p_i > 1$. Let $u \in \mathbb{R}^n$ such that $H_\delta(u) \geq 0$ for all $\delta \in R$. Fix some $q = p + \delta + \delta'$, where $\delta, \delta' \in R$, and assume that...
We claim that $\delta + \delta' \notin R$. We wish to show

$$H_q(u) = \sum_i u_i \log \frac{p_i}{q_i} \geq 0.$$ 

Consider several cases. First, if $\delta = \delta' = e_j - e_k$, it suffices to show that

$$u_j \log \frac{p_j}{p_j + 2} + u_k \log \frac{p_k}{p_k - 2} \geq 0.$$ 

We claim that

$$u_j \log \frac{p_j}{p_j + 2} + u_k \log \frac{p_k}{p_k - 2} \geq 2u_j \log \frac{p_j}{p_j + 1} + 2u_k \log \frac{p_k}{p_k - 1},$$

which would be sufficient, since the right-hand side of the above equation is $\geq 0$ by assumption. We show that

$$u_j \log \frac{p_j}{p_j + 2} \geq 2u_j \log \frac{p_j}{p_j + 1} \quad \text{and} \quad u_k \log \frac{p_k}{p_k - 2} \geq 2u_k \log \frac{p_k}{p_k - 1}. \quad (4)$$

Observe

$$u_j \log \frac{p_j}{p_j + 2} \geq 2u_j \log \frac{p_j}{p_j + 1} \iff p_j^2 + 2p_j + 1 \geq p_j^2 + 2p_j,$$

$$u_k \log \frac{p_k}{p_k - 2} \geq 2u_k \log \frac{p_k}{p_k - 1} \iff p_k^2 - 2p_k + 1 \geq p_k^2 - 2p_k.$$

Thus (5) holds, and we conclude that (4) is true in this case, as desired. If $\delta \neq \delta'$ but they share both indices, then $p = q$, and we’re done. If they do not share any indices, then we have that $H_q(u) = H_\delta(u) + H_{\delta'}(u) \geq 0$ by assumption. Suppose $\delta \neq \delta'$, and $\delta$ and $\delta'$ share one index, $j$. If $\delta = e_i - e_j$ and $\delta' = e_j - e_k$ for $i \neq j \neq k$, then $\delta + \delta' = e_i - e_k$, a contradiction to the assumption $\delta + \delta' \notin R$. It is similar when $\delta = e_j - e_i$ and $\delta' = e_k - e_j$. Suppose then that $\delta = e_i - e_j$ and $\delta' = e_j - e_k$. We wish to show that

$$u_i \log \frac{p_i}{p_i + 2} + u_j \log \frac{p_j}{p_j - 1} + u_k \log \frac{p_k}{p_k - 1} \geq 0.$$

Note then that

$$u_i \log \frac{p_i}{p_i + 2} \geq 2u_i \log \frac{p_i}{p_i + 1} \iff p_i^2 + 2p_i + 1 \geq p_i^2 + 2p_i,$$

and the last inequality always holds for positive $p_i$, so the lemma is true for this case. The case when $\delta = e_j - e_i$ and $\delta' = e_k - e_i$ is proved similarly. Since $H_q(u) \geq 0$ in all of the cases we considered, and the cases are exhaustive, we conclude that the lemma holds.\hfill$\square$

**A family of polytopes.** Using the face characterization of Theorem 17, we may compute the $f$-vectors of the logarithmic Voronoi cells for any $n$. We can also numerically calculate the $f$-vector for the logarithmic Voronoi cell of any specific point $p \in \mathcal{M}_{n,d}$ with $p_i > 1$ for all $i$ by explicitly constructing the polytope using inequalities. Of course, both of these calculations match. Below we list the $f$-vectors (which we computed in both ways) for small values $n \in \{2, 3, 4, 5, 6, 7\}$ to give the reader a sense of their behavior. The logarithmic Voronoi cells for every $\mathcal{M}_{n,d}$ are combinatorially isomorphic to the dual of the
corresponding root polytope, exactly as in the Euclidean case:

\[ n = 2(1, 2, 1), \]
\[ n = 3(1, 6, 6, 1), \]
\[ n = 4(1, 14, 24, 12, 1), \]
\[ n = 5(1, 30, 70, 60, 20, 1), \]
\[ n = 6(1, 62, 180, 210, 120, 30, 1), \]
\[ n = 7(1, 126, 434, 630, 490, 210, 42, 1). \]

Therefore we have a family of Euclidean Voronoi polytopes that tile \( \mathbb{R}^{n-1} \) and a family of logarithmic Voronoi polytopes that tile the open simplex \( \Delta_{n-1} \). This family begins

\[ n - 1 = 1, \quad n - 1 = 2, \quad n - 1 = 3, \quad \cdots \]
(line segment) (hexagon) (rhombic dodecahedron) \( \cdots \)

Root polytopes of type \( A \) have connections to tropical geometry. The rhombic dodecahedron is a polytope which has been called the 3-pyrope because of the mineral \( \text{Mg}_3\text{Al}_2(\text{SiO}_4)_3 \) whose pure crystal can take the same shape. For more on root polytopes, tropical geometry, and polytopes, see [27].

**Historical comments.** Georgy Voronoi devoted many years of his life to studying properties of 3-dimensional parallelohedra, convex polyhedra that tessellate 3-dimensional Euclidean space. His paper on the subject called *Recherches sur les parallélloèdres primitifs* [40] was a result of his twelve-year work. In a cover letter to the manuscript, he wrote, “I noticed already long ago that the task of dividing the \( n \)-dimensional analytical space into convex congruent polyhedra is closely related to the arithmetic theory of positive quadratic forms” [39]. Indeed, Voronoi was interested in studying cells of lattices in \( \mathbb{Z}^n \) with the aim of applying them to the theory of quadratic forms. This motivated us to study a lattice intersected with the probability simplex, the topic of our current section. Today, Voronoi decomposition finds applications to the analysis of spatially distributed data in many fields of science, including mathematics, physics, biology, archeology, and even cinematography. In [41], the author uses Voronoi cells to optimize search paths in an attempt to improve the final 6-minute scene of Andrei Tarkovsky’s *Offret (the Sacrifice)*. Voronoi diagrams are so versatile they even found their way into baking: Ukrainian pastry chef Dinara Kasko uses Voronoi diagrams to 3D-print silicone molds which she then uses to make cakes.\(^1\)

**Acknowledgements**

Both authors would like to thank Bernd Sturmfels for suggesting this topic during the summer of 2019 and for many helpful suggestions along the way. We also thank the Max Planck Institute of Mathematics in the Sciences for support during the summers of 2019 and 2020, and also the library staff for their exceptional support during the difficult pandemic, allowing us access to the resources we need for research. Alexandr was also supported by the Berkeley Chancellor’s fellowship.

\(^1\)http://www.dinarakasko.com
References


LOGARITHMIC VORONOI CELLS


Received 2020-06-20. Revised 2020-11-11. Accepted 2020-12-28.

YULIA ALEXANDR: yalexandr@berkeley.edu
University of California, Berkeley, CA, United States

ALEXANDER HEATON: alexheaton2@gmail.com
Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany

and

Technische Universität Berlin, Berlin, Germany

mathematical sciences publishers
THE SEMIALGEBRAIC GEOMETRY OF SATURATED OPTIMAL DESIGNS FOR THE BRADLEY–TERRY MODEL

THOMAS KAHLE, FRANK RÖTTGER AND RAINER SCHWABE

Optimal design theory for nonlinear regression studies local optimality on a given design space. We identify designs for the Bradley–Terry paired comparison model with small undirected graphs and prove that every saturated, locally $D$-optimal design is represented by a path. We discuss the case of four alternatives in detail and derive explicit polynomial inequality descriptions for optimality regions in parameter space. Using these regions, for each point in parameter space we can prescribe a locally $D$-optimal design.

1. Introduction

Consider an experimental situation in which $m$ alternatives are to be brought into a rank order. For each single observation in this experiment only two of these alternatives can be compared at a time and only a binary response can be observed which indicates the rank order of the two alternatives presented. Such experiments are known in economics as “discrete choice” experiments and in psychology as “forced choice” experiments or “ipsative measures”. The use of such experiments dates back to the work by Fechner [8] on psychophysics in a deterministic setup. In a statistical setup this situation is described by the Bradley–Terry model, which was introduced in [27] to rank chess players in tournaments and in [3] to analyze taste testing results for pork depending on different feeding patterns. See [17] for a leisurely introduction.

This model has proven popular in different areas of statistics, also outside of chess tournaments and pork tasting. In [14], Hastie and Tibshirani developed a coupling model similar to the Bradley–Terry model to study class probabilities for pairs of classes. In [24], Simons and Yao discussed the model asymptotics when the number of potential alternatives tends to infinity. Algorithms for Bradley–Terry models are discussed, for example, in [15], and asymptotics of algorithms, for example, in [7]. Besides marketing or transportation, another popular application area for the Bradley–Terry model is the world of professional sports, such as American football, car racing, matching in tournaments, card games or strategies for sport bets; see [4; 11; 1; 18]. The Bradley–Terry model is part of a broader class of models that describe statistical rankings. Specifically, it arises from marginalization of the Plackett–Luce model; see [25].

In this paper we are interested in optimal experimental designs for the Bradley–Terry model, that is, a scheme to assign a fixed number of measurements to different experimental settings such that the experiment is most informative about the parameters. Optimal experimental designs for the Bradley–Terry model were first investigated in [26], which gave an algorithmic approach to fit the model parameters.

MSC2020: primary 62K05, 62R01; secondary 13P25, 14P10, 62J02.

Keywords: nonlinear regression, optimal design, polynomial inequalities.
In [10] Graßhoff and Schwabe completely analyzed the case of three competing alternatives (with pair comparisons). They gave symbolic solutions for the design problem depending on the parameters and described the optimality areas of these design classes in the parameter space. The present paper extends the results of Graßhoff and Schwabe in two directions. We discuss the case of four competing alternatives in detail and characterize optimal saturated designs for an arbitrary number of competing alternatives, always with pair comparisons. The case of four alternatives arises also when considering a $2^2$ layout with interaction where two attributes can be set to two levels each. After a reparametrization, which does not affect the $D$-optimality, this model can be identified as a single-attribute model with four levels which can be used as alternatives in the Bradley–Terry model.

Section 2 gives the general setup. Section 5 contains an almost complete analysis of the case of four alternatives. Only one very challenging polynomial inequality system remains open (Problem 15). In Sections 3 and 4 we discuss saturated optimal designs for an arbitrary number of alternatives. Our main result is an easy combinatorial polynomial inequality description of regions in parameter space where a given saturated design is optimal, including the information for which designs these regions of optimality are empty (Theorem 11). Polynomial inequality constraints in experimental design are a recurrent topic. See [16] for a discussion of this principle for Poisson regression. Knowledge about the optimality regions can be very helpful in designing experiments. For example, a screening experiment could reveal that the estimates of the parameters are all within one region of optimality. In such a situation it is then clear which design to use. See [6] for a general class of models where local optimality is studied. In the discussion in Section 6 we compare the efficiency of our tailored designs versus uniform designs as the parameters grow in magnitude.

### 2. General setup

We consider pairs $(i, j)$ of alternatives $i, j = 1, \ldots, m$. The preference of $i$ over $j$ is modeled by a binary variable $Y(i, j)$ taking the value $Y(i, j) = 1$ if $i$ is preferred over $j$ and $Y(i, j) = 0$ otherwise. We do not consider any order effects here. The main assumption of the Bradley–Terry model is that there is a hidden ranking of the alternatives according to some numerical preference value $\pi_i > 0$, $i = 1, \ldots, m$. When presented with the pair $(i, j)$, the probability of preferring $i$ over $j$ is

$$
P(Y(i, j) = 1) = \frac{\pi_i}{\pi_i + \pi_j}.
$$

The model can be transformed into a logistic model using $\beta_i := \log(\pi_i)$. Then

$$
P(Y(i, j) = 1) = \frac{1}{1 + \exp(-(\beta_i - \beta_j))} = \eta(\beta_i - \beta_j),
$$

with $\eta(z) = (1 + \exp(-z))^{-1}$ as the inverse logit link function.

Scaling all $\pi_i$ with a constant factor leaves the preference probabilities invariant. Therefore one can without loss of generality assume that $\pi_m = 1$ or $\beta_m = 0$. This means that the number of parameters of the Bradley–Terry model is $m - 1$. The number of alternatives is the main measure of complexity of the design theory as it equals the dimension of the design space. The remaining parameters can be identified.
and \( \beta_m = 0 \) is known as control coding. We denote by \( e_i \) the \( i \)-th standard unit vector in \( \mathbb{R}^{m-1} \). To exhibit our model as a generalized linear model, the regression vector for a pair \((i, j)\) is

\[
f(i, j) = \begin{cases} 
  e_i - e_j & \text{for } i, j \neq m, \\
  e_i & \text{for } i < j, \ j = m, \\
  0 & \text{for } i = j = m.
\end{cases}
\]

With \( \beta^T = (\beta_1, \ldots, \beta_{m-1}) \) this yields \( \mathbb{P}(Y(i, j) = 1) = \eta(f(i, j)^T \beta) \), where \( f(i, j)^T \beta \) is the linear predictor.

**Remark 1.** When all probabilities

\[
p_{ij} := \mathbb{P}(Y(i, j) = 1) = \frac{\pi_i}{\pi_i + \pi_j} \quad \text{for } i, j \in [m] := \{1, 2, \ldots, m\}
\]

are treated as coordinates in \( \mathbb{R}^{m(m-1)} \), the Bradley–Terry model can be described by algebraic equations. This means that all values of the \( p_{ij} \) that arise for different values of \( \pi \) satisfy certain algebraic equations and, among the probability vectors, they are the only solutions to these equations. Theorem 7.7 of [25] shows that the model has the special geometric structure of a toric variety and its defining equations consist of binomials and linear trinomials.

The design region of the Bradley–Terry paired comparison model is

\[\mathcal{X} = \{(i, j) : i, j = 1, \ldots, m, i < j\} \]

It consists of all pairs of ordered alternatives. The pairs \((i, j)\) and \((j, i)\) bear the same information, and the comparison \((i, i)\) of two identical alternatives does not have any information at all (as can be seen easily later). Therefore, whenever there are two alternatives \( i, j \in \{1, \ldots, m\} \) we assume \( i < j \). An experimental design is an assignment of a weight \( w_{ij} \geq 0 \) to each point \((i, j)\) \( \in \mathcal{X} \), such that \( \sum_{ij} w_{ij} = 1 \) (compare, for example, [23]). Although a design could be impossible to realize with a finite number \( N \) of observations, it is common to let \( w_{ij} \in \mathbb{R} \) as opposed to \( w_{ij} \in \frac{1}{N} \mathbb{N} \). For any \( k \in \mathbb{N} \) we write

\[
\Delta_k := \left\{ w \in \mathbb{R}^k_{\geq 0} : \sum_{l} w_l = 1 \right\}
\]

for the \((k-1)\)-dimensional simplex in \( \mathbb{R}^k \) whose vertices are the standard unit vectors. It is customary to use \( \xi \) to refer to a design with weights \( w_{ij} \) and slightly abuse notation with expressions like \( \xi \in \Delta(\eta) \).

The information gained from one observation of \( Y(i, j) \) is encoded in the information matrix

\[
M((i, j), \beta) = \lambda_{ij} f(i, j) f(i, j)^T \in \mathbb{R}^{(m-1)\times(m-1)},
\]

where

\[
\lambda_{ij} := \lambda_{i,j}(\beta) = \eta'(\beta_i - \beta_j) = \frac{e^{\beta_i - \beta_j}}{(1 + e^{\beta_i - \beta_j})^2}
\]

is referred to as the intensity in [10]. It holds that \( M((i, j), \beta) = M((j, i), \beta) \) and \( M((i, i), \beta) = 0 \).
Assuming independent observations, the information matrix for a design \( \xi \) with weights \( w_{ij} \) is the \((m-1) \times (m-1)\)-matrix
\[
M(\xi, \beta) = \sum_{(i,j)} w_{ij} M((i, j), \beta) = \sum_{(i,j)} w_{ij} \lambda_{ij} f(i, j) f(i, j)^T.
\] (2-1)

The theory of optimal experimental design suggests picking weights \( w_{ij} \) that optimize a numerical function of \( M(\xi, \beta) \). Standard references that include the theory for generalized linear models are [20; 23]. One popular function to optimize is the logarithm of the determinant:

**Definition 2.** An experimental design \( \xi^* \) is locally \( D \)-optimal if
\[
\log \det(M(\xi^*, \beta)) \geq \log \det(M(\xi, \beta))
\]
for all \( \xi \in \Delta_{m^2} \).

In optimal experimental design one speaks of local optimality if the optimal choice of a design depends on the unknown parameters that one wants to learn about; see [5]. From the perspective of mathematical optimization one has a parametric family of convex optimization problems where both the optimization domain (the polytope of information matrices) and the target function depend on the parameters \( \beta \). The methods of convex optimization suggest studying the directional derivatives of the target function. The following is found in [23, Section 3.5.2].

**Definition 3.** The directional derivative (Fréchet derivative) of the \( D \)-optimality criterion at \( M_1 \) in the direction of \( M_2 \) for some \((m-1) \times (m-1)\) information matrices \( M_1, M_2 \) is
\[
F_D(M_1, M_2) = \lim_{\varepsilon \searrow 0} \frac{1}{\varepsilon} \left( \log \det((1 - \varepsilon)M_1 + \varepsilon M_2) - \log \det(M_1) \right).
\]

It is shown in [23, Sections 3.8 and 3.11] that
\[
F_D(M(\xi, \beta), M((i, j), \beta)) = \lambda_{ij} f(i, j)^T M(\xi, \beta)^{-1} f(i, j) - (m - 1). \tag{2-2}
\]
This yields the following \( D \)-optimality criterion:

**Theorem 4** (Kiefer–Wolfowitz). A design \( \xi^* \) is locally \( D \)-optimal if and only if
\[
\lambda_{ij} f(i, j)^T M(\xi^*, \beta)^{-1} f(i, j) \leq m - 1 \tag{2-3}
\]
for all \( 1 \leq i < j \leq m \).

The following corollary from [23, Corollary 3.10] is very useful.

**Corollary 5.** For design points \((i, j)\) with positive weight in \( \xi^* \), the inequalities (2-3) in Theorem 4 hold with equality.

A main observation about the Bradley–Terry model is that it is useful to represent pairs \((i, j)\) with positive weights \( w_{ij} \) as the edges of an undirected graph on the vertex set \( \{1, \ldots, m\} \). Properties of these graphs, in particular the edge density, determine the asymptotics of estimation for sparse Bradley–Terry models [13].

**Definition 6.** A graph representation of a design \( \xi \) for the Bradley–Terry model is the undirected simple graph with vertex set \( \{1, \ldots, m\} \) and edge set \( E = \{(i, j) : w_{ij} > 0\} \).
Using standard notions from graph theory, a tree is a connected graph with no cycles. A path is a tree in which every vertex is connected to at most two other vertices. We exploit the symmetry of the model. The symmetric group \( S_m \) of all bijective self-maps of \([1, \ldots, m]\) permutes the alternatives. The permutation action extends to ordered pairs by acting on both entries of the pair simultaneously (and changing the order if necessary). The action also extends naturally to designs \( \xi \) on pairs \((i, j)\) by putting \((\xi^\sigma)_{(i,j)} = \xi_{\sigma^{-1}(i,j)}\) for any \( \sigma \in S_m \). A graph representation of an entire orbit under this action is simply the unlabeled graph. Proposition 7 below expresses that for properties of the model it is irrelevant which alternative is alternative 1, which is alternative 2 and so on. One only needs to take care that upon relabeling the parameters, regression vectors, etc. are relabeled accordingly.

In our setup we have singled out the last alternative \( m \) and set \( \beta_m = 0 \) to have identifiable parameters. This changes the symmetry and needs to be accounted for. The concepts of this paper, however, are compatible with this. For example the value of the determinant of a design is equivariant:

**Proposition 7.** Let \( \sigma \in S_m \) and let \( \xi \) be any design. Then \( \xi \) is locally D-optimal for the parameters \( \beta = (\beta_1, \ldots, \beta_{m-1})^T \) if and only if \( \xi^\sigma \) is locally D-optimal for \( Q_{\sigma}^{-T} \beta \), where \( \sigma \mapsto Q_{\sigma} \) is a group homomorphism from \( S_m \) to the group of invertible \((m-1)\times(m-1)\)-matrices satisfying \( f(\sigma(i), \sigma(j)) = Q_{\sigma} f(i, j) \) for all \( \sigma \in S_m \).

**Proof.** By [21, Section 2], the design \( \xi^\sigma \) is locally optimal for the parameter \( Q_{\sigma}^{-T} \beta \) if and only if there exist matrices \( Q_{\sigma} \) as in the statement. As transpositions generate all permutations, it suffices to show the existence of such a \( Q_{\sigma} \) for all transpositions. For transpositions of \( i < m \) and \( j < m \), let \( Q_{\sigma} \) be the usual permutation matrix. For a transposition \((im)\), let \( Q_{\sigma} \) equal an identity matrix, with the \( i \)-th row replaced by the \((im)\)-row and the \( j \)-th row replaced by the row \((-1, \ldots, -1)\). Then, for an arbitrary permutation \( \sigma \), it holds that \( f(\sigma(i), \sigma(j)) = Q_{\sigma} f(i, j) \). \( \square \)

### 3. Saturated designs and graph-representation

An experimental design is saturated if its support has cardinality equal to the number of free parameters of the model. In our case of D-optimality, if a design has support size strictly smaller than \( m - 1 \), the determinant of the information matrix vanishes and optimality is impossible. A useful result about saturated designs is that their weights are completely rigid: they are all equal [23, Lemma 5.1.3] and thus only the different supports are considered. We first study which saturated designs can be D-optimal. The following simple fact is reminiscent of the connectedness of block designs with block length 2 in [22, p. 2].

**Lemma 8.** For any locally D-optimal saturated design \( \xi \) of the Bradley–Terry paired comparison model, the graph representation of the support is a tree.

**Proof.** A saturated design consists of \( m - 1 \) equally weighted comparisons. If there is a cycle \( i_1, \ldots, i_k \) in the graph representation of the design, then there is at least one alternative that does not appear in the design and therefore is represented by a disconnected vertex in the graph representation. Now, the \((m-1)\times(m-1)\)-information matrix of a saturated design is a sum of \( m - 1 \) rank-1 matrices of the form \( \lambda_{ij} f(i, j) f(i, j)^T \). For \( 1 \leq i < j \leq m - 1 \), these rank-1 matrices only have entries in the \( i \)-th and \( j \)-th rows and columns. For \( j = m \), there is only one entry \( \lambda_{im} \) in the intersection of the \( i \)-th row and \( i \)-th column. Therefore, if a saturated design contains a cycle and misses one alternative that is not \( m \), the
information matrix has no nonzero entries in either the corresponding row or the corresponding column. If alternative \( m \) is missed, it follows that every row sum of the information matrix is zero, as all rank-1 matrices are of the form \( \lambda_{ij}(e_i - e_j)(e_i - e_j)^T \). Therefore the determinant of the information matrix is zero, and the design can never be optimal. \( \square \)

Based on this fact we can determine the saturated optimal designs for the Bradley–Terry model.

**Theorem 9.** In the Bradley–Terry paired comparison model with \( m \) alternatives, if a design is saturated and locally \( D \)-optimal, then its graph representation is a path on \( [m] := \{1, 2, \ldots, m\} \).

**Proof.** Let \( \bar{\xi} \) be a saturated, locally \( D \)-optimal design for the Bradley–Terry model with \( m \) alternatives. The graph representation of \( \bar{\xi} \) is a tree by Lemma 8. Applying a suitable permutation of \( [m] \) and Proposition 7, we assume that \( \bar{\xi} \) has exactly one comparison that contains \( m \), that is, that \( m \) is a leaf. Let \( F \) be the (square) matrix of the transposed regression vectors of the design points

\[
F = \begin{pmatrix}
    f(i_1, j_1)^T \\
    f(i_2, j_2)^T \\
    \vdots \\
    f(i_{m-2}, j_{m-2})^T \\
    f(i_{m-1}, m)^T
\end{pmatrix},
\]

and define \( Q = \text{diag}(\lambda_{i_1, j_1}, \ldots, \lambda_{i_{m-1}, m}) \) as a diagonal matrix of intensities and correspondingly \( W = \text{diag}(w_{i_1, j_1}, \ldots, w_{i_{m-1}, m}) \) for the weights of the design points. Then, the information matrix is \( M(\xi, \beta) = F^T W Q F \), and inserting this into (2-2), we obtain the directional derivatives for every \( 1 \leq i < j \leq m-1 \) as

\[
\lambda_{ij} f(i, j)^T F^{-1} Q^{-1} W^{-1} F^{-T} f(i, j) - (m-1).
\]

If the design is \( D \)-optimal, this formula is nonpositive for every \( 1 \leq i < j \leq m - 1 \). Since all weights are equal to \( \frac{1}{m-1} \) this is equivalent to

\[
\lambda_{ij} f(i, j)^T F^{-1} Q^{-1} F^{-T} f(i, j) \leq 1.
\]

The proof is by downward induction. To this end, we remove one alternative and its associated design point and show that the reduced design \( \bar{\xi} \) is optimal on the reduced design space. Without loss of generality we can assume that the optimal design has only one comparison \((1, v)\) in which alternative 1 is involved. We can also assume that \( v = 2 \) using the \( S_m \) symmetry and Proposition 7. We remove alternative 1. Consider the Bradley–Terry model on the alternatives \( \{2, \ldots, m\} \). Its information matrix is a product \( F \bar{W} \bar{Q} F^T \), where \( \bar{W} \) and \( \bar{F} \) are the lower-right \((m-2) \times (m-2)\)-submatrices of \( \frac{m-1}{m-2} W \) and \( F \), respectively, and \( \bar{Q} \) is the diagonal matrix of the reduced model’s intensities \( \bar{\lambda}_{ij} \). Through our assumptions,

\[
F = \begin{pmatrix}
    1 & -1 & 0 & \cdots & 0 \\
    0 & \ddots & F \\
    0 & & \ddots & \ddots \\
    \end{pmatrix}.
\]
We show the implication
\[ \lambda_{ij} \bar{f}(i, j)^T F^{-1} Q^{-1} F^{-T} f(i, j) \leq 1 \quad \text{for all } 2 \leq i < j \leq m \]
\[ \Rightarrow \tilde{\lambda}_{ij} \tilde{f}(i, j)^T \bar{F}^{-1} \bar{Q}^{-1} \bar{F}^{-T} \tilde{f}(i, j) \leq 1 \quad \text{for all } 2 \leq i < j \leq m. \]
This implies that the design \( \tilde{\xi} \) with equal weights \( \frac{1}{m-2} \) on \( E \setminus \{1, 2\} \) is optimal for the reduced model. Since \( \tilde{\lambda}_{ij} = \lambda_{ij} \), we only have to show
\[ \tilde{f}(i, j)^T \bar{F}^{-1} \bar{Q}^{-1} \bar{F}^{-T} \tilde{f}(i, j) \leq f(i, j)^T F^{-1} Q^{-1} F^{-T} f(i, j) \]
(3-1)
for all \( 2 \leq i < j \leq m \). Now let
\[ F^{-1} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & A_1 \end{pmatrix} \]
for some \((m-2) \times (m-2)\)-matrix \( A_1 \). This leads to \( \bar{F}^{-1} = A_1 - (1/a_{11}) a_{21} a_{12}^T \). It can be checked that \( a_{21} = 0 \) and thus
\[ F^{-1} = \begin{pmatrix} 1 & a_{12} \\ 0 & A_1 \end{pmatrix}. \]
This means, that \( \bar{F}^{-1} = A_1 \). Now, as \( f(i, j)^T = (0, \tilde{f}(i, j)^T) \),
\[ f(i, j)^T F^{-1} Q^{-1} F^{-T} f(i, j) = (0, \tilde{f}(i, j)^T) \begin{pmatrix} 1 & a_{12} \\ 0 & A_1 \end{pmatrix} \begin{pmatrix} 1/\lambda_{12} \\ \bar{Q}^{-1} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ a_{12} & A_1^T \end{pmatrix} \begin{pmatrix} 0 \\ \tilde{f}(i, j) \end{pmatrix} \]
\[ = (0, \tilde{f}(i, j)^T A_1) \begin{pmatrix} 1/\lambda_{12} \\ \bar{Q}^{-1} \end{pmatrix} \begin{pmatrix} 0 \\ A_1^T \tilde{f}(i, j) \end{pmatrix} \]
\[ = \tilde{f}(i, j)^T A_1 \bar{Q}^{-1} A_1^T \tilde{f}(i, j) = \tilde{f}(i, j)^T \bar{F}^{-1} \bar{Q}^{-1} \bar{F}^{-T} \tilde{f}(i, j). \]
In fact, (3-1) is realized as an equality and the reduced saturated design is optimal. Now if \( \xi \) were not a path, iterating this procedure eventually leads to an optimal saturated design for the Bradley–Terry model on four alternatives that is not a path. Such a design does not exist by the explicit computations in Section 5. Hence, the graph representation of a saturated, locally D-optimal design is a path. \( \square \)

4. Optimality regions of saturated designs

We now describe the sets of parameters for which a saturated design from Theorem 9 is optimal. We call such a set the region of optimality of the design. Knowing these regions simplifies the experimental design problem since it can be combined with prior knowledge about the parameters (e.g., from a screening experiment). Also, knowing if the regions are big or small yields information about the robustness of designs.

Exploiting the symmetry in Proposition 7, it suffices to study a single design representing all saturated designs. This is the path \((1, 2), (2, 3), \ldots, (m-1, m)\).

Lemma 10. The optimality region of the design \((12, 23, 34, \ldots, (m-1)m)\) is defined by the inequalities
\[ g(i, j) = \lambda_{ij} \sum_{k=i}^{i-1} \frac{1}{\lambda_k(k+1)} \leq 1, \quad 1 \leq i < m, \ 1 \leq i < j \leq m. \]
Furthermore, this region is not empty.
Proof. We apply Theorem 4 to find the optimality regions of the design \((12, 23, 34, \ldots, (m-1)m)\). Therefore, one has to analyze the directional derivatives \(f(i, j)^T F^{-1} F^{-T} f(i, j) - (m-1)\), where \(f(i, j)\) are the regression vectors, \(Q\) is a diagonal matrix of the design intensities \(\lambda_{12}, \lambda_{23}, \ldots, \lambda_{(m-1)m}\) and \(F\) is the matrix of the transposed regression vectors. So,

\[
F = \begin{pmatrix}
1 & -1 & \cdots & 0 \\
0 & 1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 1
\end{pmatrix}
\quad \text{and} \quad
F^{-1} = \begin{pmatrix}
1 & \cdots & 1 \\
\vdots & \ddots & \vdots \\
0 & \cdots & 1
\end{pmatrix}.
\]

For \(i < j < m\), we have \(f(i, j) = e_i - e_j\). This leads to

\[
f(i, j)^T F^{-1} = (\mathbb{1}_{i=1}, \mathbb{1}_{i\leq 2<j}, \mathbb{1}_{i\leq 3<j}, \ldots, \mathbb{1}_{i\leq m-2<j}, 0).
\]

For \(i < j = m\), we have \(f(i, m) = e_i\). So

\[
f(i, m)^T F^{-1} = (\mathbb{1}_{i=1}, \mathbb{1}_{i\leq 2}, \mathbb{1}_{i\leq 3}, \ldots, \mathbb{1}_{i\leq m-2}, 1).
\]

This means that the directional derivative in the direction \((i, j)\) for \(j < m\) is

\[
\lambda_{ij}(m-1)(\mathbb{1}_{i\leq 1}, \mathbb{1}_{i\leq 2<j}, \ldots, \mathbb{1}_{i\leq m-2<j}, 0) \begin{pmatrix} 1/\lambda_{12} & 1/\lambda_{23} & \cdots & 1/\lambda_{(m-1)m} \end{pmatrix} \begin{pmatrix} \mathbb{1}_{i\leq 1} \\
\mathbb{1}_{i\leq 2<j} \\
\vdots \\
\mathbb{1}_{i\leq m-2<j} \end{pmatrix} = \lambda_{ij}(m-1) \sum_{k=1}^{m-2} \frac{\mathbb{1}_{i\leq k < j}}{\lambda_{k(k+1)}}
\]

and for \(j = m\) is

\[
\lambda_{im}(m-1)(\mathbb{1}_{i\leq 1}, \mathbb{1}_{i\leq 2}, \ldots, \mathbb{1}_{i\leq m-2}, 1) \begin{pmatrix} 1/\lambda_{12} & 1/\lambda_{23} & \cdots & 1/\lambda_{(m-1)m} \end{pmatrix} \begin{pmatrix} \mathbb{1}_{i\leq 1} \\
\mathbb{1}_{i\leq 2} \\
\vdots \\
\mathbb{1}_{i\leq m-2} \end{pmatrix} = \lambda_{im}(m-1) \sum_{k=1}^{m-1} \frac{\mathbb{1}_{i\leq k}}{\lambda_{k(k+1)}} = \lambda_{im}(m-1) \sum_{k=1}^{m-1} \frac{1}{\lambda_{k(k+1)}}.
\]

For \(j = i + 1\) the directional derivatives are 0 by (2-2) and Corollary 5. Let

\[
g(i, j) = \lambda_{ij} \sum_{k=i}^{j-1} \frac{1}{\lambda_{k(k+1)}}.
\]
By Theorem 4 the optimality region of the design \((12, 23, 34, \ldots, (m-1)m)\) is cut out by the inequalities 
\[ g(i, j) \leq 1 \] for \(1 \leq i < j \leq m\).

To exhibit a point in the optimality region, let \(\beta_i = i\beta_1\) and thus \(\pi_i = \pi_1^i\). This implies 
\[ \lambda_{ij} = \frac{\pi_1^{j-i}}{(1 + \pi_1^{j-i})^2}, \]
and therefore 
\[ g(i, j) = \frac{\pi_1^{j-i}}{(1 + \pi_1^{j-i})^2} \sum_{k=i}^{j-1} \frac{(1 + \pi_1)^2}{\pi_1} = \frac{(j-i)\pi_1^{j-i-1}(1 + \pi_1)^2}{(1 + \pi_1^{j-i})^2}, \]
which is at most 1 for all \(1 \leq i < j \leq m\) if just \(\pi_1\) is sufficiently large. \(\square\)

**Theorem 11.** The optimality regions of all saturated designs corresponding to paths, i.e., of all optimal saturated designs, are in the \(S_m\)-orbit of the saturated design for \((12, 23, 34, \ldots, (m-1)m)\). The optimality regions are defined by the inequalities 
\[ \{g(\sigma(i), \sigma(j)) \leq 1 : 1 \leq i < m, i < j \leq m\}, \]
where \(\sigma \in S_m\) is a permutation turning \((12, 23, 34, \ldots, (m-1)m)\) into the given path.

**Proof.** Theorem 9 shows that the saturated optimal designs correspond to paths. By Proposition 7, we can choose any representative for the orbit of path designs. We choose \((12, 23, 34, \ldots, (m-1)m)\) and plug in the results from Lemma 10. \(\square\)

## 5. Explicit solutions for four alternatives

This section studies the optimal designs for the Bradley–Terry paired comparison model with four alternatives, as it arises for example in [9]. We first deal with the case of saturated designs, i.e., optimal designs whose supports consist of only three design points. The unsaturated case with four, five or six support points follows in Section 5.2.

The Bradley–Terry paired comparison model with four alternatives has three identifiable parameters \(\beta_1, \beta_2, \beta_3\). As above we use \(\beta_i := \log(\pi_i)\) and \(\beta_4 = 0\). Our goal is to cover all of \(\mathbb{R}^3\) with regions of optimality of specific explicit designs. The regression vectors for four alternatives are 
\[ f(1, 2) = (1, -1, 0)^T, \quad f(1, 3) = (1, 0, -1)^T, \quad f(1, 4) = (1, 0, 0)^T, \]
\[ f(2, 3) = (0, 1, -1)^T, \quad f(2, 4) = (0, 1, 0)^T, \quad f(3, 4) = (0, 0, 1)^T. \]

**5.1. Saturated designs.** For saturated designs with nonsingular information matrices, the optimality criterion in Theorem 4 yields a system of inequalities in the intensities \(\lambda_{ij}\). We find these first. According to [23, Lemma 5.1.3], a saturated design has three positive weights whose values are all \(\frac{1}{3}\), the remaining weights being zero. There are \(\binom{6}{3} = 20\) possible saturated designs. Exactly 16 of them have a nonsingular information matrix. Among the 16, only 12 have a nonempty region of optimality. We find that they are in bijection with the paths on four vertices. The following theorem is the base case to which the proof of Theorem 9 reduces.
Theorem 12. For the Bradley–Terry model with four alternatives there are 20 saturated designs. Among those

- 8 have an empty region of optimality,
- 12 have optimal experimental designs.

The 12 designs with nonempty region of optimality correspond to the 12 labelings of the path $P_4$. The region of optimality of the path $(1,2),(2,3),(3,4)$ is constrained by

$$\lambda_{13}(\lambda_{12} + \lambda_{23}) - \lambda_{12}\lambda_{23} \leq 0,$$

$$\lambda_{24}(\lambda_{23} + \lambda_{34}) - \lambda_{23}\lambda_{34} \leq 0,$$

$$\lambda_{14}(\lambda_{12}\lambda_{23} + \lambda_{12}\lambda_{34} + \lambda_{23}\lambda_{34}) - \lambda_{12}\lambda_{23}\lambda_{34} \leq 0.$$

The regions of optimality for other paths arise from this by relabeling.

Since the $D$-optimality criterion is equivariant under the $S_4$ action by Proposition 7, it suffices to study one labeling for each unlabeled graph with three edges on four vertices. The proof of Theorem 12 is split into a discussion of information matrices for the three graphs in Figure 1.

5.1.1. Paths. Consider the path in Figure 1. Its edge set is $\{(1, 2), (2, 3), (3, 4)\}$. A corresponding saturated design can only be optimal if its weights are $w_{12} = w_{23} = w_{34} = \frac{1}{3}$ and $w_{13} = w_{14} = w_{24} = 0$.

The information matrix of this design is

$$M = \frac{1}{3}(\lambda_{12} f(1, 2) f(1, 2)^T + \lambda_{23} f(2, 3) f(2, 3)^T + \lambda_{34} f(3, 4) f(3, 4)^T)$$

$$= \frac{1}{3} \begin{pmatrix} \lambda_{12} & -\lambda_{12} & 0 \\ -\lambda_{12} & \lambda_{12} + \lambda_{23} & -\lambda_{23} \\ 0 & -\lambda_{23} & \lambda_{23} + \lambda_{34} \end{pmatrix}.$$ 

We apply Theorem 4. The directional derivatives are

$$g_{ij}(\lambda) := \lambda_{ij} f(i, j)^T M^{-1} f(i, j) - 3.$$ 

The region of optimality is

$$\{\lambda \in \mathbb{R}^X_{\geq 0} : g_{ij}(\lambda) \leq 0, \ 1 \leq i < j \leq 4\}.$$ 

This region is a semialgebraic set, that is, defined constructively by polynomial inequalities. To see this we use Mathematica. Corollary 5 simplifies the description because it says that for design points with positive weights the conditions become equations, and those equations have no free variables, as the weights in a saturated design are fixed. Using Mathematica’s Reduce functionality we derived (5-1).
The inequalities in (5-1) can be compared to [10, Theorem 2]. The structure is similar, but for four alternatives a cubic inequality appears. For more alternatives even higher degree inequality constraints appear according to Theorem 11. These conditions can be expressed in $\beta$-coordinates. The resulting regions of optimality are displayed in Figure 2 on the left.

5.1.2. The claw graph $K_{1,3}$. We now show that the graph in the middle of Figure 1, sometimes known as a claw, leads to an empty region of optimality. After symmetry reduction it suffices to show that the design $(12, 13, 14)$ cannot be $D$-optimal. This design would be optimal in the following region given by the three directional derivatives corresponding to the nonedges $(23, 24, 34)$:

$$
\lambda_{23} \leq \frac{\lambda_{12}\lambda_{13}}{\lambda_{12} + \lambda_{13}} \land \lambda_{24} \leq \frac{\lambda_{12}\lambda_{14}}{\lambda_{12} + \lambda_{14}} \land \lambda_{34} \leq \frac{\lambda_{13}\lambda_{14}}{\lambda_{13} + \lambda_{14}}.
$$

Plugging in the formulas for the $\lambda_{ij}$ in terms of the $\pi_i$, this becomes

$$
(\pi_2 + \pi_3)(\pi_1^2 + \pi_2\pi_3) \leq \pi_1(\pi_2 - \pi_3)^2,
$$

$$
(\pi_2 + 1)(\pi_1^2 + \pi_2) \leq \pi_1(\pi_2 - 1)^2,
$$

$$
(\pi_3 + 1)(\pi_1^2 + \pi_3) \leq \pi_1(\pi_3 - 1)^2.
$$

Using Mathematica, we find that these conditions are incompatible with $\pi_1 > 0$, $\pi_2 > 0$, $\pi_3 > 0$. It would be interesting to find a short certificate for the infeasibility of this system. Such a certificate always exists by the Positivstellensatz from real algebraic geometry (see [2]). This means that if the inequality system has no solution, then one can combine the inequalities to produce an explicit contradiction. There are computational tools to search for such certificates, but our attempts with SOStools [19] were not successful.

5.1.3. Singular designs. Designs corresponding to the rightmost graph in Figure 1 have singular information matrices and can thereby not be $D$-optimal.

Proof of Theorem 12. Since there are 12 distinct labelings of the path on four vertices, the theorem follows from the computations in Sections 5.1.1–5.1.3.

□
5.2. **Unsaturated designs.** We now examine the designs whose support contains at least four pairs. In this case the weights $w_{ij}$ of an optimal design are not necessarily uniform. Instead we find formulas that express the weights in terms of the parameters. These formulas might look complicated, but they are very symmetric and can easily be handled by computer algebra systems. Our approach is again via Theorem 4: optimality of a design $\xi^*$ is equivalent to

$$\lambda_{ij} f(i, j)^T M(\xi^*, \beta)\delta f(i, j) - 3 \leq 0, \quad 1 \leq i < j \leq 4. \quad (5-2)$$

Furthermore, by Corollary 5, there is equality for any pair $i, j$ such that $w_{ij} > 0$ in $\xi^*$. We distinguish cases according to the size of the support.

5.2.1. **Full support.** Full support means that all weights of a design are positive. Then all inequalities (5-2) hold with equality and we have a system of six equations in the variables $w_{ij}, \lambda_{ij}$ for $1 \leq i < j \leq 4$. We used Mathematica to solve the system and to express the weights $w_{ij}$ as functions of the intensities $\lambda_{ij}$:

$$w_{ij} = \frac{1}{A} \left( \lambda_{ik} \lambda_{il} \lambda_{jk} \lambda_{jl} \delta \lambda_{ij} \lambda_{ik} \lambda_{il} \lambda_{jk} \lambda_{jl} - \lambda_{ij} \lambda_{ik} \lambda_{il} \lambda_{jk} \lambda_{jl} \lambda_{kl} - \lambda_{ij} \lambda_{ik} \lambda_{il} \lambda_{jk} \lambda_{jl} \lambda_{kl} - \lambda_{ij} \lambda_{ik} \lambda_{il} \lambda_{jk} \lambda_{jl} \lambda_{kl} - \lambda_{ij} \lambda_{ik} \lambda_{il} \lambda_{jk} \lambda_{jl} \lambda_{kl} - \lambda_{ij} \lambda_{ik} \lambda_{il} \lambda_{jk} \lambda_{jl} \lambda_{kl} \right)$$

where $(i, j, k, l)$ is any permutation of $(1, 2, 3, 4)$. The term $A$ is the normalization that ensures $\sum_{i<j} w_{ij} = 1$:

$$A = 3 \left( \lambda_{ij} \lambda_{ik} \lambda_{il} \lambda_{jk} \lambda_{jl} \lambda_{kl} + \lambda_{ij} \lambda_{ik} \lambda_{il} \lambda_{jk} \lambda_{jl} \lambda_{kl} - \lambda_{ij} \lambda_{ik} \lambda_{il} \lambda_{jk} \lambda_{jl} \lambda_{kl} - \lambda_{ij} \lambda_{ik} \lambda_{il} \lambda_{jk} \lambda_{jl} \lambda_{kl} - \lambda_{ij} \lambda_{ik} \lambda_{il} \lambda_{jk} \lambda_{jl} \lambda_{kl} - \lambda_{ij} \lambda_{ik} \lambda_{il} \lambda_{jk} \lambda_{jl} \lambda_{kl} \right)$$

This design is locally optimal for some $\beta$ when $w_{ij} > 0$ for all $1 \leq i < j \leq 4$. Figure 3 shows the optimality region of six-point designs on the left.

**Example 13.** A simple example for a design with full support arises when $\beta_i = 0$ for all $1 \leq i \leq 4$. Then $\lambda_{ij} = \frac{1}{4}$ for all $1 \leq i < j \leq 4$ and therefore $w_{ij} = \frac{1}{6}$; that is, assigning the same number of repetitions to each comparison, is optimal. Figure 3 and the continuity of the formulas for $w_{ij}$ illustrate that, whenever all $\beta_i$ are sufficiently small, an optimal design will assign almost equal number of repetitions to each pair $(i, j)$.

**Remark 14.** When working with polynomial equations, Gröbner bases are a powerful tool. The expressions of the $w_{ij}$ in terms of the $\lambda_{ij}$ can also be found using elimination theory. For example, the computer algebra system Macaulay2 [12] makes this easy.
then the optimality conditions become

\[ w_{13} = \frac{2\lambda_{14}\lambda_{34}(\lambda_{14}\lambda_{34} - \lambda_{13}(\lambda_{14} + \lambda_{34}))}{3\left(\lambda_{13}^2(\lambda_{14} - \lambda_{34})^2 - 2\lambda_{13}\lambda_{14}\lambda_{34}(\lambda_{14} + \lambda_{34}) + \lambda_{14}^2\lambda_{34}^2\right)}, \]

\[ w_{14} = \frac{2\lambda_{13}\lambda_{34}(\lambda_{13}(\lambda_{34} - \lambda_{14}) - \lambda_{14}\lambda_{34})}{3\left(\lambda_{13}^2(\lambda_{14} - \lambda_{34})^2 - 2\lambda_{13}\lambda_{14}\lambda_{34}(\lambda_{14} + \lambda_{34}) + \lambda_{14}^2\lambda_{34}^2\right)}, \]

\[ w_{23} = \frac{2\lambda_{24}\lambda_{34}(\lambda_{24}\lambda_{34} - \lambda_{23}(\lambda_{24} + \lambda_{34}))}{3\left(\lambda_{23}^2(\lambda_{24} - \lambda_{34})^2 - 2\lambda_{23}\lambda_{24}\lambda_{34}(\lambda_{24} + \lambda_{34}) + \lambda_{24}^2\lambda_{34}^2\right)}, \]

\[ w_{24} = \frac{2\lambda_{23}\lambda_{34}(\lambda_{23}(\lambda_{34} - \lambda_{24}) - \lambda_{24}\lambda_{34})}{3\left(\lambda_{23}^2(\lambda_{24} - \lambda_{34})^2 - 2\lambda_{23}\lambda_{24}\lambda_{34}(\lambda_{24} + \lambda_{34}) + \lambda_{24}^2\lambda_{34}^2\right)}, \]

and

\[ w_{34} = \frac{1}{B} \left( 3\lambda_{13}^2\lambda_{14}^2\lambda_{23}^2\lambda_{24}^2 - 4\lambda_{13}\lambda_{14}\lambda_{23}\lambda_{24}\lambda_{34}^4 - 2\lambda_{13}\lambda_{14}\lambda_{23}^2\lambda_{24}^2\lambda_{34}^2 + 4\lambda_{13}\lambda_{14}\lambda_{23}\lambda_{24}\lambda_{34}^2 - 2\lambda_{13}\lambda_{14}\lambda_{23}^2\lambda_{24}\lambda_{34} + 4\lambda_{13}\lambda_{14}\lambda_{23}\lambda_{24}\lambda_{34} \right)
\]

with

\[ B = 3(\lambda_{13}^2\lambda_{14}^2 - 2\lambda_{13}\lambda_{14}\lambda_{34}^2 - 2\lambda_{13}\lambda_{14}\lambda_{34}^2 - 2\lambda_{13}\lambda_{14}\lambda_{34} + \lambda_{13}^2\lambda_{23}^2 + \lambda_{13}^2\lambda_{24}^2) \cdot (\lambda_{23}^2\lambda_{24}^2 - 2\lambda_{23}\lambda_{24}\lambda_{34} - 2\lambda_{23}\lambda_{24}\lambda_{34}^2 - 2\lambda_{23}\lambda_{24}\lambda_{34} + \lambda_{23}^2\lambda_{34}^2 + \lambda_{24}^2\lambda_{34}^2). \]
These designs are optimal if the directional derivative in (1, 2)-direction is smaller than or equal to zero, which is equivalent to

\[
\lambda_{12}(\lambda_{13}(\lambda_{23}(\lambda_{24}-\lambda_{34})-\lambda_{24}\lambda_{34})+\lambda_{34}(\lambda_{23}(\lambda_{34}-\lambda_{24})-\lambda_{24}\lambda_{34})) - \lambda_{14}\lambda_{34}(\lambda_{23}(\lambda_{24}+\lambda_{34})-\lambda_{24}\lambda_{34}) \geq -2\lambda_{13}\lambda_{14}\lambda_{23}\lambda_{24}\lambda_{34}.
\]

This inequality together with the formulas for the weights and the condition that all the weights except \(w_{12}\) are positive gives the design region. This region is nonempty. A plot in \(\beta\)-coordinates is on the right in Figure 3.

5.2.3. Four-point designs. We now discuss designs whose support contain exactly four points. There are \(\binom{4}{2} = 15\) possibilities for such designs which each have two zero weights, \(w_{ij} = w_{kl} = 0\). The four-point designs form two orbits under the action of \(S_4\), distinguished by whether the two nonedges in the graph representation share a vertex or not, that is, whether \(|\{i, j, k, l\}| = 4\), that is, \(i, j, k, l\) are all distinct or \(|\{i, j, k, l\}| = 3\), that is, exactly two are equal. In the first case, there are three different design classes.

We believe that these designs cannot be \(D\)-optimal, as the condition \(w_{ij} = w_{kl} = 0\) with \(|\{i, j, k, l\}| = 4\) implies that a third weight is zero, which would lead to a saturated design. A proof of this statement eludes us so far. Using Mathematica, it follows from the equivalence theorem that such a design satisfies

\[
\lambda_{ik}\left(w_{ik}^2 - \frac{w_{ik}}{3}\right) = \lambda_{ij}\left(w_{ij}^2 - \frac{w_{ij}}{3}\right) = \lambda_{jl}\left(w_{jl}^2 - \frac{w_{jl}}{3}\right) = \lambda_{jk}\left(w_{jk}^2 - \frac{w_{jk}}{3}\right),
\]

with 0 < \(w_{ik}, w_{ij}, w_{jk}, w_{jl}\) < \(\frac{1}{3}\) and additionally the inequalities

\[
\frac{\lambda_{ij}(3w_{il} + w_{jl}) - 2)(3w_{il} + w_{jl}) - 1}{\lambda_{jl}w_{jl}(3w_{jl} - 1)} \leq 3,
\]

\[
\frac{\lambda_{kl}(3w_{jk} + w_{jl}) - 2)(3w_{jk} + w_{jl}) - 1}{\lambda_{jl}w_{jl}(3w_{jl} - 1)} \leq 3.
\]

Among the solutions of (5-3) there are the saturated designs. If one of the weights equals \(\frac{1}{3}\), then (5-3) implies that another weight is zero, i.e., the design is saturated. Since the saturated cases have been dealt with in Theorem 12, we only look for solutions whose weights all lie in the open interval \((0, \frac{1}{3})\). There are solutions of (5-3) that satisfy this, for example, if the weights and corresponding intensities are equal. In all the cases we examined, the inequalities (5-4) and (5-5) are not satisfied.

Problem 15. Show that independent of the \(\lambda_{ij}\), a simultaneous solution of (5-3), (5-4), and (5-5) is a saturated design.

Finally we analyze the orbit of four-point designs with \(w_{ij} = w_{kl} = 0\) with \(|\{i, j, k, l\}| = 3\). Consider the representative with \(w_{12} = w_{13} = 0\). Then,

\[
w_{14} = \frac{1}{3},
\]

\[
w_{23} = \frac{2\lambda_{24}\lambda_{34}(-\lambda_{23}\lambda_{24} - \lambda_{23}\lambda_{34} + \lambda_{24}\lambda_{34})}{3(\lambda_{23}^2\lambda_{24}^2 - 2\lambda_{23}\lambda_{24}\lambda_{34} - 2\lambda_{23}\lambda_{24}\lambda_{34}^2 - 2\lambda_{23}\lambda_{24}\lambda_{34}^2 + \lambda_{23}^2\lambda_{34}^2 + \lambda_{24}^2\lambda_{34}^2)},
\]

\[
w_{24} = \frac{2\lambda_{23}\lambda_{34}(-\lambda_{23}\lambda_{24} - \lambda_{23}\lambda_{34} + \lambda_{24}\lambda_{34})}{3(\lambda_{23}^2\lambda_{24}^2 - 2\lambda_{23}\lambda_{24}\lambda_{34} - 2\lambda_{23}\lambda_{24}\lambda_{34}^2 - 2\lambda_{23}\lambda_{24}\lambda_{34}^2 + \lambda_{23}^2\lambda_{34}^2 + \lambda_{24}^2\lambda_{34}^2)},
\]

\[
w_{34} = \frac{2\lambda_{23}\lambda_{24}(-\lambda_{23}\lambda_{24} - \lambda_{23}\lambda_{34} + \lambda_{24}\lambda_{34})}{3(\lambda_{23}^2\lambda_{24}^2 - 2\lambda_{23}\lambda_{24}\lambda_{34} - 2\lambda_{23}\lambda_{24}\lambda_{34}^2 - 2\lambda_{23}\lambda_{24}\lambda_{34}^2 + \lambda_{23}^2\lambda_{34}^2 + \lambda_{24}^2\lambda_{34}^2)}.
\]
Figure 4. Assembling optimality regions for the Bradley–Terry model.

\begin{align*}
w_{24} &= \frac{2\lambda_{23}\lambda_{34}(-\lambda_{23}\lambda_{24} + \lambda_{23}\lambda_{34} - \lambda_{24}\lambda_{34})}{3(\lambda_{23}^2\lambda_{24}^2 - 2\lambda_{23}\lambda_{24}\lambda_{34} - 2\lambda_{23}\lambda_{24}\lambda_{34}^2 - 2\lambda_{23}\lambda_{34}\lambda_{34} + \lambda_{23}^2\lambda_{34}^2 + \lambda_{24}^2\lambda_{34}^2)}, \\
w_{34} &= \frac{2\lambda_{23}\lambda_{24}(\lambda_{24}\lambda_{24} - \lambda_{23}\lambda_{34} - \lambda_{24}\lambda_{34})}{3(\lambda_{23}^2\lambda_{24}^2 - 2\lambda_{23}\lambda_{24}\lambda_{34} - 2\lambda_{23}\lambda_{24}\lambda_{34}^2 - 2\lambda_{23}\lambda_{34}\lambda_{34} + \lambda_{23}^2\lambda_{34}^2 + \lambda_{24}^2\lambda_{34}^2)}.
\end{align*}

This design is optimal if the directional derivatives along \((1, 2)\) and \((1, 3)\) are smaller than 3, so if

\[
\frac{3\lambda_{12}(\lambda_{14} + \lambda_{24})}{\lambda_{14}\lambda_{24}} \leq 3 \quad \& \quad \frac{3\lambda_{13}(\lambda_{14} + \lambda_{34})}{\lambda_{14}\lambda_{34}} \leq 3.
\]

This optimality region for this four-point design is visualized in Figure 2 on the right. For each point in the optimality region, the specific weights are computed by the equations above.

Having discussed all cases, it suffices to apply the symmetry to each of these regions and then \(\mathbb{R}^3\) can be pieced together. Figure 4 gives an idea of this puzzle. Because of continuity, the boundaries between any two regions always belong to the region with fewer design points. Therefore, the yellow amoeba is open, the red regions for saturated designs are closed (by the nonstrict inequalities in Theorem 12), and all other regions have both open and closed boundaries.

**Remark 16.** Figures 3 and 2 are reminiscent of the amoebas in tropical geometry. It would be interesting to investigate if the logarithmic algebraic geometry that arises in \(\beta\)-space from the polynomial constraints in \(\lambda\)-space offers new insights.

6. Discussion

This paper explains the parameter regions of optimality for experimental design of the Bradley–Terry model, with the strongest results for four alternatives. In practical applications this knowledge can be put to use as follows: First, with a screening experiment, initial knowledge of approximate parameters is attained. The initial guess lies in one of the full-dimensional regions illustrated in Figure 4. Depending on which region it is, one can use specific knowledge about the optimal design weights \(w_{ij}\). For example,
there are explicit polynomial formulas for how the optimal weights depend on the location in parameter space. Section 5 contains explicit such formulas for the case of four alternatives.

In the case that a screening experiment reveals parameters in a region where saturated designs are optimal, the solution becomes particularly pleasant: one only needs to assign equal weights to \( m - 1 \) of the pairs. The characterization of regions of optimality of saturated designs is complete for any number of alternatives (Theorem 11).

We illustrate the effect of choosing the right design by computing the efficiency of the uniform design (assigning equal weights to all pairs) in the case of four alternatives. Consider the line in parameter space that is specified by \( 2\beta_2 = \beta_1 \), \( 4\beta_3 = 5\beta_1 \). Figure 5 shows the efficiency of the uniform design along that line. At \( \beta = (\beta_1, \beta_2, \beta_3) = (0, 0, 0) \) the uniform design is optimal. As \( \beta \) grows, the efficiency decreases. First the weights should be adjusted, and starting at approximately 1.4 a five-point design would be optimal. Around 2.1 a four-point design becomes optimal and finally, from 2.9, a saturated design is optimal. Clearly, working with a uniform design in the case that the support should be smaller is inefficient. In the limit \( \beta \to \infty \) the uniform design requires twice as many observations as the optimal saturated design.

We outline some further research directions now. For full-support designs, by Corollary 5 the region of optimality is given by the equations

\[
\lambda_{ij} f(i, j)^T M(\xi^*, \beta)^{-1} f(i, j) = (m - 1)
\]

and positivity constraints \( \lambda_{ij} > 0 \). We hope that tools from real algebraic geometry can shed further light on such semialgebraic sets, especially for designs with full support, as their semialgebraic sets contain no complicated inequalities.

The Bradley–Terry model considered here is only for the \( m \) levels of one attribute and an extension to more attributes is conceivable. The computational challenges of finding optimal designs are formidable and a nice geometry as in the present case is not expected.

In the case of optimality, the equations above express the weights of \( \xi \) in terms of the parameters. We conjecture that the equations can be solved in the following sense.

**Conjecture 17.** The \( \binom{m}{2} \) weights of a fully supported D-optimal design are rational functions in the intensities and of numerator degree \( \binom{m}{2} + m - 1 \).
An example of such expressions are the degree-9 equations in Section 5.2.1.

**Remark 18.** The solution for the four-dimensional case reveals that the numerator of a weight $w_{ij}$ is a sum of ten monomials. These monomials can be described combinatorially as follows. For simplicity, let $i = 1$ and $j = 2$. Then eight of the ten monomials are products of the squarefree monomial $\lambda_{12}\lambda_{13}\lambda_{14}\lambda_{23}\lambda_{24}\lambda_{34}$ with monomials of the form $\lambda_{ij}\lambda_{ik}\lambda_{kl}$, where $(ij, ik, kl)$ are edges of the eight graphs that are either paths or trees on four vertices and that do not contain the edge $(1, 2)$. Furthermore, the monomials that come from a graph with a node of degree 3 have a positive sign, while the monomials from paths have a negative sign. The remaining two monomials do not show such an easy structure and it remains open why they are of the form $\lambda_{12}^2\lambda_{14}^2\lambda_{23}^2\lambda_{24}^2(\lambda_{12} + 2\lambda_{34})$. The complete design is generated by permutations acting on the indices of the numerator described above, while the denominator of the weights is just the sum of all the numerators, that is, a normalization.

From the structure in the case of four alternatives, one can at least partially conjecture the structure of a solution in higher dimensions. In the case of five alternatives, we conjecture that for full-support designs the function that expresses $w_{ij}$ in the intensities $\lambda_{ij}$ satisfies the following rules: It is of the form of a polynomial divided by a normalization. The numerator polynomial is of degree $(m^2 - m) + m - 1$ (i.e., 14 for $m = 5$) and composed as follows. Start with the monomial $\lambda_{12}\lambda_{13}\cdots\lambda_{m-1,m}$. To construct the weight for the comparison $(1, 2)$, multiply it with a square-free product of $m - 1$ of the variables $\lambda_{ij}$, where $ij$ is an edge in a spanning tree on $[m]$ which does not contain $(1, 2)$. Sum these monomials over all trees that do not contain $(1, 2)$. For $n = 5$, only 50 out of the 125 trees qualify. In this summation, trees of maximal degree 2 receive a negative sign, the others a positive sign. Additionally, we may have to add monomials of a still unknown structure as in Remark 18 above. We expect a similar structure in the denominator for five alternatives as for four, so that there is a sum of monomials in the denominator that is multiplied by 4. As there are 125 trees, this would make 500 monomials from the tree-structure. This coincides with having 50 monomials from trees in the numerator, as there are ten weights for five alternatives. In comparison, for four alternatives, there are $3 \cdot 22 = 66$ monomials in the denominator, but only $6 \cdot 8 = 48$ come from the described graph structure. The implications of these observations are still unknown.

**Acknowledgement**

The authors are supported by the Deutsche Forschungsgemeinschaft DFG under grant 314838170, GRK 2297 MathCoRe.

**References**


Received 2020-08-09. Revised 2021-01-05. Accepted 2021-01-22.

THOMAS KAHLE: thomas.kahle@ovgu.de
Fakultät für Mathematik, Otto-von-Guericke Universität Magdeburg, Magdeburg, Germany

FRANK RÖTTGER: frank.roettger@unige.ch
Research Center for Statistics, University of Geneva, Geneva, Switzerland

RAINER SCHWABE: rainer.schwabe@ovgu.de
Fakultät für Mathematik, Otto-von-Guericke Universität Magdeburg, Magdeburg, Germany

mathematical sciences publishers
**Guidelines for Authors**

Authors may submit manuscripts in PDF format on-line at the submission page.

**Originality.** Submission of a manuscript acknowledges that the manuscript is original and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

**Language.** Articles are usually in English or French, but articles written in other languages are welcome.

**Required items.** A brief abstract of about 150 words or less must be included. It should be self-contained and not refer to bibliography keys. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and a Mathematics Subject Classification for the article, and, for each author, affiliation (if appropriate) and email address.

**Format.** Authors are encouraged to use \LaTeX and the standard amsart class, but submissions in other varieties of \TeX, and exceptionally in other formats, are acceptable. Initial uploads should normally be in PDF format; after the refereeing process we will ask you to submit all source material.

**References.** Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of Bib\TeX is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

**Figures.** Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages — Mathematica, Adobe Illustrator, Corel Draw, MATLAB, etc. — allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to graphics@msp.org with as many details as you can about how your graphics were generated.

Bundle your figure files into a single archive (using zip, tar, rar or other format of your choice) and upload on the link you been provided at acceptance time. Each figure should be captioned and numbered so that it can float. Small figures occupying no more than three lines of vertical space can be kept in the text (“the curve looks like this:”). It is acceptable to submit a manuscript with all figures at the end, if their placement is specified in the text by means of comments such as “Place Figure 1 here”. The same considerations apply to tables.

**White Space.** Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal’s preferred fonts and layout.

**Proofs.** Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.
Gröbner bases for staged trees
LAMPRINI ANANIADI and ELIANA DUARTE

Equivariant Hilbert series for hierarchical Models
AIDA MARAJ and UWE NAGEL

1-Wasserstein distance on the standard simplex
ANDREW FROHMADER and HANS VOLKMER

Association and Simpson conversion in $2 \times 2 \times 2$ contingency tables
SVANTE LINUSSON and MATTHEW T. STAMPS

Logarithmic Voronoi cells
YULIA ALEXANDR and ALEXANDER HEATON

The semialgebraic geometry of saturated optimal designs for the Bradley–Terry model
THOMAS KAHALE, FRANK RÖTTLER and RAINER SCHWABE