AN ALGEBRAIC MONTE-CARLO ALGORITHM
FOR THE PARTITION ADJACENCY MATRIX REALIZATION PROBLEM

Éva Czabarka, László Székely, Zoltán Toroczkai and Shanise Walker

The graphical realization of a given degree sequence and given partition adjacency matrix simultaneously, is a relevant problem in data driven modeling of networks. Here we formulate common generalizations of this problem and the exact matching problem, and solve them with an algebraic Monte-Carlo algorithm that runs in polynomial time if the number of partition classes is bounded by a constant. We note, that no deterministic polynomial time algorithm is known for any of these problems.

1. Introduction

In data driven modeling of complex networks one often needs to sample from ensembles of graphs that share characteristics with an observed network. These characteristics act as constraints for the sampling procedure and they may be reproduced exactly (“sharp constraints”) in every sampled graph or in expected value (average constraints) over the ensemble. One such natural characteristic is the degree sequence. The degree sequence, however, has many graphical realizations in general, with varying properties, e.g., either showing assortativity or disassortativity (the extent to which vertices of similar degrees are connected or not). For example, social networks tend to be assortative, while biological and technical networks tend to be disassortative. Thus, in order to model such situations, one also has to specify the degree correlations. The simplest way of achieving that is via providing the so-called joint degree matrix (JDM), whose entries are the number of edges between degree $i$ and degree $j$ vertices, for all $i$ and $j$ degree values. Note that the JDM also specifies the degree sequence itself, uniquely [7]. The JDM received considerable attention in the literature [1; 2; 7; 8; 10; 24; 26; 28; 30] and it is well understood [2; 7; 24; 30]. Reference [3] presents an exact algorithm for constructing simple graphs with a prescribed JDM.

However, to model real world networks even JDM level constraints are not always sufficient. In particular, Orsini et al. [22] demonstrate this on several networks including the internet (autonomous systems level), the distributed PGP cryptosystem, US airport network, protein interactions, brain fMRI functional networks and an English word adjacency network. To capture most of the “signal” found in the structure of a real-world network they also had to include correlations beyond degree-degree correlations, such as clustering coefficients, i.e., small subgraph correlations, called collectively as $dk$-series ($1k$ is purely degree distribution, $2k$ is joint degree distribution, etc). When trying to generate and sample graphs

Székely was supported in part by the NSF DMS, grant numbers 1300547 and 1600811.

MSC2010: primary 05D15; secondary 05C82, 05D40, 68W20.

Keywords: perfect matching, exact matching, degree sequence, bipartite graph, Hall’s theorem, joint degree matrix, partition adjacency matrix, Monte-Carlo algorithm, computational complexity.

115
with these prescribed correlations, these authors observe that already at the \( d = 3 \) level the process does not converge and modeling essentially fails. As briefly discussed in [22], this graph generation process can be described as a Boolean constraint satisfaction SAT problem, in which the variables are the elements of the adjacency matrix whose values need to be set (to 0 or 1) such that a set of constraints expressed in terms of functions of the marginals (degrees) are satisfied. From this point of view it is thus expected that the problem eventually becomes NP-complete (3-SAT), which indeed was experienced in [22], through the failure of the algorithms to converge.

The graph construction problems above all have their constraints related to some graph structural properties. However, in many real world situations there are also externally imposed constraints, such as group membership that is not modeled by the approaches above. For example, one might study a social network at different levels of resolution: we may look at a large organization as a network of interactions between teams or departments but also at the connections between the individuals throughout the organization and ask questions related to the performance of the organization as a whole as function of these networks. One can certainly think of the teams and departments as a partitioning of the individuals into groups/classes and the connections between them as a partitioning of the edges. Similarly, we may study social networks in relation to gender partitions, political party, religious, or any other affiliations. While these are natural questions, their studies in the literature are lacking due to the fact that these externally imposed partitions make the problems very difficult, as also illustrated here. To include such classes of partition induced constraints in network modeling, the first author, in 2014 introduced the concept of the partition adjacency matrix (PAM) [6]. Given a simple graph and a partition of its vertex set, entries of the PAM count the number of edges between, and within the partition classes. If the partition consists of singleton sets of vertices, then it specializes to the familiar adjacency matrix of the graph, while the JDM of a graph is a special PAM, with all classes composed of vertices having the same degree. Another motivation for working with PAM constraints is related to the construction of Chung–Lu type models [5] (used extensively in modeling applications), where edges are selected randomly and independently obtaining, in expectation, graph samples with a desired degree sequence. It is relatively simple to define a Chung–Lu type model constrained by an expected degree sequence and a desired PAM on a specified partition. Should we want to construct a Chung–Lu type model based on given expected degree sequence and expected JDM, however, we would run into difficulties. This is because, due to the random and independent selection of the edges, the exact degree of a vertex, and therefore the degree partition of the graph, changes from sample to sample, making the model difficult to specify, parametrically. This can be resolved with a proper PAM formulation.

Reference [7] shows that the space of all realizations of a given JDM over a fixed vertex set is connected via swapping certain pairs of edges (giving a Markov basis), and gives a simple condition that characterizes which matrices are JDMs (a JDM determines the degree sequence). One can extend these results [6] to graphs with a given degree sequence when replacing JDM with PAM, as long as degrees within a partition class do not differ by more than 1. It was also shown that the space of realizations of graphs with a given degree sequence and PAM is no longer connected via such simple swaps, if we allow larger degree differences.
In contrast with the JDM, however, much less is known about PAMs, which are, as explained above, an important notion in data driven modeling of networks. While a JDM determines the degree sequence of a realizing graph, a PAM does not, in general. Similarly to JDM problems, PAM problems include existence or realization (Is there a simple graph with a given degree sequence and given PAM?), construction (Provide an algorithm that constructs such simple graphs!), sampling (Formulate an algorithm that can sample such graphs with prescribed probability distribution!) and counting problems (How many simple graphs are there to realize a given degree sequence and given PAM?), in increasing order of their difficulty. Here we focus on the existence and construction problems.

**Partition adjacency matrix realization problem.** Given a set $W$ and natural numbers $d(w)$ associated with $w \in W$, a $W_i : i \in I$ partition of $W$, and natural numbers $c(W_i, W_j)$ associated with unordered pairs of partition classes, is there a simple graph $G$ on the vertex set $W$ with degree $d(w)$ for every $w \in W$, and with exactly $c(W_i, W_j)$ edges with endpoints in $W_i$ and $W_j$? If yes, we say that $G$ is a realization for the given degree sequence $\{d(w)\}$ and given PAM constraints $\{c(W_i, W_j)\}$.

**Partition adjacency matrix construction problem.** Construct such a graph, if the answer to the realization problem is affirmative.

Reference [11] conjectures that the realization problem is NP-complete, and here we also support this conjecture. The skeleton of a PAM is the graph, whose vertices are the partition classes, and two partition classes, $W_i$ and $W_j$ are joined by an edge, if $c(W_i, W_j) > 0$. Reference [11] found polynomially solvable instances of the realization problem for two partition classes ($|I| = 2$), and also for loopless unicyclic skeleton graphs. In the Bipartite PAM problem the skeleton graph is bipartite and loopless.

A stronger version of the problems above is when there is also a forbidden subgraph that all graphical realizations must avoid. Such problems arise in part for algorithmic reasons in direct construction algorithms that add edges sequentially: the existing edges forbid the addition of further edges between the same pairs of vertices in the graph being constructed [3; 16]. Thus, we formulate:

**Partition adjacency matrix realization/construction problems in the presence of a blue graph.** In addition to the contraints of the PAM realization problem, a graph $B$ (the blue graph) is given on the vertex set $W$. Is there a realization that is not using any edges from $B$? If yes, construct such a graph.

To simplify the discussion, we assume that for any PAM realization/construction problem, the obvious and easy-to-check necessary conditions that $d$ is the degree sequence of a simple graph, and that

$$\sum_{v \in W_i} d(v) = c(W_i, W_i) + \sum_j c(W_i, W_j),$$

$$\sum_{i} d(v) = \sum_i c(W_i, W_i) + \sum_i \sum_j c(W_i, W_j)$$

hold. In an earlier version of this manuscript [9], we gave an algebraic Monte-Carlo algorithm for the blue graph version of the bipartite PAM realization problem. This algorithm runs in polynomial time if the number of partition classes is bounded by a constant. We are indebted to András Frank (Budapest), who
kindly called our attention to the analogous exact matching problem,\footnote{http://lemon.cs.elte.hu/egres/open/Exact_matching_in_red-blue_bipartite_graphs} and asked if the two problems admit a common generalization, i.e., a third problem, of which the first two problems are specific instances, such that third problem allows algorithmic solution like the first two. Let us recall:

**Exact matching problem.** Given a graph $G$, whose edges are colored red or green, is there a perfect matching with exactly $m$ red edges in the matching?

A matching in a graph $G$ is a set of independent edges from $G$, i.e., no two edges of this set share a common vertex. A matching $M$ is maximal if no more independent edges can be added to it from $G$. A maximum matching is a maximal matching of largest cardinality and this number $\nu(G)$ is called the matching number of $G$. We have $\nu(G) \leq n/2$, where $n$ is the number of vertices of $G$. Any matching (if it exists) with $\nu(G) = n/2$ is called a perfect matching (and thus only graphs with even number of vertices can have perfect matching). In a graph with perfect matching all vertices are incident on an edge from the perfect matching (all vertices are matched).

The exact matching problem originates from Papadimitriou and Yannakakis \cite{PAP}, and Lovász proposed a Monte-Carlo algorithm for it. Lovász’s algorithm, which was not published, is based on the general ideas in his paper \cite{LOV}, and is described by Mulmuley, Vazirani and Vazirani in \cite[p. 111]{MUL}. No deterministic polynomial time algorithm is known for the exact matching problem. Here we provide the promised common generalization:

**Dominating f-factor problem.** Given a graph $G$ on $n$ vertices, disjoint subsets $E_1, \ldots, E_k \subset E(G)$, integers $m_1, \ldots, m_k$, and prescribed degrees $d_1, \ldots, d_n$ associated with the vertices $v_1, \ldots, v_n$ of $G$, is there a subgraph $G'$ of $G$, such that $v_i$ has degree $d_i$ in $G'$ for all vertices, and $G'$ has at least $m_j$ edges from the edge set $E_j$, for all $j = 1, \ldots, k$?

**Dominating matching problem.** Given a graph $G$, disjoint subsets $E'_1, \ldots, E'_k \subset E(G)$, integers $m_1, \ldots, m_k$, is there a perfect matching in $G$, which uses at least $m_j$ edges from the edge set $E'_j$, for all $j = 1, \ldots, k$?

Clearly the dominating matching problem is a special case of the dominating f-factor problem, where every degree is one. In Section 2, we will show using the Tutte gadget that the dominating f-factor problem can be solved through solving a dominating matching problem on about $n^2$ vertices. The exact matching problem is an instance of the dominating matching problem, where $k = 2$, $E_1$ is the set of red edges, $E_2$ is the set of green edges, $m_1 = m$, $m_2 = |E(G)| - m$. The PAM realization problem is an instance of the dominating f-factor problem in the following way: $G$ is the complement of the blue graph $B$, the disjoint edge subsets are $E_{ij} = \{u, v \in E(G) : u \in W_i, v \in W_j\}$ for $i \leq j$ (assuming without loss of generality that $I$ is an ordered set), $m_{ij} = c(W_i, W_j)$ for $i < j$ and $m_{ii} = c(W_i, W_i)$.

Here we provide an algebraic Monte-Carlo algorithm for the dominating matching problem, and hence for the dominating f-factor problem, which runs in polynomial time under the assumption that

$$\prod_{i \in I} (m_i + 1) = O(\text{polynomial}(n)).$$
This assumption certainly holds if \(|I|\) stays bounded by a constant, while \(n\) grows. If the algorithm returns \(\text{TRUE}\), then the sought-after graph exists, if the algorithm returns \(\text{FALSE}\), then with high probability (whp) such a graph does not exist. The realization algorithm is described in Section 3, together with its correctness, which hinges on the Schwartz–Zippel lemma [29]. We will also conclude that constructing an actual solution is not harder than the decision problem. We end the paper with some complexity results in Section 4.

2. The Tutte gadget

Clearly, the standard degree sequence realization problem is a relaxation of the PAM problem, where we do not care about satisfying the \(c_{ij}\) conditions. Havel [15] and Hakimi [14] solved the degree sequence realization problem and Ryser [27] solved the bipartite degree sequence realization problem. We next use a result of Tutte [33] to connect degree sequence realization to the existence of a perfect matching in a bigger graph, the Tutte gadget.

Initially, we are given a degree sequence realization problem on the vertices in \(V\), i.e., for each \(v \in V\) we are given a proposed degree \(d(v)\). We are also given a set of blue — or forbidden — edges \(B\) that our realization is not allowed to use. For a vertex \(v \in V\), let \(N_B(v) = \{u : \{u, v\} \in B\}\) denote the set of blue neighbors of \(v\), and \(S_v = V \setminus (\{v\} \cup N_B(v))\) denote the set of allowed neighbors. Without restrictions in the degree sequence realization problem, we have \(S_v = V \setminus \{v\}\). The setup of this problem implies that \(u \in S_v\) if and only if \(v \in S_u\), and we will also assume further that for each \(v \in V\), \(|S_v| \geq d(v)\) holds (otherwise a realization obviously cannot exist).

The Tutte gadget of the degree sequence realization problem with a set \(B\) of blue edges is a graph \(T\) such that

\[
V(T) = \{v^u : v \in V, u \in S_v\} \cup \{a^v_1, \ldots, a^v_{|S_v| - d(v)} : v \in V\},
\]

and

\[
E(T) = \{(v^u, u^v) : v \in V, u \in S_v\} \cup \{(v^u, a^v_i) : v \in V, u \in S_v, i = 1, 2, \ldots, |S_v| - d(v)\}.
\]

For the degree sequence realization problem, with \(B = \emptyset\), for each \(v \in V\), \(S_v = V \setminus \{v\}\) and the degree condition becomes \(d(v) \leq |V| - 1\). The Tutte gadget is a graph with \(2n(n - 1) - \sum_v d(v)\) vertices.

The Tutte gadget is relevant for the following property: it has a perfect matching if and only if a graph solves the corresponding degree sequence realization problem; furthermore, if some \(\{w^u, u^w\}\) edges are present in the perfect matching, then the corresponding \(\{w, u\}\) edges provide a graph solving this degree sequence realization problem, and if some \(\{w, u\}\) edges provide a graph solving the degree sequence realization problem, then the corresponding \(\{w^u, u^w\}\) edges in \(T\) are part of a perfect matching of \(T\). This property is well-known and is also easy to verify.

As mentioned earlier, the PAM realization problem is an instance of the dominating f-factor problem.

In the dominating f-factor problem we essentially have a degree sequence realization problem in the presence of a blue graph (the complement of \(G\)) with additional sets \(E_i \subseteq E(G)\) and parameters \(m_i\).

The dominating f-factor problem is equivalent to a dominating matching problem on the Tutte gadget \(T\), where we define the edge sets \(E'_i\) by \(E'_i = \{\{w^u, u^w\} \in E(T) : \{u, w\} \in E_i\}\).
3. The dominating matching problem

We denote the $j$-th entry of the $i$-th row in matrix $A$ by $i[A]_j$. Let $A$ be a skew-symmetric matrix, i.e., $A = -A^T$, and assume that $A$ has an even order $2n$. The Pfaffian of $A$ is defined as

$\text{Pf}(A) = \sum_{\pi} \text{sign}(\pi) \cdot i_1[A]_{j_1} \cdot i_2[A]_{j_2} \cdots i_n[A]_{j_n},$

where $\pi$ runs through permutations of the form

$\pi = \begin{pmatrix} 1 & 2 & 3 & 4 & \cdots & 2n-1 & 2n \\ i_1 & j_1 & i_2 & j_2 & \cdots & i_n & j_n \end{pmatrix}$

under the assumptions $i_1 < j_1$, $i_2 < j_2$, ..., $i_n < j_n$ and $i_1 < i_2 < \cdots < i_n$, and $\text{sign}(\pi) = \pm 1$, the sign of the permutation $\pi$. For more background on the Pfaffian, see [18]. Cayley [4] and Muir [19; 20] proved that $(\text{Pf}(A))^2 = \text{det}(A)$. Note that the summation for $\pi$ can be thought of as a summation over the perfect matchings of $2n$ elements.

Assume now that we are given a graph $G$ for the dominating matching problem. We will assume that the graph has an even number of vertices, say $2n$, otherwise it cannot have a perfect matching. Fix an arbitrary orientation $\vec{G}$ of the graph $G$. For the arc $i \to j$ introduce a variable $x_{ij}$, and define $A$ by

$i[A]_j = \begin{cases} x_{ij} & \text{if } i \to j, \\ -x_{ij} & \text{if } j \to i, \\ 0 & \text{otherwise.} \end{cases}$

The variables $x_{ij}$ are independent of each other. There is a unique way to list a perfect matching of the elements of $[2n]$ in the form $i_1j_1, i_2j_2, \ldots, i_nj_n$ such that for all $k$ we have $i_k < j_k$ and $i_1 < i_2 < \cdots < i_n$. Therefore the nonzero terms of $\text{Pf}(A)$ in the summation and the perfect matchings in $G$ correspond to each other. As cancellation of terms is not possible, $G$ has a perfect matching if and only if the polynomial $\text{Pf}(A)$ is not termless, i.e., not the zero polynomial. Tutte’s theorem [32], that $G$ has a perfect matching if and only if the polynomial $\text{det}(A)$ is not the zero polynomial follows from Cayley’s theorem that $(\text{Pf}(A))^2 = \text{det}(A)$. Introduce now additional new variables, $z_\ell$ associated with the edge set $E^\ell$, for $\ell = 1, 2, \ldots, k$. Define the matrix $A^*$ by the substitutions $x_{ij} \leftarrow x_{ij}z_\ell$ for all $(i, j) \in E^\ell$ in $A$, and not changing $x_{ij}$ if $(i, j) \notin \bigcup_{\ell=1}^k E^\ell$. A matching that defines a term in $\text{Pf}(A)$ solves the dominating matching problem if and only if for every $\ell$, the exponent of $z_\ell$ is at least $m_\ell$, for $\ell = 1, 2, \ldots, k$.

Now we need some properties of the difference operator acting on multivariate polynomials. For a polynomial $f(x, y, z, \ldots)$, set

$\nabla_x f = f(x, y, z, \ldots) - f(x - 1, y, z, \ldots).$

We will use products of these operators to indicate juxtaposition, and consequently $\nabla_x^k$ will denote the repetition of the operator $\nabla_x$ $k$ times; $\nabla_x^0$ is the identity operator. Note that unless $f$ is identically zero, applying $\nabla_x$ strictly decreases the degree of $x$ in the polynomial. Therefore, if the degree of $x$ in $f$ is less
than \( k \), then \( \nabla_x^k f \) is identically 0, and \( \nabla_x^k x^k = k! \neq 0 \). It is well-known (see [31, p. 10, (7) and (8)]) that

\[
\nabla_x^k f(x) = \sum_{\ell=0}^{k} (-1)^\ell \binom{k}{\ell} f(x - \ell).
\]

Furthermore, as

\[
\nabla_x \nabla_y f = f(x, y, z, \ldots) - f(x - 1, y, z, \ldots) - f(x, y - 1, z, \ldots) + f(x - 1, y - 1, z, \ldots) = \nabla_y \nabla_x f,
\]

the order of \( \nabla \) operators associated with different variables is freely interchangeable. For any function \( f \) in variables \( z_\ell \), and possibly other variables not shown, the following iterated difference, which is put into product notation, can be computed formally:

\[
\left( \prod_{\ell=1}^{k} \nabla_{z_\ell}^{m_\ell} \right) f(z_1, \ldots, z_\ell, \ldots, z_k) = \sum_{u_1=0}^{m_1} \cdots \sum_{u_\ell=0}^{m_\ell} \cdots \sum_{u_k=0}^{m_k} \left( \prod_{\ell=1}^{k} (-1)^u_\ell \binom{m_\ell}{u_\ell} \right) f(z_1 - u_1, \ldots, z_\ell - u_\ell, \ldots, z_k - u_k). \tag{2}
\]

We are ready to claim the key fact behind our algorithm: the polynomial

\[
\left( \prod_{\ell=1}^{k} \nabla_{z_\ell}^{m_\ell} \right) \text{Pf}(A^*(z_1, \ldots, z_k)) = \sum_{u_1=0}^{m_1} \cdots \sum_{u_\ell=0}^{m_\ell} \cdots \sum_{u_k=0}^{m_k} \left( \prod_{\ell=1}^{k} (-1)^u_\ell \binom{m_\ell}{u_\ell} \right) \text{Pf}(A^*(z_1 - u_1, \ldots, z_\ell - u_\ell, \ldots, z_k - u_k)) \tag{3}
\]

is not identically 0 if and only if the dominating matching problem has a solution, as no monomial can be a multiple of another. Thus, the dominating matching problem boils down to checking whether the polynomial (3) is identically zero or not. Make random substitutions into all variables of (3), if this polynomial is not identically zero, then whp after a number of substitutions we obtain a nonzero value. In this case the answer to the problem is a (correct) TRUE. If we always get zero values, the answer returned is a FALSE, and it is correct whp. (We give a more detailed analysis below.) From a computational point of view, the issue is whether we can compute substituted values of (3) in polynomial time. Note that the Pfaffian with integer entries (or with entries from an integral domain) can be evaluated efficiently, similarly to the evaluation of a determinant [12].

While the polynomial \( \text{Pf}(A^*) \) is not directly computable in polynomial time, the result of substituting numbers into all variables is. Indeed, (3) expanded in (2) with \( f(z_1, \ldots, z_\ell, \ldots, z_k) = \text{Pf}(A^*) \) is just a weighted sum of values of \( \text{Pf}(A^*) \) after the substitutions \( z_\ell \leftarrow z_\ell - u_\ell \) (\( \ell = 1, 2, \ldots, k \)) for every \( 0 \leq u_\ell \leq m_\ell \). In other words, for every attempt to substitute random numbers, we have to evaluate \( \prod_{\ell=1}^{k} (m_\ell + 1) \) numerical Pfaffians, a polynomial number of steps in \( n \).

Recall the Schwartz–Zippel lemma [29], where nonzero polynomial means that at least one term comes with nonzero coefficient.
**Lemma 1.** For a field $\mathbb{F}$, let $f \in \mathbb{F}[x_1, x_2, \ldots, x_t]$ be a nonzero polynomial of degree $d$ and $\Omega \subseteq \mathbb{F}$ a finite set, $|\Omega| = N$. Let $Z(f, \Omega)$ denote the set of roots from $\Omega$, i.e.,

$$Z(f, \Omega) = \{ (\alpha_1, \alpha_2, \ldots, \alpha_t) \in \Omega^t : f(\alpha_1, \alpha_2, \ldots, \alpha_t) = 0 \}.$$ 

Then $|Z(f, \Omega)| \leq dN^{t-1}$, and the probability that $f$ vanishes on random elements of $\Omega^t$, where each coordinate is selected independently and uniformly at random is at most $d/N$.

The polynomial (3) has degree at most $2n$. Let $p$ be a prime number, such that $p \geq 2n^2$. One can find such a prime using Bertrand’s postulate (better estimates on gaps between primes exist) and prime testing the numbers one after the other. Set $\mathbb{F} = \Omega = \text{GF}(p)$. We compute (3) in $\text{GF}(p)$, i.e., we do the calculations mod $p$. Note that the polynomial (3) is nonzero over $\text{GF}(p)$ as well if a solution to the dominating matching problem exists, since after taking the derivatives we get coefficients at the terms that are products of numbers at most $2n$.

Substituting randomly and uniformly selected elements of $\text{GF}(p)$ into the variables of $A^*$ and its translates, the probability of getting a 0 value for the expression (3) if it is not the identically 0 polynomial, is at most $2n/N = 2n/p \leq 1/n$, according to the lemma.

**Theorem 2.** There is a Monte-Carlo algorithm for the dominating matching problem and the dominating $f$-factor problem, which runs in polynomial time under the assumption that $\prod_{i \in I}(m_i + 1) = O(\text{polynomial}(n))$, which certainly holds if $|I|$ stays bounded by a constant. If the algorithm returns TRUE, then the sought after graph exists, if the algorithm returns FALSE, then with high probability (whp) such a graph does not exist.

An actual solution easily can be found by testing iteratively whether an edge can be included in the matching in a modified problem, a standard approach [21; 16].

**Theorem 3.** There is a Monte-Carlo algorithm to construct a solution for the dominating matching problem or the dominating $f$-factor problem, which runs in polynomial time under the assumption that $\prod_{i \in I}(m_i + 1) = O(\text{polynomial}(n))$, which certainly holds if $|I|$ stays bounded by a constant. If the algorithm returns a construction, then it is a correct solution, and if a correct solution exists, a construction is found whp.

### 4. Concluding remarks

Our algorithm for the dominating matching problem, if specialized for the exact matching problem, is different from Lovász’s algorithm [21]. We believe, however, that the same techniques may also be used to solve the dominating matching problem.

Here we did not focus on optimizing the algorithms and estimating their running times, as that is planned for subsequent work. We repeat here that no deterministic polynomial time algorithm is known for the exact matching problem, not even for bipartite graphs. Hence no deterministic polynomial time algorithm is known for the dominating $f$-factor and dominating matching problems.

We are thankful to Stefan Lendl (Graz) for bringing to our attention reference [25]. That article shows that given a bipartite graph $G$ and a partition $V_1, \ldots, V_s$ and $U_1, \ldots, U_\ell$ of the partition classes, the
decision problem whether a perfect matching $M$ exists with at most one edge between any pair of partition classes is NP-complete. It is easy to see that this problem is equivalent to the following instance of the dominating f-factor problem: the graph is $G$, the prescribed degree is $d_G(v) - 1$ for vertex $v$, the $E_{ij}$ edge sets are $E(G) \cap (V_i \times U_j)$, and $m_{ij} = |E(G) \cap (V_i \times U_j)| - 1$. Hence the dominating f-factor problem is also NP-complete.

Reference [13] noted that the 3-dimensional perfect matching problem in 3-partite graphs can be reduced to the problem of finding a multicolored perfect matching in an $n$-colored bipartite graph $K_{n,n}$. This gives another proof for the fact that the dominating matching problem is NP-complete. Reference [11] conjectures that the PAM realization problem (with empty blue graph) is already NP-complete.

References


Received 2019-09-29. Revised 2021-04-07. Accepted 2021-06-07.

ÉVA CZABARKA: czabarka@math.sc.edu
Department of Mathematics, University of South Carolina, Columbia, SC, United States
and
Visiting Professor, Department of Pure and Applied Mathematics, University of Johannesburg, South Africa

LÁSZLÓ SZÉKELY: szekely@math.sc.edu
Department of Mathematics, University of South Carolina, Columbia, SC, United States
and
Visiting Professor, Department of Pure and Applied Mathematics, University of Johannesburg, South Africa

ZOLTÁN TOROCZKAI: toro@nd.edu
Department of Physics, Department of Computer Science and Engineering, University of Notre Dame, IN, United States

SHANISE WALKER: walkersg@uwec.edu
Department of Mathematics, University of Wisconsin-Eau Claire, WI, United States
BIMONOTONE SUBDIVISIONS OF POINT CONFIGURATIONS IN THE PLANE

ELINA ROBEVA AND MELINDA SUN

Bimonotone subdivisions in two dimensions are subdivisions all of whose sides are vertical or have nonnegative slope. They correspond to statistical estimates of probability distributions of strongly positively dependent random variables. The number of bimonotone subdivisions compared to the total number of subdivisions of a point configuration provides insight into how often the random variables are positively dependent. We give recursions as well as formulas for the numbers of bimonotone and total subdivisions of $2 \times n$ grid configurations in the plane. Furthermore, we connect the former to the large Schröder numbers. We also show that the numbers of bimonotone and total subdivisions of a $2 \times n$ grid are asymptotically equal. We then provide algorithms for counting bimonotone subdivisions for any $m \times n$ grid. Finally, we prove that all bimonotone triangulations of an $m \times n$ grid are connected by flips. This gives rise to an algorithm for counting the number of bimonotone (and total) triangulations of an $m \times n$ grid.

1. Introduction

In this paper we study bimonotone subdivisions in the plane, which are intimately related to nonparametric density estimation (see Section 1.1). Let $\mathcal{A} = \{a_1, \ldots, a_m\} \subset \mathbb{R}^2$ be a point configuration in the plane. A subdivision of $\mathcal{A}$ is a collection of convex polygons whose vertices lie in $\mathcal{A}$ such that the union of the polygons is the convex hull of $\mathcal{A}$ and each pair of polygons either does not intersect or intersects at a common vertex or side. A triangulation of $\mathcal{A}$ is a subdivision of $\mathcal{A}$ for which all polygons are triangles. For example, in Figure 1, the leftmost and rightmost drawings are subdivisions, and the rightmost is also a triangulation. The second is not a subdivision because two distinct polygons intersect in their interiors, and the third is not a subdivision because one of the polygons is not convex. Note that not all points in $\mathcal{A}$ need to be used as vertices of the polygons in the subdivision/triangulation. For more details on subdivisions and triangulations, please refer to the textbook [5].

MSC2020: 52C05.
Keywords: bimonotone, subdivisions, triangulations, log-concave densities, log-supermodular densities, MTP$_2$. 

Figure 1. Examples of subdivisions and nonsubdivisions.
Definition 1. A bimonotone polygon is a polygon for which all edges have either vertical or nonnegative slope. A bimonotone subdivision is a subdivision for which all component polygons of the subdivision are bimonotone.

For example, Figure 2 below shows two subdivisions, one of which is bimonotone. Bimonotone polytopes are precisely those convex polytopes that are closed under taking coordinatewise minima and maxima of pairs of points. They were studied in the 1970s by George Bergman and discussed in [1] and [19]. Bimonotone polytopes were later studied in computer science [14] and in discrete geometry under the name distributive [8].

1.1. Motivation. For a point configuration \( A = \{a_1, \ldots, a_m\} \) and a set of heights, or tent poles, \( \{h_1, \ldots, h_m\} \), one above each of the points in \( A \), we can define a tent function as the smallest concave function whose value at \( a_i \) is at least as big as \( h_i \) for each \( i = 1, \ldots, m \). In other words, the tent function is formed by spreading a “tarp” over the poles, see Figure 3. Each tent function induces a subdivision of \( A \) composed of the polygons in the plane above which the tent function is linear. In fact, the subdivisions of \( A \) that arise from a tent function are called regular subdivisions. For most point configurations, there exist subdivisions that are not regular [5].

Tent functions show up in the field of nonparametric statistics [21] as the estimates of log-concave densities [4]. (A function is log-concave if its logarithm is concave.) More precisely, if the points in \( A \)
are the samples drawn from an unknown log-concave density \( p \), then, the maximum likelihood estimate of \( p \) will be \( \hat{p} = \exp(f) \), where \( f \) is a tent function with tent poles centered at the samples \( \mathcal{A} \) [4]. It was recently shown that if the unknown density \( \hat{p} \) is log-concave and log-supermodular (also known as multivariate totally positive of order 2, or MTP\(_2\), see Definition 2), then the maximum likelihood estimate of \( p \) is a density \( \hat{p} = \exp(f) \), where \( f \) is a tent function which induces a bimonotone subdivision of the point configuration \( \mathcal{A} \).

**Definition 2.** A function \( f : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty\} \) is supermodular if

\[
f(x) + f(y) \leq f(\min(x, y)) + f(\max(x, y))
\]

for all \( x, y \) in the domain of \( f \). A density \( p \) is log-supermodular, or multivariate totally positive of order 2 (MTP\(_2\)) if \( p = \exp(f) \), where \( f \) is supermodular.

If the density \( p \) of a random vector \( X = (X_1, \ldots, X_d) \) is MTP\(_2\), then, the coordinates of \( X \) are strongly positively dependent on each other. In fact, MTP\(_2\) implies another strong form of positive dependence called positive association [9]. Despite it being such a strong form of dependence, the MTP\(_2\) property holds for a variety of real-world distributions as well as many well-studied families of distributions [2; 6; 7; 12; 18].

For this reason, the authors of [15] study the problem of estimating a log-concave and MTP\(_2\) density. They show that if the samples lie in \( \mathbb{R}^2 \), the maximum likelihood estimate \( \hat{p} \) equals \( \exp(f) \), where \( f \) is a tent function, which induces a bimonotone subdivision [15]. One of the important remaining questions is that of characterizing how large the family of log-concave and MTP\(_2\) densities is, especially compared to the family of log-concave densities. Knowing this would shed light on the statistical complexity of the problem of estimating log-concave and MTP\(_2\) densities. Log-concave density estimation is a well-studied problem [17]. The maximum likelihood estimator was described by Walther in one dimension [20] and by Cule, Samworth and Stewart in higher dimensions [4]. A lot of work has gone into finding the sample complexity rate of the maximum likelihood estimator, most notably in [11] and later in [3]. The nonasymptotic lower bounds found in [11] show that the supremum risk of the log-concave maximum likelihood estimator for the squared Hellinger loss function is at least order \( n^{-2/(d+1)} \) in dimension \( d \) with \( n \) samples. Since the dimension \( d \) is in the exponent, this problem suffers from the so called curse of dimensionality, i.e., one needs too many samples in order to obtain an estimate close to the true density in high dimensions. A review of recent progress on log-concave density estimation can be found here [17].

If the log-concave and MTP\(_2\) family is smaller, then its statistical complexity could be lower than that of the log-concave family. We here do not compute the sample complexity of the maximum likelihood estimator of log-concave and MTP\(_2\) densities. Instead, we compare how common bimonotone subdivisions are among all subdivisions of a given point configuration. This measures the abundance of tent functions which are supermodular. We show that in dimension 2 the number of bimonotone subdivisions of a grid configuration is asymptotically equal to the number of all of the regular subdivisions of such a configuration (see Theorems 5 and 6). This suggests that the two families discussed above may be asymptotically the same in size. Our findings are consistent with consequent work in progress [16] that shows that in dimension 2 the sample complexities of estimating log-concave densities and log-concave
and MTP₂ densities are asymptotically equal. We leave the same computation in dimension \( d \geq 3 \) to future work, conjecturing that in these cases the MTP₂ and log-concave family is significantly smaller than the log-concave family.

Bimonotone polytopes and subdivisions were studied in computer science [14] and in discrete geometry under the name *distributive* [8]. More notably for us, bimonotone subdivisions are tightly related to supermodular set functions. Such a function can equivalently be thought of as a supermodular function on \( \{0, 1\}^d \). The Lovász extension theorem [13] shows that such a supermodular function on the vertices \( \{0, 1\}^d \) of the unit hypercube \([0, 1]^d\) can be extended to a concave supermodular function on the whole hypercube. This is done precisely by considering the values of the function on \( \{0, 1\}^d \) as tent poles and taking the tent function that they give rise to [4]. For more on the combinatorics and geometry of submodular functions please refer to [10]. Triangulations of polygons can be counted using the flip graph method [5]. It is also known that the number of triangulations of an \( n \)-sided polygon is equal to the \((n-2)\)-th Catalan number [5]. However, little research has been conducted on the number of subdivisions and triangulations of grids, or of bimonotone subdivisions and triangulations of a general point configuration. We focus on this in the present work.

1.2. Organization of the paper. In Section 2, we study point configurations \( \mathcal{A} \subset \mathbb{R}^2 \) whose points lie on two rows of a rectangular grid. We provide a recursion for the number of bimonotone and total subdivisions of grids with two rows of possibly different numbers of points. We use these recursions in Section 3 to find a general formula for these numbers. Furthermore, we show that the number of bimonotone subdivisions of a \( 2 \times n \) grid is equal to \( 2^{n-2} \) multiplied by the \( n \)-th Schröder number. In Section 4 we show that bimonotone triangulations using all of the vertices of a grid configuration \( \mathcal{A} \) form a connected flip graph. Using this result, in Section 5, we present two algorithms for counting the numbers of bimonotone and total subdivisions and triangulations of two-dimensional grids. In Section 6 we conclude with further research questions.

2. Recursions

In this section we derive recursions for the number of bimonotone subdivisions and total subdivisions for grids consisting of two rows. Let \( P_{m,n} \) denote a grid with 2 rows that has \( m \) points in the top row and \( n \) points in the bottom, aligned at the left, and let the bottom left point be at the origin. Let \( B_{m,n} \) be the number of bimonotone subdivisions of this configuration. We set up a recursion to count \( B_{m,n} \).

**Lemma 3.** The number of bimonotone subdivisions \( B_{m,n} \) of \( P_{m,n} \) is \( 2B_{m,n-1} + 2B_{m-1,n} - 2B_{m-1,n-1} \) if \( m > n \), \( 2B_{m,n-1} \) if \( m = n \), and 0 if \( m < n \).

**Proof.** For \( m > n \), if the top right vertex \((m-1, 1)\) is not connected to any vertex other than its left neighbor \((m-2, 1)\) and the bottom right vertex \((n-1, 0)\), as shown in Figure 4, then there are \( 2B_{m-1,n} \) bimonotone subdivisions. This is because if we pair each bimonotone subdivision with the subdivision that has the edge connecting \((n-1, 0)\) and \((m-2, 1)\) toggled, then each pair corresponds to the unique bimonotone subdivision of \( P_{m-1,n} \) with the same internal edges.
By the same reasoning, there are $2B_{m,n-1}$ bimonotone subdivisions when the bottom right vertex $(n-1,0)$ is not connected to any points but $(n-2,0)$ and $(m-1,1)$, as shown in Figure 5.

It is not possible for both vertices to be connected to points other than their external neighbors, as the edges would intersect at a point not in the configuration. And when both are not connected, as in Figure 6, there are $2B_{m-1,n-1}$ bimonotone subdivisions because similarly to above, the bimonotone subdivisions correspond to bimonotone subdivisions of $P_{m-1,n-1}$. So subtracting the overlap, $B_{m,n} = 2B_{m,n-1} + 2B_{m-1,n} - 2B_{m-1,n-1}$.

When $m < n$, there are 0 bimonotone subdivisions because the edge connecting $(m-1,1)$ and $(n-1,0)$ has a negative slope.

For $m = n$, as in Figure 7, it is impossible for the bottom right vertex to connect to any point but its neighbors as the slope would be negative. So just as in the first case of $m > n$, $B_{m,n} = 2B_{m,n-1}$.

In summary,

$$B_{m,n} = \begin{cases} 
2B_{m,n-1} + 2B_{m-1,n} - 2B_{m-1,n-1}, & m > n, \\
2B_{m,n-1}, & m = n, \\
0, & m < n.
\end{cases}$$
Lemma 4. The number of subdivisions of $P_{m,n}$ is $A_{m,n} = 2A_{m,n-1} + 2A_{m-1,n} - 2A_{m-1,n-1}$.

Proof. We again consider the connectivity of the top right and bottom right vertices and use the inclusion-exclusion principle. Using the same reasoning as for the $m > n$ case for $B_{m,n}$, we get the same recursion.

Expressions for $A_{m,n}$ can now be found similarly, as shown in the right column of Table 1.

### 3. General form

Next, we use the recursions from Section 2 to find the general forms of the numbers of bimonotone and total subdivisions of $P_{m,n}$. We begin with bimonotone subdivisions.

Theorem 5. The number of bimonotone subdivisions of $P_{m,n}$ is given by $B_{m,n} = \frac{2^{m-2}}{(n-2)!} (2P_{n-1}(m) - P_{n-1}(m - 1)) + 2^{m-n}B_{m-1,n}$.

Proof. We use induction. First, $B_{m,1} = 2^{m-2}$, because for each top vertex except $(0, 1)$ and $(m - 1, 1)$, the vertex may connect to the bottom vertex or may not. Now we consider $B_{m,n}$ in terms of $P_{n-1}(m)$. From the recursion,

$$B_{m,n} = \frac{2^{m-2}}{(n-2)!} \left( 2P_{n-1}(m) - P_{n-1}(m - 1) + 2^{m-n}B_{m-1,n} \right)$$

Plugging in $B_{m-1,n}$ and so on gives

$$B_{m,n} = \frac{2^{m-2}}{(n-2)!} \left( 2P_{n-1}(m) + \sum_{i=n+1}^{m-1} (-P_{n-1}(i) + 2P_{n-1}(i) - P_{n-1}(n)) \right) + 2^{m-n}B_{m-1,n}$$

$$= \frac{2^{m-2}}{(n-2)!} \left( 2P_{n-1}(m) + \sum_{i=n}^{m-1} -P_{n-1}(i) + 2P_{n-1}(i) \right)$$

$$= \frac{2^{m-2}}{(n-2)!} \left( P_{n-1}(m) + \sum_{i=n}^{m} P_{n-1}(i) \right).$$
Let $S(m, p)$ be the sum of the $p$-th powers of the first $m$ positive integers. Faulhaber’s formula gives this as

$$S(m, p) = \sum_{k=1}^{m} k^p = \frac{m^{p+1}}{p+1} + \frac{1}{2} m^p + \sum_{k=2}^{p} \frac{B_k}{k!} p^{k-1} m^{p-k+1},$$

where $p^{k-1} = p!/[p - (k - 1)!]$ and the $B_k$ are the Bernoulli numbers [23].

Let $P_{n-1}(m) = m^{n-2} + \sum_{i=0}^{n-3} a_i m^i$. Then from Faulhaber’s formula,

$$B_{m,n} = \frac{2^{m-2}}{(n-2)!} \left( P_{n-1}(m) + (S(m, n-2) - S(n, n-2)) + \sum_{i=0}^{n-3} a_i (S(m, i) - S(n, i)) \right).$$

Only $S(m, n-2)$ contains a $m^{n-1}$ term, which is $m^{n+1}/(n+1)$ from Faulhaber’s formula. None of the other terms contain a higher degree term. Thus the $m^{n-1}$ term of $B_{m,n}$ is $(2^{m-2}/(n-1)) m^{n-1}$. Since all other terms are polynomial, this is a polynomial.

Thus $B_{m,n}$ can be expressed in the form $B_{m,n} = (2^{m-2}/(n-1)!) P_n(m)$, where $P_n(m)$ is monic and of degree $n - 1$. \qed

We prove a similar result for the total number of subdivisions.

**Theorem 6.** The number of subdivisions of $P_m$ is given by $A_{m,n} = (2^{m-2}/(n-1)!) Q_n(m)$, where $Q_n(m)$ is some monic polynomial of degree $n - 1$.

**Proof.** We again use induction. Exactly as for bimonotone subdivisions, we find $A_{m,1} = 2^{m-2}$, satisfying the base. Then, considering $A_{m,n}$ in terms of $Q_{n-1}(m)$,

$$A_{m,n} = \frac{2^{m-2}}{(n-2)!} \left( 2 Q_{n-1}(m) - Q_{n-1}(m-1) \right) + 2 A_{m-1,n}.$$

Plugging in $A_{m-1,n}$ and so on gives

$$A_{m,n} = \frac{2^{m-2}}{(n-2)!} \left( 2 Q_{n-1}(m) + \sum_{i=1}^{m-1} - Q_{n-1}(i) + 2 Q_{n-1}(i) \right)$$

$$= \frac{2^{m-2}}{(n-2)!} \left( Q_{n-1}(m) + \sum_{i=1}^{m-1} Q_{n-1}(i) \right).$$

This proceeds in the same way as the proof of Theorem 5 without the $-S(n, n)$ term, which does not affect the result. Thus $A_{m,n}$ can be expressed in the form $A_{m,n} = (2^{m-2}/(n-1)!) Q_n(m)$, where $Q_n(m)$ is monic and of degree $n - 1$. \qed

Note that the two forms are (asymptotically) identical.

**Corollary 7.** The number of bimonotone subdivisions of $P_m$ is asymptotically equivalent to the total number of subdivisions for large values of $m$.

This Corollary suggests that estimating a two-dimensional log-concave and MTP$_2$ distribution requires the same sample size as estimating a two-dimensional log-concave distribution. This was also found in the
current work [16]. It remains an open problem to show that in higher dimensions estimating log-concave and MTP$_2$ densities requires much fewer samples than estimating log-concave densities.

Now, we connect bimonotone subdivisions to the large Schröder numbers [24]. The $n$-th large Schröder number $S_n$ is the number of paths from $(0, 0)$ to $(n, n)$ where unit steps can be taken north, east, or northeast and no points on the path lie above the line $y = x$.

**Theorem 8.** The number of bimonotone subdivisions of the $2 \times n$ lattice grid $P_{m,n}$ is equal to $2^{n-2}$ multiplied by the $(n-1)$-th large Schröder number.

**Proof.** We strongly induct on $n$. The large Schröder numbers are known to follow the relation

$$S_n = S_{n-1} + \sum_{k=0}^{n-1} S_k S_{n-1-k}.$$ 

First, $B_1$, the number of bimonotone subdivisions of a $2 \times 1$ grid, has little meaning, so we define it to be $\frac{1}{2}$ to fit the relationship with the Schröder numbers.

For a $2 \times n$ grid, the bimonotone subdivisions can be divided into the cases where there are no internal vertical edges, or where the leftmost vertical edge occurs at $x = k + 1$.

The first case, shown in Figure 8, corresponds to the bimonotone subdivisions of a $2 \times (n-1)$ grid. We can see this by considering the $2 \times (n-1)$ grid’s bimonotone subdivisions after shifting the top vertices of the grid one unit to the right. The shifted subdivision cannot contain vertical lines, as that would make the original not bimonotone. These shifted bimonotone subdivisions are also all those of the original $2 \times n$ grid, except the leftmost and rightmost edges of the $2 \times (n-1)$ grid can be present or not in the $2 \times n$ grid. This makes the number of bimonotone subdivisions in this case $4B_{n-1}$.

If the leftmost internal vertical edge occurs at $x = k + 1$, as shown in Figure 9, then we can separately consider the $2 \times (k+1)$ grid to the left and $2 \times (n-k)$ to the right. By the same reasoning as in the first case, the number of bimonotone subdivisions of the left side, which has no vertical edges, is $4B_k$. Note that for $k = 1$, there are 2 bimonotone subdivisions of the left side, which agrees with our definition of $B_1$ as $\frac{1}{2}$. The number of bimonotone subdivisions of the right side is simply $B_{n-k}$. Thus the total is $4B_k B_{n-k}$. 

![Figure 8. No vertical edges.](image)

![Figure 9. Leftmost vertical edge at $x = k + 1$.](image)
Figure 10. Every bimonotone triangulation that uses all grid points can be flipped to this triangulation.

Adding these, placing $2B_{n-1}$ into the summation as $4B_1B_{n-1}$, and plugging in $B_k = 2^{k-2}S_{k-1}$ for $k < n$, we get

$$B_n = 2B_{n-1} + 4 \sum_{k=1}^{n-1} B_k B_{n-k}$$

$$= 2^{n-2} \left( S_{n-2} + \sum_{k=0}^{n-2} S_k S_{n-2-k} \right)$$

$$= 2^{n-2}S_{n-1}. \quad \Box$$

4. Bimonotone triangulations

In this section, we consider the more specific problem of counting bimonotone triangulations. Because triangulations can be counted using the flip method [5], we apply the method to the bimonotone case.

**Definition 9.** A *flip* takes a quadrilateral in a triangulation and switches which of its two diagonals is included in the triangulation.

**Theorem 10.** Every bimonotone triangulation using all vertices of an equally spaced $m \times n$ lattice grid can be flipped to every other bimonotone triangulation of the grid.

**Proof.** We will show that every bimonotone triangulation can be flipped to the triangulation with every vertical edge and positive-slope diagonal $(i, j)\to(i+1, j+1)$ present (see Figure 10).

Take the longest diagonal of a particular triangulation. Let this range from $(0, 0)$ to $(x_d, y_d)$. We know that $x_d$ and $y_d$ are relatively prime, as the diagonal would intersect a lattice point otherwise.

We will first prove that the two vertices completing the two triangles using the diagonal must be the two points closest to the edge on either side. This will allow us to prove that the quadrilateral that has this diagonal can be flipped and remain bimonotone.

Consider the closest points below the line for each value of $x$. If $s$ is the slope of the line, then these points’ vertical distances from the diagonal are the fractional parts $\{s\}, \{2s\}, \ldots, \{(x_d - 1)s\}$. As each of these has numerator $k y_d \mod x_d$ and $y_d$ is relatively prime to $x_d$, this is an ordering of $1/x_d, 2/x_d, \ldots, (x_d - 1)/x_d$. 
Figure 11. No points can lie within the two shaded triangles.

Figure 12. The closest point must lie in the two gray triangles, not the red rectangle.

Let the closest point be at \((x_c, y_c)\). The vertical distance from this point to the diagonal is \(1/x_d\). We must prove that no lattice points lie strictly within the two triangles bounded by the extensions of the two lines connecting the closest vertex and one of the diagonal’s endpoints (see the gray triangles in Figure 11). Otherwise, a triangulation could use a point in this region as the third vertex instead of the closest point. The greatest vertical distance within the region is \(1/x_c\) for the right triangle and \((1/x_d) \cdot y_d/(y_d - y_c)\) for the left triangle, by similar triangles.

For each integer \(2 \leq k \leq x_d - 1\), we ensure the point with distance \(k/x_d\) does not lie within the two triangles. For now, only consider whether the point is in the triangle on the right. Either \(x_c < x_d/k\) or \(x_c \geq x_d/k\). If \(x_c < x_d/k\), then by similar triangles the point with distance \(ky_d/x_d\) lies on the line connecting the left vertex of the diagonal and the closest point, to the right of the closest point. Thus it is not strictly within the triangle. If \(x_d/k \leq x_c\), then \(k/x_d \geq 1/x_c\), the maximum vertical distance in the triangle, so the point cannot lie within the triangle. By symmetry, the same is true for the left triangle. Therefore, only the closest point on one side of the diagonal can be the third vertex for a triangle using it.

By symmetry, the closest points on either side of the diagonal are in mirrored positions, so that the quadrilateral formed is a parallelogram. We will now prove that when this quadrilateral is flipped, its new diagonal is shorter and remains bimonotone.
Consider the vertex \((x_c, y_c)\) of the triangle below the diagonal. For the flipped diagonal to be bimonotone, the vertex must be either above or to the left of the midpoint of the diagonal (nonstrictly), i.e., the gray regions in Figure 12. This condition is always met. If \(s\) is the slope of the diagonal, the closest point to the diagonal in the region strictly below and to the right of the midpoint, i.e., the red region in Figure 12, is either \(\frac{1}{2} + \frac{s}{2}\) or \(1 + \frac{s}{2}\) or \(\frac{1}{2} + s\) vertically away from the diagonal. There are guaranteed to be points closer than \(\frac{1}{2}\) in the desired (gray) regions: By symmetry, if the closest point in the \(i\)-th column of points is \(t\) vertically below the diagonal, then the closest point in the \((x_d-i+1)\)-th row is \(1-t\) below. So there exist points less than or equal to \(\frac{1}{2}\) vertically below the diagonal, and, therefore, the closest point to the diagonal has to be in the desired (gray) regions. Therefore, we can flip the diagonal to replace it with one that is still bimonotone but shorter.

This can be repeated for all diagonals until the configuration with only horizontal and vertical edges and unit diagonals is reached. Therefore, all bimonotone triangulations of an \(m \times n\) grid are connected by flips.

\[\square\]

5. Algorithms

We now present an algorithm to count all subdivisions and all bimonotone subdivisions of an \(m \times n\) grid. The algorithm considers all possible internal edges between the vertices of the grid. To count bimonotone subdivisions, only edges of nonnegative or vertical slope are considered. All possible combinations of edges are then tested for being a subdivision. For each pair of edges, we check if they intersect within the grid outside a vertex. If none do, then we check convexity: For each vertex in the interior of the grid, the angles between consecutive edges protruding out of the vertex must be less than or equal to \(\pi\). Successful subdivisions are added to the count until all are counted. See Algorithm 1.

The output of the algorithm for small values of \(m, n\) is shown in Table 2.

**Algorithm 1:** Count all subdivisions and all bimonotone subdivisions of an \(m \times n\) grid.

<table>
<thead>
<tr>
<th>Input</th>
<th>The sizes (m) and (n) of an (m \times n) grid.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output</td>
<td>The numbers of all subdivisions and all bimonotone subdivisions of the grid.</td>
</tr>
</tbody>
</table>

1. for All subsets of (bimonotone) edges between the vertices of the \(m \times n\) grid do
2. if Any two edges intersect at a point not on the grid then
3. | Continue. |
4. end
5. end
6. for Each vertex from the grid with an edge adjacent to it do
7. if The angles between consecutive edges protruding out of the vertex are at most \(\pi\) then
8. | Add 1 to the count. |
9. end
10. end

| Output | The numbers of bimonotone and all subdivisions. |
Table 2. Left: the number of Subdivisions of $2 \times n$ grids. Right: the number of Subdivisions of $3 \times n$ grids.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$B_{2,n}$</th>
<th>$A_{2,n}$</th>
<th>$n$</th>
<th>$B_{3,n}$</th>
<th>$A_{3,n}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>12</td>
<td>26</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>26</td>
<td>3</td>
<td>528</td>
<td>2224</td>
</tr>
<tr>
<td>4</td>
<td>88</td>
<td>252</td>
<td>4</td>
<td>34152</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>720</td>
<td>2568</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>6304</td>
<td>26928</td>
<td>6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We now present a second algorithm to count the number of bimonotone triangulations of an $m \times n$ grid for which every point in the grid is a vertex of a triangle. This is a breadth-first search algorithm using flips. Starting from the arrangement including every vertical, horizontal, and positively sloped unit diagonal edge, quadrilaterals can be flipped so that the diagonal present is switched. This is attempted for every possible flip, and resulting arrangements are checked if they are different from previous ones. If new, they are added to the list. See Algorithm 2.

6. Future Research

It would be interesting to prove this conjecture on the total number of subdivisions of a $2 \times n$ grid:

**Conjecture 11.** The number of subdivisions of a $2 \times n$ grid is equal to $2^{n-2}$ multiplied by the $(n-1)$-th central Delannoy number.

The central Delannoy numbers [22] are related to the large Schröder numbers, which are similarly involved in the expression for the number of bimonotone subdivisions of a $2 \times n$ grid given in Theorem 8.
The \( n \)-th central Delannoy number counts the number of paths from \((0, 0)\) to \((n, n)\) where steps can be taken east, north, or northeast. The difference between these and the large Schröder numbers is that these paths can cross the line \(y = x\).

We have confirmed that this conjecture holds for small values of \(n\), up to \(n = 6\).

This conjecture could not be proved in the same way as Theorem 8 because the central Delannoy numbers lack the kind of recurrence relation that the Schröder numbers have. There is a recurrence relation using the noncentral Delannoy numbers for rectangular grids, but the noncentral numbers do not appear to have any simple relation with the number of subdivisions of \(P_{m,n}\).

Acknowledgements

The research presented here took place mostly in 2018 via the MIT PRIMES USA program. We are grateful to Dr. Tanya Khovanova, Dr. Slava Gerovitch, and Prof. Pavel Etingof for organizing the program. At the time Robeva was supported by an NSF postdoctoral fellowship (DMS-170-3821) and is currently supported by a National Sciences and Engineering Research Council of Canada Discovery Grant (DGECR-2020-00338).

References


Received 2020-09-13. Revised 2021-08-10. Accepted 2021-08-28.

ELINA ROBEVA: erobeva@math.ubc.ca

Department of Mathematics, University of British Columbia, Vancouver, BC, Canada

MELINDA SUN: mmsun@mit.edu

Massachusetts Institute of Technology, Cambridge, MA, United States
A GENERALIZATION FOR THE EXPECTED VALUE OF THE EARTH MOVER’S DISTANCE

WILLIAM Q. ERICKSON

The earth mover’s distance (EMD), also called the first Wasserstein distance, can be naturally extended to compare arbitrarily many probability distributions, rather than only two, on the set \([n] = \{1, \ldots, n\}\). We present the details for this generalization, along with a highly efficient algorithm inspired by combinatorics; it turns out that in the special case of three distributions, the EMD is half the sum of the pairwise EMDs. Extending the methods of Bourn and Willenbring (2020), we compute the expected value of this generalized EMD on random tuples of distributions, using a generating function which coincides with the Hilbert series of the Segre embedding. We then use the EMD to analyze a real-world data set of grade distributions.

1. Introduction

We generalize a result appearing in [5], in which the authors compute the expected value of the earth mover’s distance (EMD) between two probability distributions by means of a generating function. The EMD can be viewed as the solution to a problem in transport theory, first considered in [20] by French geometer Gaspard Monge in 1781. (Although the term “earth mover” seems to have been coined only in the 1990s, it is pointed out in Villani’s monumental reference [23] that the title of Monge’s original treatise translates, more or less, as “On the theory of material extracted from the earth and input to a new construction.” Monge, then, truly was the original earth mover.) Nearly 200 years later, in [15], Monge’s name was given to a critical property of certain cost arrays for which his problem can be solved by a greedy algorithm. Throughout the late 1980s and 1990s, in [2] and [4], this Monge property was generalized to higher-dimensional arrays. (See also [18] for a more recent treatment.) It is an essential fact in this paper (whose proof is reserved for Section 8 at the end) that the multidimensional cost array associated with our generalized EMD has this Monge property.

Section 2, written for those readers unfamiliar with the EMD, presents a simple example and points out all the relevant details which will reappear in our generalization.

We begin in Section 3 by defining an earth mover’s “distance” \(\operatorname{EMD}_d\) between \(d\) distributions; the classical EMD treated in [5] coincides with \(\operatorname{EMD}_2\). We actually find that on three distributions, \(\operatorname{EMD}_3\) equals half the sum of the three \(\operatorname{EMD}_2\) values, although no such relationship holds for \(d > 3\).

In Section 4, en route to constructing a generating function, we define a discrete version of \(\operatorname{EMD}_d\) which compares histograms instead of probability distributions, and we describe an efficient computational method using a generalization of the RSK correspondence from combinatorics.

**MSC2020:** primary 05E14, 13P25; secondary 05E40.

**Keywords:** earth mover’s distance, generating functions, Wallach representations.
In Section 5, we encode the values of the discrete EMD\(_d\) in a generating function, which we manipulate in order to extract the expected value. Translating this discrete result back into the continuous setting, we prove the main theorem of this paper (Theorem 7), which is a recursive formula to compute the expected value of EMD\(_d\). We then apply our theory in Section 6 to analyze a real-world data set of grade distributions.

Finally, in Section 7, we interpret our generating function in terms of the Segre embedding in algebraic geometry. This leads us to realize the EMD\(_2\) situation as an infinite-dimensional representation of the Lie algebra \(u(n, n)\), whose action corresponds to manipulating the two distributions compared by our EMD.

Since the appearance of [5], the problem of finding the expected value of EMD\(_2\) has been solved from an analytical approach in [11]. The setup has also been specialized in [19] to a data set of distributions with a fixed average value.

We believe that the result in this paper—a method to evaluate the “closeness” of arbitrarily many distributions—has great potential as a tool in data analysis. In evaluating teaching and assessment practices at the university level, for instance, we can now assign a single value to an entire course by evaluating the EMD between the individual sections, and then track the behavior of that course’s EMD for different groups of instructors, different course coordinators, fall vs. spring semesters, and other variables. We can even assign EMD values to individual exams and other assessments using the grade distributions in various sections; or in the other direction, we can compare different courses to each other, both within and outside a given department. In all of these settings, we believe that the generalized EMD can contribute to an interesting cluster analysis of the kind proposed in [5].

2. EMD between 2 distributions: summary and an example

For readers unfamiliar with the classical EMD, we summarize the idea here. Consider two probability distributions on the finite set of integers \([n] = \{1, \ldots, n\}\). (More vividly, in place of a “probability distribution,” imagine \(n\) bins of earth whose combined mass is one unit, located at 1, \ldots, \(n\) on the number line.) Intuitively, the EMD between the two distributions measures the “cheapest” way to move earth between the bins so as to equalize the distributions, where the “cost” of moving one unit of earth is the distance of the move. For example, the cost of moving 0.25 units of earth from bin 2 to bin 5 is \(0.25 \cdot (5 - 2) = 0.75\). To make this precise, we define the cost function \(C : [n] \times [n] \to \mathbb{Z}_{\geq 0}\), where \(C(i, j)\) is the cost of moving one unit of earth from bin \(i\) to bin \(j\). In this case, clearly \(C(i, j) = |i - j|\).

Any solution which equalizes the two distributions—whether or not it is the optimal solution—can be encoded in an \(n \times n\) matrix \(J\). Necessarily, the row sums of \(J\) will correspond to one distribution, and the column sums to the other distribution, so the entries of \(J\) must sum to 1. To find the total cost of earth moved, we simply multiply each matrix entry \(J_{ij}\) by its distance \(|i - j|\) from the main diagonal, and sum over all entries.

As a trivial example, in the case of two equal distributions, we can take \(J\) to be the matrix whose diagonal entries are the values of the distributions, with zeros elsewhere. Since all nonzero entries of \(J\) have coordinates of the form \((i, i)\), they contribute zero cost, and so the total cost is likewise zero, just as we would expect for two identical distributions.
To interpret this process of multiplying entries by their distance from the diagonal, our next example
describes exactly how the entries of $J$ give (possibly ambiguous, but equivalent) step-by-step instructions
to equalize the two distributions. This treatment is admittedly overkill in the case of two distributions,
but it will provide the best intuition when we generalize to $d$ distributions in the next section. The
less-than-rigorous descriptions below will be formalized in the next section in terms of the taxicab metric.

**Example.** Consider the two distributions $\mu_1 = (0.3, 0.3, 0.4)$ and $\mu_2 = (0.1, 0, 0.9)$. Hence $n = 3$. Then one matrix (among infinitely many) with the prescribed row and column sums is

$$J = \begin{bmatrix} .1 & 0 & .2 \\ 0 & 0 & .3 \\ 0 & 0 & .4 \end{bmatrix}.$$ 

The nonzero entries of $J$ correspond to moving earth as follows:

- $J_{1,1} = 0.1$. Note that the coordinates $(1, 1)$ are already equal to each other, so we do not have to move the 0.1 units of earth at all.

- $J_{1,3} = 0.2$. Now the coordinates $(1, 3)$ are not equal; in order to make them equal with as little cost as possible, we have three valid options, all of which have cost 2:
  - In the first coordinate, we could add 2 to make the change $1 \rightarrow 3$. This corresponds to moving the 0.2 units of earth in $\mu_1$, from bin 1 to bin 3.
  
  - In the second coordinate, we could subtract 2 to make the change $3 \rightarrow 1$. This corresponds to moving the 0.2 units of earth in $\mu_2$, from bin 3 to bin 1.
  
  - We could add 1 to the first coordinate $(1 \rightarrow 2)$ and subtract 1 from the second coordinate $(3 \rightarrow 2)$. This corresponds to moving 0.2 units of earth in $\mu_1$ from bin 1 to bin 2, and then moving 0.2 units of earth in $\mu_2$ from bin 3 to bin 2.

- $J_{2,3} = 0.3$. The cheapest ways to equalize the coordinates $(2, 3)$ are the following two options, each with cost 1:
  - In the first coordinate, we could add 1 to make the change $2 \rightarrow 3$. This corresponds to moving the 0.3 units of earth in $\mu_1$, from bin 2 to bin 3.
  
  - In the second coordinate, we could subtract 1 to make the change $3 \rightarrow 2$. This corresponds to moving the 0.3 units of earth in $\mu_2$, from bin 3 to bin 2.

- $J_{3,3} = 0.4$. Since the coordinates $(3, 3)$ are already equal, we do not have to move the 0.4 units of earth at all.

Now, depending upon which of the above options we choose at each step, this process can result in
any of six distinct pairs of final distributions $\mu_1'$ and $\mu_2'$. But within each possible pair, as the reader can
check, we always finish with $\mu_1' = \mu_2'$, as desired. Furthermore, the total cost of all the earth moved is
independent of our choices, since all options above minimized the cost at each step. (Also note that the
cost at each step was always equal to \(|i - j|\), coinciding with the cost function \(C\) we defined earlier.) In this case, the total cost of the earth moved was

\[
0.1(0) + 0.2(2) + 0.3(1) + 0.4(0) = 0.7.
\]

The EMD between \(\mu_1\) and \(\mu_2\) is, by definition, the infimum (actually the minimum) of the set of total costs, taken over all possible matrices \(J\) with the prescribed row and column sums. In this example, although not obvious at first glance, 0.7 is in fact the least possible cost, and so \(\text{EMD}(\mu_1, \mu_2) = 0.7\). This turns out to be a consequence of the fact that the support of \(J\) lies in a chain: in other words, if we put the product order \(\leq\) on \([n] \times [n]\), we see that

\[(1, 1) \leq (1, 3) \leq (2, 3) \leq (3, 3);\]

this pairwise comparability is what we mean by a chain in \([n] \times [n]\). This fact — that support in a chain implies minimality — is equivalent, on a deeper level (see [15]), to the fact that our cost function \(C\), if considered as an \(n \times n\) array, has the “Monge property” alluded to in the introduction; in this case, the greedy algorithm to solve the earth mover’s problem (known as the “northwest corner rule”; see [4]) eliminates one row or column at each step, meaning the support of the solution matrix \(J\) is always a chain.

There is one phenomenon here in the \(d = 2\) case which will not generalize to \(d > 2\): in the above example, we could have removed any ambiguity by deciding that we would move earth within \(\mu_1\) exclusively, so that both final distributions would equal \(\mu_2\). Therefore, we could interpret the problem as finding the cheapest way to transport material from a “source” or “supply vector” (\(\mu_1\)) to a “sink” or “demand vector” (\(\mu_2\)). For \(d > 2\), however, the optimal solution at each step may require moving earth in any or all of the distributions, and so we lose the binary supply-demand interpretation of the problem.

Having presented the big picture, without details, in the \(d = 2\) case, we now proceed to build up the general case for arbitrary \(d\). Throughout the next section, the reader can verify that the definitions and results coincide with those found in this simple example where \(d = 2\).

3. Extending EMD to \(d\) distributions

3.1. Definitions and notation. Let \(\mathcal{P}_n\) denote the set of probability distributions on \([n]\). Assume the uniform probability measure on the \(d\)-fold product \(\mathcal{P}_n \times \cdots \times \mathcal{P}_n\), defined by its embedding into \(\mathbb{R}^{d\cdot n}\). Our goal is to compare \(d\) elements of \(\mathcal{P}_n\), written as the \(d\)-tuple \(\mu := (\mu_1, \ldots, \mu_d)\). We should keep in mind that each \(\mu_i\) is itself an \(n\)-tuple whose components sum to 1. Throughout this paper, we write the sum of a vector’s components using absolute value bars, so in this case, \(|\mu_i| = 1\). We will denote the \(k\)-th component of \(\mu_i\) by \(\mu_i(k)\), which is just the value of the distribution \(\mu_i\) at \(k \in [n]\). To each \(\mu\) there corresponds the set \(\mathcal{J}_\mu\) of joint distribution arrays, defined as follows.

For an array \(J\), we will write \(J(m_1, \ldots, m_d)\) for the entry at position \((m_1, \ldots, m_d)\). Now, we define \(\mathcal{J}_\mu\) as the set containing all those arrays \(J \in \mathbb{R}_{\geq 0}^{n \times \cdots \times n}\) whose sums within the coordinate hyperplanes coincide with \(\mu\). Specifically, fixing \(m_i = k\), we must have

\[
\sum_{m_1, \ldots, m_i, \ldots, m_d = 1}^n J(m_1, \ldots, k, \ldots, m_d) = \mu_i(k).
\]
A GENERALIZATION FOR THE EXPECTED VALUE OF THE EARTH MOVER’S DISTANCE

Figure 1. An illustration of the conditions in (1), in the case where \( d = 3 \) and \( n = 4 \).

Given some \( \mu = (\mu_1, \mu_2, \mu_3) \), every array in \( \mathcal{J}_\mu \) satisfies the above relations, where each arrow represents the sum of the entries in the designated plane.

In other words, summing all the entries whose positions in the array have \( k \) as their \( i \)-th coordinate, we obtain the \( k \)-th component of \( \mu_i \). In the familiar \( d = 2 \) case, \( i = 1 \) gives us the row sums, and \( i = 2 \) the column sums. For \( d = 3 \), see Figure 1 for an illustration.

Any array \( J \in \mathcal{J}_\mu \) can be regarded as a solution to the earth mover’s problem for \( n \) bins, determined by the distributions in \( \mu \). This means we need a \( d \)-dimensional analog of the “cost” function from Section 2, and the natural candidate arises from the taxicab metric on \([n]^d\). Specifically, at each position in a \( d \)-dimensional array, we want the cost to be the taxicab distance to the main diagonal, i.e., to the nearest position in the array whose coordinates are all equal. (This “equality of coordinates” property of the main diagonal, as we recall from Section 2, corresponded to zero earth being moved.) Roughly speaking, this cost is the fewest number of \( \pm 1 \) we need to add in order to equalize all the coordinates. For example, the most efficient way to equalize the coordinates of the position \((5, 4, 5, 5, 5, 7, 5)\) is to add 1 to the 4, and then to subtract 2 from the 7, for a total cost of 3. This is precisely the taxicab distance to the main diagonal, specifically to the position \((5, 5, 5, 5, 5, 5)\). Just as in the example from the previous section, this distance-finding exercise corresponds to moving earth:

• When we added 1 to the 2nd coordinate to make the change \( 4 \rightarrow 5 \), we moved a unit of earth in the 2nd distribution \( \mu_2 \) from bin 4 to bin 5.

• When we subtracted 2 from the 6th coordinate to make the change \( 7 \rightarrow 5 \), we moved a unit of earth in the 6th distribution \( \mu_6 \) from bin 7 to bin 5.

Example. Consider the three distributions

\[
\mu_1 = (0.5, 0.1, 0.4), \quad \mu_2 = (0.5, 0.2, 0.3), \quad \mu_3 = (0.7, 0.2, 0.1).
\]

Then one array in \( \mathcal{J}_\mu \) is, for instance,

\[
J = \begin{bmatrix}
[.5, 0, 0] & [0, 0, 0] & [0, 0, 0] \\
[0, 0, 0] & [.1, 0, 0] & [0, 0, 0] \\
[0, 0, 0] & [.1, 0, 0] & [.0, .2, .1]
\end{bmatrix},
\] (2)
flattened so that the first coordinate specifies the row, the second coordinate specifies one of the three main columns, and the third coordinate specifies the position inside the triple at that position. The nonzero entries are

\[ J(1, 1, 1) = 0.5, \quad J(2, 2, 1) = 0.1, \quad J(3, 2, 1) = 0.1, \quad J(3, 3, 2) = 0.2, \quad J(3, 3, 3) = 0.1. \]

This information tells us how to arrive at the solution corresponding to \( J \):

- The cost of \((1, 1, 1)\) is 0 since it is already on the main diagonal, so we do not move the 0.5 at all.
- The cost of \((2, 2, 1)\) is 1, since in the 3rd coordinate we must make the change \(1 \rightarrow 2\). This means that in the 3rd distribution \( \mu_3 \), we move 0.1 from bin 1 to bin 2. Currently \( \mu_3' = (0.6, \ 0.3, \ 0.1) \).
- The cost of \((3, 2, 1)\) is 2, since we equalize the coordinates most efficiently by subtracting 1 from the 1st coordinate \((3 \rightarrow 2)\) and adding 1 to the 3rd coordinate \((1 \rightarrow 2)\). Hence, we move 0.1 from bin 3 to bin 2 in \( \mu_1 \), and from bin 1 to bin 2 in \( \mu_3 \). Now \( \mu_1' = (0.5, \ 0.2, \ 0.3) \) and \( \mu_3' = (0.5, \ 0.4, \ 0.1) \).
- The cost of \((3, 3, 2)\) is 1, by adding 1 to the 3rd coordinate. This corresponds to moving 0.2 from bin 2 to bin 3 in \( \mu_3 \). Now \( \mu_3' = (0.5, \ 0.2, \ 0.3) \).
- The cost of \((3, 3, 3)\) is 0, so we do not move the 0.1 anywhere.

Note that our final result is that all three distributions are the same, as desired:

\[ \mu_1' = \mu_2' = \mu_3' = (0.5, \ 0.2, \ 0.3). \]

Also note that we rigged this example, unlike that in Section 2, so that none of the steps would present more than one optimal option, although in general there certainly might exist several different solutions for the same array \( J \). But of course in each case the total cost is the same.

The natural computation now is to find that total cost, by multiplying the amount of earth moved at each step by the number of bins it was moved; in other words, multiply each entry in \( J \) by the cost of its position, then add these products together:

\[ 0.5(0) + 0.1(1) + 0.1(2) + 0.2(1) + 0.1(0) = 0.5. \]

This completes the example.

Of course, \textit{a priori} there is no reason why the particular array \( J \) in the example should give the least costly way to equalize the three distributions. When we finally define our generalized EMD, it will be defined as the least possible cost for any \( J \in \mathcal{J}_M \). First, however, we should record a formula for the cost of an array position, to improve upon the somewhat sloppy method by inspection we have used so far.

The formula for the \( d \)-dimensional taxicab distance from a point to a line is derived in [7]. In our case, the line of interest is the main diagonal, which passes through \((1, \ldots, 1)\) in the direction \((1, \ldots, 1)\). This distance, and therefore our cost function \( C \), turns out to be

\[ C(m_1, \ldots, m_d) = \min_{i \in [d]} \left\{ \sum_{j \neq i} |m_i - m_j| \right\}. \quad (3) \]
This cost function $C$ can also naturally be thought of as an $n \times \cdots \times n$ array, so we will occasionally refer to the “cost array” in this paper.

There is also a more direct way to compute $C$, which will be convenient later. Let $\mathbf{m} := (m_1, \ldots, m_d)$, and let $\hat{\mathbf{m}}$ denote the vector whose components are those of $\mathbf{m}$ rearranged in ascending order; e.g., if $\mathbf{m} = (7, 4, 5, 3, 1)$, then $\hat{\mathbf{m}} = (1, 3, 4, 5, 7)$.

**Proposition 1.** Equation (3) can be computed as $C(\mathbf{m}) = \sum_{i=1}^{\lfloor d/2 \rfloor} \hat{m}_{d-i+1} - \hat{m}_i$.

As an example before the proof, take $\mathbf{m} = (7, 4, 5, 3, 1)$ as above. Then by the proposition, to compute $C(\mathbf{m})$, we instead look at $\hat{\mathbf{m}}$ and sum up the pairwise differences working outside-in:

$$(7-1=6)$$

$$(5-3=2)$$

$\hat{\mathbf{m}} = (1, 3, 4, 5, 7),$$

therefore $C(\mathbf{m}) = 6 + 2 = 8$.

**Proof.** For fixed $i \in [d]$, let $k$ be such that $\hat{m}_i = m_i$. Then we have

$$\sum_{j \neq i} |m_i - m_j| = (\hat{m}_2 - \hat{m}_1) + 2(\hat{m}_3 - \hat{m}_2) + 3(\hat{m}_4 - \hat{m}_3) + \cdots + (k - 1)(\hat{m}_k - \hat{m}_{k-1})$$

$$+ (\hat{m}_{d} - \hat{m}_{d-1}) + 2(\hat{m}_{d-1} - \hat{m}_{d-2}) + 3(\hat{m}_{d-2} - \hat{m}_{d-3}) + \cdots + (d - k)(\hat{m}_{k+1} - \hat{m}_k),$$

which is minimized when $k = \lfloor (d + 1)/2 \rfloor$. Making this evaluation in the displayed sum, we find that the sum telescopes: when $d$ is even, we obtain

$$-\hat{m}_1 - \hat{m}_2 - \cdots - \hat{m}_{\lfloor (d+1)/2 \rfloor} + \hat{m}_{\lfloor (d+1)/2 \rfloor + 1} + \cdots + \hat{m}_d,$$

and when $d$ is odd, we obtain

$$-\hat{m}_1 - \hat{m}_2 - \cdots - \hat{m}_{\lfloor (d+1)/2 \rfloor - 1} + \hat{m}_{\lfloor (d+1)/2 \rfloor + 1} + \cdots + \hat{m}_d.$$

In either case, this simplifies as

$$C(\mathbf{m}) = \sum_{i=1}^{\lfloor d/2 \rfloor} \hat{m}_{d-i+1} - \hat{m}_i.$$ 

\[\square\]

**Remark.** A recent paper [18] defines a different cost function than ours for the earth mover’s problem, namely $C'(\mathbf{m}) := \max\{m_i\} - \min\{m_i\}$. We can see from Proposition 1 that $C'$ agrees with our $C$ for $d = 2$ and $d = 3$, but not for $d > 3$. For example, letting $\mathbf{m} = (1, 1, 2, 2)$, we have $C(\mathbf{m}) = 2$ but $C'(\mathbf{m}) = 1$. For our purposes, we have chosen our $C$ because it counts every earth-movement required to equalize the distributions. For example, keeping $\mathbf{m} = (1, 1, 2, 2)$, consider the distributions $\mu_1 = \mu_2 = (1, 0)$ and $\mu_3 = \mu_4 = (0, 1)$. Then one solution is given by the array whose only nonzero entry is a 1 at position $\mathbf{m}$. Intuitively, we want the EMD of these four distributions to be 2, not 1, since we must first move a unit of earth by 1 bin, and then move another unit by 1 bin.

Having built up the necessary intuition and formulas, we are finally ready to make our main definition:
Definition. Let $\mu$ be a $d$-tuple of probability distributions, as above. Then the generalized earth mover’s distance is defined as

$$\text{EMD}_d(\mu) := \min_{J \in J(\mu)} \sum_{m \in [n]^d} C(m) J(m). \quad (4)$$

Remark. For $d > 2$, the term “coefficient” would perhaps be more fitting than “distance,” but in this paper we retain “distance” as a nod to tradition.

3.2. Existence of a greedy algorithm. As mentioned in the first two sections, finding the right-hand side of (4) is equivalent to finding the optimal solution to a $d$-dimensional transport problem. It is shown in [4] that there exists a greedy algorithm to find this solution in $O(d^2 n)$ time, precisely when the cost array $C$ has the Monge property mentioned in the introduction:

Definition. A $d$-dimensional array $A$ has the Monge property if, for all $x = (x_1, \ldots, x_d)$ and $y = (y_1, \ldots, y_d)$, we have

$$A(\min\{x_1, y_1\}, \ldots, \min\{x_d, y_d\}) + A(\max\{x_1, y_1\}, \ldots, \max\{x_d, y_d\}) \leq A(x) + A(y).$$

Remark. If we regard an array as a function on a lattice, then the Monge property is equivalent to submodularity. Characterized by their “diminishing returns” property, submodular functions have found a vast expanse of modern applications, especially in machine learning (see [3]).

We now state the crucial proposition, whose proof we will give in Section 8.

Proposition 2. The cost array $C$ defined in (3) has the Monge property.

This proposition, then, guarantees the existence of a greedy algorithm to compute $\text{EMD}_d$. (This justifies our writing “min” instead of “inf” in our definition; we also could have used a compactness argument as in [5].) The greedy algorithm described in [4] is a generalization of the two-dimensional “northwest corner rule.” Just as in the $d = 2$ case (see Section 2), for generic $d$ this algorithm arrives at its solution in the form of an array $J \in J(\mu)$ whose support is a chain, i.e., pairwise comparable under the product order on $[n]^d$. (In [5, Proposition 4], the “straightening” procedure that converts the support of any $J$ into a chain, without increasing the total cost, is valid precisely because the cost array $C(i, j) = |i - j|$ has the Monge property.) Rather than describe this greedy algorithm, which is already well-known (see [4] or [18]), our goal is instead to find the expected value of $\text{EMD}_d$. To this end, the importance of the algorithm is the following:

Corollary 3. The minimum in (4) occurs for some $J \in J(\mu)$ whose support is a chain in $[n]^d$.

Since there is nothing special about the condition $|\mu_i| = 1$ from the perspective of transport problems, Corollary 3 also holds in a discrete setting using integer compositions in place of probability distributions. We will take this discrete approach in the next section, where we use a highly efficient combinatorial method to find the optimal array $J$ for any $\mu$. 
4. A discrete approach

We follow the method from [5], with a view toward constructing a generating function in the next section. In place of $P_n$, we temporarily turn our attention to $C(s, n)$, the set of (weak) integer compositions of some positive integer $s$ into $n$ parts. That is, elements of $C(s, n)$ are $n$-tuples of nonnegative integers whose sum is $s$ (whereas before, the elements of $P_n$ were $n$-tuples of nonnegative real numbers whose sum was 1); we can also think of compositions as histograms. The cost function $C$, however, remains the same as before, since it still describes distances among the $n$ bins in each of the $d$ compositions.

In this section, $\mu = (\mu_1, \ldots, \mu_d)$ denotes a sequence of compositions $\mu_i \in C(s, n)$. It is tempting simply to adjust the definition of $J_\mu$ so that arrays in the set must have nonnegative integer entries summing to $s$; then we could just reuse the definition (4) to obtain a definition for the discrete EMD$_d$. Although this is one viable approach, nevertheless, in light of Corollary 3, we need only consider those arrays whose support is a chain; therefore we will work with the following set of arrays from this point forward:

$$J_{(n^d)}^s := \left\{ J \in (\mathbb{Z}_{\geq 0})^{n \times \cdots \times n} \mid \sum_{m \in [n]^d} J(m) = s, \text{ and the support of } J \text{ is a chain in } [n]^d \right\}.$$

We will now show that each $d$-tuple of compositions $\mu \in C(s, n) \times \cdots \times C(s, n)$ corresponds to a unique array $J_\mu \in J_{(n^d)}^s$. Therefore, by the end of the next subsection, we will have a direct computational definition for the discrete version of the EMD, which avoids taking the minimum over a set of arrays as we must in the definition (4) of the continuous EMD. Once we have this definition for the discrete EMD, we will be able to recover the continuous version by scaling all the $\mu_i$ by $1/s$. Near the end of the paper, we will do exactly this, and then let $s \to \infty$, in order to translate discrete results back into the continuous setting.

4.1. Generalized RSK correspondence. The authors of [5] use the Robinson–Schensted–Knuth correspondence to great effect in order to determine a unique optimal matrix $J_{(\mu_1, \mu_2)}$ for an ordered pair of compositions $(\mu_1, \mu_2)$. We now apply this same idea to $d$ compositions in order to uniquely determine an optimal $d$-dimensional array. This will allow us to calculate the discrete EMD$_d$ directly (and even more efficiently, in many cases, than by using the greedy algorithm mentioned above).

For nonexperts, we summarize here a special case of the correspondence. (For full details, see Chapter 4 of [12].) The Robinson–Schensted–Knuth (RSK) correspondence furnishes a bijection:

$$\left\{ \begin{array}{l}
\text{ordered pairs of semistandard} \\
\text{Young tableaux of the same} \\
\text{shape, with entries in } [n]
\end{array} \right\} \leftrightarrow \left\{ \begin{array}{l}
n \times n \text{ matrices with} \\
\text{nonnegative integer entries}
\end{array} \right\}.$$

For our purposes, we will restrict our attention to the special case of one-row tableaux, since any composition in $C(s, n)$ corresponds uniquely to a one-row tableau containing $s$ boxes with entries from $[n]$. As an example, consider the two compositions,

$$\mu_1 = (1, 2, 3, 4), \quad \mu_2 = (5, 0, 2, 3),$$
in $C(10, 4)$. We associate to each composition $\mu_i$ a one-row tableau $T(\mu_i)$, which we fill so that the entry $k$ appears $\mu_i(k)$ times:

$$T(\mu_1) = \begin{bmatrix} 1 & 2 & 2 & 3 & 3 & 4 & 4 & 4 & 4 \end{bmatrix}, \quad T(\mu_2) = \begin{bmatrix} 1 & 1 & 1 & 1 & 3 & 3 & 4 & 4 & 4 \end{bmatrix}.$$  

Regarding these tableaux as the two rows of a $2 \times s$ matrix, we then have

$$M(\mu_1, \mu_2) = \begin{bmatrix} 1 & 2 & 2 & 3 & 3 & 4 & 4 & 4 & 4 \\
1 & 1 & 1 & 1 & 3 & 3 & 4 & 4 & 4 \end{bmatrix}.$$  

Finally, we fill in an $n \times n$ array $J(\mu_1, \mu_2)$ whose $(i, j)$ entry equals the number of times the column $\begin{bmatrix} i \\ j \end{bmatrix}$ appears in $M(\mu_1, \mu_2)$. For example, $\begin{bmatrix} 4 \\ 4 \end{bmatrix}$ appears three times, so we write a 3 in position $(4, 4)$. Filling in the rest of the array, we obtain the correspondence

$$(\mu_1, \mu_2) \leftrightarrow (T(\mu_1), T(\mu_2)) \leftrightarrow M(\mu_1, \mu_2) \leftrightarrow J(\mu_1, \mu_2) = \begin{bmatrix} 1 & 0 & 0 & 0 \\
2 & 0 & 0 & 0 \\
2 & 0 & 1 & 0 \\
0 & 0 & 1 & 3 \end{bmatrix}.$$  

Note that we can also reverse the procedure, starting with the array $J(\mu_1, \mu_2)$, translating its entries into a two-row matrix, and finally recovering the original pair of tableaux (and hence the pair of compositions). Therefore this is indeed a bijection. In the context of the EMD, the array $J(\mu_1, \mu_2)$ has two significant properties:

- The row and column sums coincide with the original compositions $\mu_1$ and $\mu_2$, so $J(\mu_1, \mu_2)$ is a solution to the discrete earth mover’s problem for $\mu_1$ and $\mu_2$.
- Since both rows of $M(\mu_1, \mu_2)$ are nondecreasing, the support of $J(\mu_1, \mu_2)$ is a chain in $[n] \times [n]$.

In summary, we have the following bijective correspondence in the case $d = 2$:

$$C(s, n) \times C(s, n) \leftrightarrow J_s(n^2).$$

This RSK correspondence extends naturally to $d$-tuples of compositions in $C(s, n)$. (For experts, details about the existence of this multivariate RSK generalization can be found in [6].) Given $\mu = (\mu_1, \ldots, \mu_d)$, the tableaux corresponding to the $\mu_i$ uniquely determine a $d \times s$ matrix $M_\mu$, which in turn determines a unique $n \times \cdots \times n$ array $J_\mu \in J_{s(n^d)}$. This correspondence is again bijective, establishing the following special case of the generalized RSK correspondence:

$$C(s, n) \times \cdots \times C(s, n) \leftrightarrow J_{s(n^d)}^s, \quad \mu \mapsto J_\mu.$$  

This correspondence leads us to the following definition of the discrete EMD:

**Definition.** For positive integers $d$, $n$, and $s$, with $d \geq 2$, let $\mu = (\mu_1, \ldots, \mu_d)$ with each $\mu_i \in C(s, n)$. Let $J_\mu$ be the unique array corresponding to $\mu$, as in (5). Let $C$ be the cost function on $[n]^d$ as in (3). Then we define the *discrete generalized earth mover’s distance* to be

$$\text{EMD}_d^s(\mu) := \sum_{m \in [n]^d} C(m)J_\mu(m),$$  

as in (6).
where we write the superscript $s$ to distinguish this discrete version from the continuous version.

This definition in (6) is largely conceptual; in practice, we can calculate $\text{EMD}_d^s(\mu)$ directly from the matrix $M_\mu$, since the support of $J_\mu$ is determined by the columns of $M_\mu$:

**Theorem 4.** Let $\mu = (\mu_1, \ldots, \mu_d)$ with each $\mu_i \in C(s, n)$, and let $C$ be the cost function in (3). Let $M_\mu$ be the unique $d \times s$ array corresponding to $\mu$ via the generalized RSK correspondence, as described above, and let $M_\mu(\bullet, j)$ denote the $j$-th column vector in $M_\mu$. Then

$$\text{EMD}_d^s(\mu) = \sum_{j=1}^s C(M_\mu(\bullet, j)).$$

**Proof.** Consider the definition of $\text{EMD}_d^s$ in (6). By definition, $J_\mu(m)$ equals the number of occurrences of $m$ as a column vector of the matrix $M_\mu$. Therefore $J_\mu(m) = 0$ unless $m$ is one of those column vectors, and so we can simply sum over the $s$ column vectors to obtain the result. \hfill \square

**Remark.** This construction via RSK is more efficient than the aforementioned greedy algorithm for computing $\text{EMD}_d^s$ when $d$ or $n$ is sufficiently large: rather than filling a $d$-dimensional array in $O(d^2 n)$ time, we need to consider only the $s$ column vectors of a $d \times s$ matrix.

**Example.** As an example for $d = 3$, consider the three compositions

$$\mu_1 = (4, 0, 1), \quad \mu_2 = (1, 2, 2), \quad \mu_3 = (0, 5, 0)$$

in $C(5, 3)$. These correspond to the tableaux $\begin{bmatrix} 1 & 1 & 1 & 1 & 3 \end{bmatrix}$, $\begin{bmatrix} 1 & 2 & 2 & 3 & 3 \end{bmatrix}$, and $\begin{bmatrix} 2 & 2 & 2 & 2 & 2 \end{bmatrix}$, respectively. Stacking these tableaux vertically gives us the matrix

$$M_\mu = \begin{bmatrix} 1 & 1 & 1 & 3 \\ 1 & 2 & 2 & 3 & 3 \\ 2 & 2 & 2 & 2 & 2 \end{bmatrix}.$$ 

Now using Theorem 4 on the five columns of $M_\mu$, we compute that

$$\text{EMD}_3^s(\mu) = C(1, 1, 2) + C(1, 2, 2) + C(1, 2, 2) + C(1, 3, 2) + C(3, 3, 2)$$

$$= 1 + 1 + 1 + 2 + 1$$

$$= 6.$$

Before advancing to the main problem of the paper, we show that $\text{EMD}_3$ can actually be expressed in terms of the classical $\text{EMD}_2$. (In the following proposition, we suppress the superscript $s$ because the result holds for both the discrete and the continuous version of $\text{EMD}$: the equality is independent of $s$, and therefore still holds after dividing both sides by $s$ and letting $s \rightarrow \infty$.)

**Proposition 5.** The value of $\text{EMD}_3$ is half the sum of the three pairwise $\text{EMD}_2$ values:

$$\text{EMD}_3(\mu_1, \mu_2, \mu_3) = \frac{1}{2} \left( \text{EMD}_2(\mu_1, \mu_2) + \text{EMD}_2(\mu_1, \mu_3) + \text{EMD}_2(\mu_2, \mu_3) \right).$$
Proof. Let \( \mu = (\mu_1, \mu_2, \mu_3) \) as usual. In each column \( j \) of the matrix \( M_\mu \), call the three entries \( a_j, b_j, c_j \), labeled so that \( a_j \leq b_j \leq c_j \). Each of the three pairs \((a_j, b_j), (a_j, c_j), \) and \((b_j, c_j)\) corresponds naturally to one of the pairs \((\mu_1, \mu_2), (\mu_1, \mu_3), \) and \((\mu_2, \mu_3)\), where row \( i \) corresponds to \( \mu_i \). Therefore by Theorem 4, we have

\[
\text{EMD}_2(\mu_1, \mu_2) + \text{EMD}_2(\mu_1, \mu_3) + \text{EMD}_2(\mu_2, \mu_3) = \sum_{j=1}^{s} C(a_j, b_j) + C(a_j, c_j) + C(b_j, c_j)
\]

\[
= \sum_{j} (b_j - a_j) + (c_j - a_j) + (c_j - b_j)
\]

\[
= \sum_{j} 2c_j - 2a_j
\]

\[
= 2 \sum_{j} c_j - a_j
\]

\[
= 2 \sum_{j} C(a_j, b_j, c_j)
\]

\[
= 2 \cdot \text{EMD}_3(\mu_1, \mu_2, \mu_3).
\]

This relationship does not generalize to \( d > 3 \), because in general, the telescoping summand in the proof does not reduce in terms of a higher-dimensional cost function. For example, when \( d = 4 \), the analog of the third line above is \( \sum_{j} 3d_j + c_j - b_j - 3a_j \), or \( \sum_{j} C(a_j, b_j, c_j, d_j) + 2(d_j - a_j) \).

5. Expected value of \( \text{EMD}_d \)

Again we will follow and extend the methods used in [5] for arbitrary values of \( d \). First we define a generating function, in which the exponents encode the values of \( \text{EMD}_d \), and which we will then differentiate in order to sum up all of these values. This will allow us to compute expected value for \( \text{EMD}_d \) by reading off the coefficients from this derivative.

A generating function is a formal power series whose terms encode combinatorial information, without any regard to convergence. As a famous example, the Fibonacci numbers \( F_i \) may be encoded in a formal power series \( \sum_{i=0}^{\infty} F_i x^i = x + x^2 + 2x^3 + 3x^4 + 5x^5 + 8x^6 + 13x^7 + \cdots \), whose generating function has the closed form \( x/(1 - x - x^2) \). The generating function we will define, however, will contain two variables (one to keep track of composition size \( s \), and one to keep track of \( \text{EMD}_d \) values), and unlike the example above, we will have to settle for a recursive definition, because writing down the closed form quickly becomes intractable even for \( n > 3 \).

5.1. Generating function for the discrete setting. Because we are about to make a recursive definition, we must momentarily consider \( d \)-tuples \( \mu \) consisting of compositions with different numbers of bins — i.e., different values \( n_i \) such that each \( \mu_i \in C(s, n_i) \). Therefore, \( n = (n_1, \ldots, n_d) \) will denote this vector of bin numbers. Note that \( \text{EMD}_d \) is still defined on \( C(s, n_1) \times \cdots \times C(s, n_d) \), since we can append zeros to the end of each \( n_i \)-tuple as needed; in this way, as in [5], we regard the compositions as sharing a common bin number (the maximum of the \( n_i \)). We do not, however, consider any statistical meaning
A recursive definition of this generating function, for the

The coefficient of

is the vector whose \(i\)-th component is 1 if \(i \in A\) and 0 otherwise. For example, if \(d = 5\), and \(A = \{2, 4, 5\}\), then \(e(A) = (0, 1, 0, 1, 1)\).

For fixed \(n = (n_1, \ldots, n_d)\), we define a generating function in two indeterminates \(z\) and \(t\):

\[
H_n(z, t) := \sum_{s=0}^{\infty} \left( \sum_{\mu \in \mathcal{C}(s, n_1) \times \cdots \times \mathcal{C}(s, n_d)} z^{\text{EMD}_d^s(\mu)} \right) t^s.
\]  

(7)

The coefficient of \(z^r t^s\) is the number of elements \(\mu \in \mathcal{C}(s, n_1) \times \cdots \times \mathcal{C}(s, n_d)\) such that \(\text{EMD}_d^s(\mu) = r\). A recursive definition of this generating function, for the \(d = 2\) case, is derived in [5, Theorem 3]. Our generalization for \(d > 2\) follows:

**Proposition 6.** Fix \(n = (n_1, \ldots, n_d)\). The generating function \(H_n := H_n(z, t)\) has the following recursive definition, where the sum is over all nonempty subsets \(A \subseteq [d]\):

\[
H_n = \frac{\sum_A (-1)^{|A|-1} \cdot H_{n-e(A)}}{1 - z |C(n)| t},
\]

where \(H_{(t^s)} = 1/(1 - t)\), and \(H_{n-e(A)} = 0\) if \(n - e(A)\) contains \(0\).

We reserve the proof for Section 9.

**Example.** We write out this recursive definition in a concrete case, where \(d = 3\) and \(n = (5, 2, 2)\). It is easiest to arrange the terms of the numerator according to the size of the subset \(A\). First, for \(|A| = 1\), we add together all possible \(H_{n'}\), where \(n'\) equals \(n\) with exactly 1 coordinate decreased; then for \(|A| = 2\), we subtract all possible \(H_{n''}\) where \(n''\) equals \(n\) with exactly 2 coordinates decreased; finally, for \(|A| = 3\), we add the one possible \(H_{n'''}\) where \(n'''\) equals \(n\) with all 3 coordinates decreased. As for the denominator, \(C\) is the same cost function we defined in (3), meaning that \(C(5, 2, 2) = 5 - 2 = 3\). Then the recursion for \(H_n\) looks like this:

\[
H_{(5,2,2)} = \frac{H_{(4,2,2)} + H_{(5,1,2)} + H_{(5,2,1)} - H_{(4,1,2)} - H_{(4,2,1)} - H_{(5,1,1)} + H_{(4,1,1)}}{1 - z^2 t}.
\]

Having seen an example, we now study the important (and very well-studied) specialization that results from setting \(z = 1\). In this case, the coefficient of \(t^s\) in \(H_n(1, t)\) is the total number of \(d\)-tuples \(\mu\), which is \(\prod_{i=1}^{d} |C(s, n_i)|\):

\[
H_n(1, t) = \sum_{s=0}^{\infty} \prod_{i=1}^{d} \binom{s+n_i-1}{n_i-1} t^s.
\]  

(8)
It is shown in [9] that the closed form of (8), after adjusting the index to match our setup, and writing 
\(|\mathbf{n}| := n_1 + \cdots + n_d\), is
\[
H_n(1, t) = \frac{W_n(t)}{(1 - t)^{|\mathbf{n}| - d + 1}},
\]
where the numerator \(W_n(t)\) is a polynomial whose coefficients are the “Simon Newcomb” numbers. (For 
more on this natural generalization of Eulerian numbers to multisets, see [1; 9; 21].) Specifically, denoting
the coefficient of \(t^i\) in \(W_n\) by the symbol \([t^i]W_n\), we have
\[
[t^i]W_n = \# \text{ permutations of the multiset } \{1^{n_1-1}, \ldots, d^{n_d-1}\} \text{ containing } i \text{ descents.}
\]
It follows that the evaluation \(W_n(1)\) is equal to the total number of permutations of the multiset 
\(\{1^{n_1-1}, \ldots, d^{n_d-1}\}\), a fact we will need later:
\[
W_n(1) = \frac{(\sum_{i=1}^{d} (n_i - 1))!}{\prod_{i=1}^{d} (n_i - 1)!} = \frac{(|\mathbf{n}| - d)!}{\prod (n_i - 1)!}.
\]

5.2. A partial derivative. Next, in order to transfer the EMD values from the exponents of \(z\) into coefficients, we take the partial derivative of \(H_n\) with respect to \(z\). Applying the quotient rule to our definition of
\(H_n\) in Proposition 6, we obtain the following, where the sum still ranges over nonempty subsets \(A \subseteq [d]::
\[
\frac{\partial H_n}{\partial z} = \frac{(1 - z^{C(n)}t) \left( \sum_{A} (-1)^{|A|-1} \cdot \frac{\partial H_n-e(A)}{\partial z} \right) + C(n) \cdot z^{C(n)-1} \cdot t \cdot \left( \sum_{A} (-1)^{|A|-1} \cdot H_n-e(A) \right)}{(1 - z^{C(n)}t)^2}
\]
Now that the exponents have been changed into coefficients of \(z\), we can set \(z = 1\):
\[
H_n' := \left. \frac{\partial H_n}{\partial z} \right|_{z=1} = \sum_{s=0}^{\infty} \left( \sum_{\mu \in \mathcal{C}(s, n_1) \times \cdots \times \mathcal{C}(s, n_d)} \text{EMD}_d^s(\mu) \right) t^s
\]
\[
= \frac{(1 - t) \left( \sum_{A} (-1)^{|A|-1} \cdot H_n'-e(A) \right) + t \cdot C(n) \left( \sum_{A} (-1)^{|A|-1} \cdot H_n-e(A) \right)}{(1 - t)^2}
\]
At this point, \(z\) has played out its role, and so from now on we will write \(H_n\) in place of \(H_n(1, t)\).

Note that the coefficient of \(t^s\) in \(H_n'\) is the sum of the values \(\text{EMD}_d^s(\mu)\) for all valid \(d\)-tuples \(\mu\). This
means that our goal is now in sight: to find the expected value of \(\text{EMD}_d^s\) for fixed \(\mathbf{n}\), we need to divide
the sum of all possible \(\text{EMD}_d^s\) values (i.e., the coefficient of \(t^s\) in \(H_n'\) by the total number of possible inputs \(\mu\) (i.e., the coefficient of \(t^s\) in \(H_n\)). Therefore, once we find a way to simplify (11), we will be able 
to compute the result
\[
\mathbb{E}(\text{EMD}_d^s) = \frac{[t^s]H_n'}{[t^s]H_n} = \frac{[t^s]H_n'}{\prod_{i=1}^{d} \binom{s+n_i-1}{n_i-1}},
\]
where \([t^s]\) again denotes the coefficient of \(t^s\) in a series.

In order to make the expression (11) for \(H_n'\) more tractable to program, we will now focus only on the numerators of \(H_n\) and \(H_n'\). We have already defined the numerator of \(H_n\) as \(W_n(t)\) in the previous
subsection. We will let $N_n(t)$ denote the numerator of $H'_n$. By using software and observing patterns for small $n$, we anticipate that the denominator of $H'_n$ has exponent $|n| - d + 2$, and so we now set both

$$ W_n := (1 - t)^{|n| - d + 1} H_n \quad \text{and} \quad N_n(t) := (1 - t)^{|n| - d + 2} H'_n. \quad (13) $$

Therefore, we can clear denominators in (11) by multiplying both sides by $(1 - t)^{|n| - d + 2}$. Proceeding carefully and clearing the remaining denominators using (13), the pattern becomes clear:

$$ N_n = \sum_A (-1)^{|A|-1} (1 - t)^{|A| - 1} N_{n-e(A)} + t \cdot C(n) \cdot (1 - t)^{|n| - d} \cdot (1 - t) \cdot H_n $$

$$ = \sum_A (t - 1)^{|A|-1} N_{n-e(A)} + t \cdot C(n) \cdot W_n. \quad (14) $$

This provides us with a quick recursive code to obtain $N_n$, after which we need only divide by $(1 - t)^{|n| - d + 2}$ to recover $H'_n$. The rest is just a matter of extracting coefficients in order to apply the result in (12).

5.3. **Expected value for continuous version of EMD$_d$.** Now that we have a way to determine the expected value for the discrete EMD, we aim to find a formula for the expected value in the continuous setting.

Starting with the expected value from (12), we scale by $1/s$ to normalize, and then let $s$ grow asymptotically:

$$ E_n := \mathbb{E}(\text{EMD}_d) = \lim_{s \to \infty} \frac{1}{s} \cdot \mathbb{E}(\text{EMD}'_d) $$

$$ = \lim_{s \to \infty} \frac{1}{s} \cdot [t^s] H'_n \left/ \prod_{i=1}^d \left( s + n_i - 1 \right) \right|_{n_i - 1} \right). $$

First we focus on the $[t^s] H'_n$ part, namely the coefficient of $t^s$ in $H'_n = N_n(t)/(1 - t)^{|n| - d + 2}$. Now, the coefficient of $t^s$ in the series $1/(1 - t)^{|n| - d + 2}$ is just

$$ \left( \frac{s + |n| - d + 1}{|n| - d + 1} \right) = \frac{s^{|n|-d+1}}{(|n| - d + 1)!} + \text{lower-order terms in } s. $$

Meanwhile, $N_n(t)$ is a polynomial, with some finite degree $b$. Now, as $s \to \infty$, we have $s - b \to \infty$, and so the coefficient of $t^s$ in $H'_n$ is asymptotic to $s^{|n|-d+1}/(|n| - d + 1)!$ multiplied by the sum of the coefficients of $N_n(t)$. But this sum is just $N_n(1)$, and so we have

$$ [t^s] H'_n \sim N_n(1) \cdot \frac{s^{|n|-d+1}}{(|n| - d + 1)!}. $$

Accounting for the $1/s$, we currently have the following:

$$ E_n = \lim_{s \to \infty} N_n(1) \cdot \frac{s^{|n|-d}}{(|n| - d + 1)! \prod_{i=1}^d (s + n_i - 1)} \cdot \frac{s^{|n|-d+1}}{(|n| - d + 1)!}. $$

Now, since

$$ \prod_{i} \left( \frac{s + n_i - 1}{n_i - 1} \right) \sim \prod_{i} \frac{s^{n_i - 1}}{(n_i - 1)!} = \frac{s^{|n|-d}}{\prod_{i} (n_i - 1)!}, $$
We confirm this in Mathematica; see Table 1.

But when we evaluate $N_n(1)$ from (14), the terms with $(t - 1)$ all disappear; hence we need only consider subsets $A \subseteq [d]$ with one element, meaning we are now summing from 1 to $d$:

$$N_n(1) = \sum_{i=1}^{d} N_{n-e(i)}(1) + C(n)W_n(1).$$

Substituting for $W_n(1)$ using (10), we have

$$N_n(1) = \sum_{i=1}^{d} N_{n-e(i)}(1) + \frac{C(n) \cdot (|n| - d)!}{\prod_{i=1}^{d} (n_i - 1)!}.$$

Finally, returning to (15) and plugging this all in for $N_n(1)$, we conclude with the recursive definition

$$E_n = \left[ \sum_{i=1}^{d} N_{n-e(i)}(1) + \frac{C(n) \cdot (|n| - d)!}{\prod_{i=1}^{d} (n_i - 1)!} \right] \cdot \prod_{i=1}^{d} (n_i - 1)! \cdot \frac{(|n| - d + 1)!}{(|n| - d)!}$$

$$= \sum_{i=1}^{d} (n_i - 1)E_{n-e(i)} + C(n),$$

where $E_{n-e(i)} = 0$ if $n - e(i)$ contains a 0.

We record this as the main theorem of this paper. We state the result only in the case of equal bin numbers, writing $(n^d) = (n, \ldots, n)$, so that the statistical meaning of $\text{EMD}_{d}$ is well-defined.

**Theorem 7.** The expected value of $\text{EMD}_{d}$ on $\mathcal{P}_n \times \cdots \times \mathcal{P}_n$ is $E_{(n^d)}$ as defined in (16).

**Remark.** Recall from Proposition 5 the special relationship between $\text{EMD}_3$ and $\text{EMD}_2$, namely, $\text{EMD}_3$ equals half the sum of the three pairwise $\text{EMD}_2$ values. This leads us to anticipate that

$$E_{(n^3)} = E(\text{EMD}_3) = \frac{1}{2}(\text{EMD}_2 + \text{EMD}_2 + \text{EMD}_2)$$

$$= \frac{1}{2} \cdot E(\text{EMD}_2)$$

$$= \frac{3}{2} \cdot E(\text{EMD}_2)$$

$$= \frac{3}{2} \cdot E_{(n^2)}.$$

We confirm this in Mathematica; see Table 1.

### 5.4. Unit normalized $\text{EMD}$

It is often convenient to unit normalize the value of $\text{EMD}_d$ so that its value falls between 0 and 1. To this end, we claim that for a given $n$, the maximum value of $\text{EMD}_d$ is $[d/2](n-1)$. To see this, observe that the maximum value of the discrete $\text{EMD}_d^s(\mu)$ occurs when in every column of the matrix $M_{\mu}$ (given by the RSK correspondence), half the entries are 1 and the other half are $n$; if $d$ is odd, then “half” means $[d/2]$, with the leftover entry being irrelevant by Proposition 1. For such a $\mu$, then, $\text{EMD}_d^s(\mu)$ equals the cost $[d/2](n-1)$ multiplied by $s$ (the number of columns). After
A GENERALIZATION FOR THE EXPECTED VALUE OF THE EARTH MOVER’S DISTANCE

\[ n \cdot \mathcal{E}(n^2) \leq \mathcal{E}(n^3) \leq \frac{\mathcal{E}(n^3)}{\mathcal{E}(n^2)} \]

<table>
<thead>
<tr>
<th>n</th>
<th>( \mathcal{E}(n^2) )</th>
<th>( \mathcal{E}(n^3) )</th>
<th>( \mathcal{E}(n^3)/\mathcal{E}(n^2) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.3333</td>
<td>0.5000</td>
<td>1.5</td>
</tr>
<tr>
<td>3</td>
<td>0.5333</td>
<td>0.8000</td>
<td>1.5</td>
</tr>
<tr>
<td>4</td>
<td>0.6857</td>
<td>1.0286</td>
<td>1.5</td>
</tr>
<tr>
<td>5</td>
<td>0.8127</td>
<td>1.2191</td>
<td>1.5</td>
</tr>
<tr>
<td>6</td>
<td>0.9235</td>
<td>1.3853</td>
<td>1.5</td>
</tr>
<tr>
<td>7</td>
<td>1.0230</td>
<td>1.5345</td>
<td>1.5</td>
</tr>
<tr>
<td>8</td>
<td>1.1139</td>
<td>1.6709</td>
<td>1.5</td>
</tr>
<tr>
<td>9</td>
<td>1.1982</td>
<td>1.7972</td>
<td>1.5</td>
</tr>
<tr>
<td>10</td>
<td>1.2770</td>
<td>1.9155</td>
<td>1.5</td>
</tr>
</tbody>
</table>

Table 1. Mathematica verification that \( \mathcal{E}(n^3) = \frac{3}{2} \cdot \mathcal{E}(n^2) \).

<table>
<thead>
<tr>
<th>n</th>
<th>( \hat{E}(n^d) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.3333</td>
</tr>
<tr>
<td>3</td>
<td>0.2667</td>
</tr>
<tr>
<td>4</td>
<td>0.2286</td>
</tr>
<tr>
<td>5</td>
<td>0.2032</td>
</tr>
<tr>
<td>6</td>
<td>0.1847</td>
</tr>
<tr>
<td>7</td>
<td>0.1705</td>
</tr>
<tr>
<td>8</td>
<td>0.1591</td>
</tr>
<tr>
<td>9</td>
<td>0.1498</td>
</tr>
<tr>
<td>10</td>
<td>0.1419</td>
</tr>
</tbody>
</table>

Table 2. The unit normalized expected value \( \hat{E}(n^d) \).

dividing by \( s \) to pass to the continuous setting, we see that the maximum value of \( \text{EMD}_d \) is \( \lfloor d/2 \rfloor (n-1) \), as claimed. Therefore we present definitions for the unit normalized \( \text{EMD}_d \) and its expected value:

\[
\hat{\text{EMD}}_d(\mu) := \frac{\text{EMD}_d(\mu)}{[d/2](n-1)} \quad \text{and} \quad \hat{\mathcal{E}}(n^d) := \frac{\mathcal{E}(n^d)}{[d/2](n-1)}.
\]  

(17)

Using Theorem 7 and the definition in (17), we record the unit normalized expected value \( \hat{\mathcal{E}}(n^d) \) in Table 2, for the first few values of \( n \) and \( d \). We observe a curious phenomenon in this table when we fix \( n \) and let \( d \) increase: the unit normalized expected value alternately increases (\( d \) changing from even to odd) and decreases (\( d \) changing from odd to even). This suggests that an even number of distributions are more likely to be “closer” together (in the sense of the EMD) than an odd number of distributions. We can explain this dependence on parity from both a geometric and statistical perspective.

Geometrically, in odd dimensions, the taxicab distance from a point to the main diagonal remains unchanged as we vary the median of the point’s coordinates, and the variability of the median increases as the distance increases (i.e., as the other coordinates become more spread out). For example, in a 3-dimensional array with side lengths \( n \), the greatest possible distance from the main diagonal is...
Figure 2. A plot of $\hat{E}_{(3^d)}$ (vertical axis), for values of $d$ up to 100 (horizontal axis). The disparity between even and odd values of $d$ is significant when $d$ is small, but becomes negligible for $d$ sufficiently large.

$\left\lfloor \frac{3}{2} \right\rfloor (n - 1) = n - 1$; this is attained by all those positions whose coordinates (up to reordering) are $(1, \_ , n)$, where the blank can take any value in $[n]$. In a 4-dimensional array, however, there are far fewer positions at maximal distance $\left\lfloor \frac{4}{2} \right\rfloor (n - 1) = 2(n - 1)$ from the main diagonal: namely, those positions whose coordinates (up to reordering) are $(1, 1, n, n)$. In effect, the free “odd-man-out” coordinate does not exist in an even number of dimensions, so there are fewer array positions farther away from the main diagonal compared to an odd number of dimensions. This remains true when we restrict to arrays with support in a chain, and so we expect to see the even–odd disparity reflected in the EMD$_d$ data.

Likewise, statistically, three distributions achieve maximum EMD$_3$ value when one of them is $(1, 0, \ldots, 0)$, another is $(0, \ldots, 0, 1)$, and the third is anything at all. But the maximum EMD$_4$ value is attained only when two distributions are $(1, 0, \ldots, 0)$ and the other two are $(0, \ldots, 0, 1)$. For $d$ odd, the variability in the free “odd-man-out” distribution is greater for higher EMD$_d$ values, and so it is no surprise that there is a greater proportion of higher EMD$_d$ values when $d$ is odd.

Although this even-odd disparity is significant when $d$ is relatively small (as in Table 2), the difference becomes negligible as $d$ increases, and the values of $\hat{E}_{(n^d)}$ do stabilize. We plot these values in Figure 2, fixing $n = 3$ and letting $d$ grow from 2 to 100 on the horizontal axis. On the far left side of the plot, the disparity between even and odd values of $d$ is indeed striking — and the variability is much greater when $d$ is even than when $d$ is odd — but we see that this difference quickly becomes negligible, with the sequence of expected values converging to approximately 0.39.

6. Real-world data

As a basic example, we apply our generalized discrete EMD to four mathematics courses at the University of Wisconsin–Milwaukee, during the fall 2019, spring 2020, and fall 2020 semesters. (Needless to say, these three semesters are of added interest because of the drastic changes brought on by the Covid-19 pandemic in spring 2020.) This data is contained in the “Section attrition and grade report,” published by
the Office of Assessment and Institutional Research at UWM; final letter grades are A, B, C, D, and F, so
\( n = 5 \). The four courses are MATH 105 (Introduction to College Algebra) along with 231, 232, and 233 (Calculus I–III), and we analyze the grade distributions from the individual sections of each course. In
Table 3, we record the number \( d \) of course sections in each semester, along with the (unit normalized) EMDs of those sections.

Before computing the EMD values using Theorem 4, we needed to account for the unequal enrollments among course sections; after all, EMDs is defined only on compositions with a common value of \( s \), and in this case \( s \) is the number of students in a section (typically between 20 and 30). Our solution is to scale every grade distribution so that \( s = 100 \), rounding as necessary to maintain integer compositions. (Of course the value 100 is arbitrary, and can be increased if we desire to lessen the effect of the rounding; typically, however, percentage points are sufficient for practical applications.)

The results in the table are also plotted in Figure 3, from which we can make one immediate observation: the EMD of each of the calculus courses increased from one semester to the next, whereas the EMD of the college algebra course decreased. In fact, in the semester before the pandemic, college algebra had a higher EMD than any of the calculus courses, but its EMD was lower than all of the calculus courses by the second semester of the pandemic. Of course, the EMD is only one statistic among many, and we would need to examine other measures (particularly the grade-point averages) and more previous semesters before drawing definitive conclusions. But we can say that the grade distributions of the various sections

<table>
<thead>
<tr>
<th>course</th>
<th>fall 2019</th>
<th>spring 2020</th>
<th>fall 2020</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>sections</td>
<td>EMD</td>
<td>sections</td>
</tr>
<tr>
<td>MATH 105</td>
<td>29</td>
<td>0.2059</td>
<td>13</td>
</tr>
<tr>
<td>MATH 231</td>
<td>10</td>
<td>0.1095</td>
<td>6</td>
</tr>
<tr>
<td>MATH 232</td>
<td>7</td>
<td>0.1308</td>
<td>7</td>
</tr>
<tr>
<td>MATH 233</td>
<td>7</td>
<td>0.1650</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 3. Discrete EMD applied to four mathematics courses at the University of Wisconsin–Milwaukee.

Figure 3. The unit normalized values of EMDs for four mathematics courses, tracked during three consecutive semesters.
calculus sections (within each course) had grown farther apart since the shift to online learning in spring 2020, while the distributions of the various sections of college algebra had become closer together. This raises further questions:

- To what extent has the shift to online learning caused the recent trends in EMD?
- To what extent are these trends a reflection of instructors’ and/or course coordinators’ and/or students’ response to the online format?
- Are students’ grades trending higher or lower (or neither) in each course since the pandemic began?
- How does EMD typically behave in spring vs. fall semesters before and after the pandemic began?

Finally, we compare these real-world EMD values to the expected value defined in Theorem 7. From Table 2, for \( n = 5 \), we observe that the approximate range of \( \widehat{E}(5^d) \) is from 0.25 to 0.3 for the section numbers \( d \) in our data set. The actual EMD values lie significantly below this range (with the sole exception of Calculus II in fall 2020), especially considering that the expected value would be slightly higher in the discrete setting than it is in the continuous setting. We should not be too surprised by this, of course, since college grades are (hopefully) not assigned at random, and therefore not all grade distributions are equally likely.

7. Connection to algebraic geometry and representation theory

In this section, as a sort of appendix, we shift away from statistics and observe the connection between the EMD and the Segre embedding from algebraic geometry. As a result, we show how the entire EMD^2 setting can be realized as a representation of the Lie algebra \( u(n, n) \).

7.1. A determinantal variety. In the \( d = 2 \) case, as indicated in [5], the series \( H_{(n,n)} := H_{(n,n)}(1, t) \), from (8), is in fact the Hilbert series of the determinantal variety

\[ D_{n}^{\leq 1} := \{ M \in M_{n}(\mathbb{C}) \mid \text{rank } M \leq 1 \} \]

consisting of \( n \times n \) complex matrices with rank at most 1. To see this, it suffices to show that for a given nonnegative integer \( s \), the number of elements in \( \mathcal{C}(s, n) \times \mathcal{C}(s, n) \) equals the dimension of \( \mathbb{C}[D_{n}^{\leq 1}]^s \), the space of homogeneous degree-\( s \) polynomial functions on \( D_{n}^{\leq 1} \). To this end, let \( w_{ij} \) be the coordinate functions on a generic \( n \times n \) matrix. Since all \( 2 \times 2 \) minors vanish on any matrix in \( D_{n}^{\leq 1} \), we observe that \( \mathbb{C}[D_{n}^{\leq 1}] \simeq \mathbb{C}[w_{11}, \ldots, w_{nn}]/\mathcal{I} \), where \( \mathcal{I} \) is the determinantal ideal generated by the quadratics of the form \( w_{ij}w_{i'j'} - w_{i'j}w_{ij} \) for \( i < i' \) and \( j < j' \). It follows that a basis for \( \mathbb{C}[D_{n}^{\leq 1}]^s \) is given by the set of monomials

\[ B^s = \left\{ \prod_{k=1}^{s} w_{i_k,j_k} \mid \{ (i_k, j_k) \} \text{ is a chain in } [n] \times [n] \right\}. \]

Now we have an obvious bijective correspondence between the sets \( B^s \) and \( J_{s}^s_{(n,n)} \):

\[ \prod_{i,j \in [n]} w_{ij}^{f_{ij}} \leftrightarrow J \in J_{s}^s_{(n,n)}. \]
But $J_{(n,n)}^{(t)}$ is in bijective correspondence with $C(s,n) \times C(s,n)$, as is clear from our RSK correspondence in (5). This proves our claim that $H_{(n,n)}$ is the Hilbert series of $D_n^{\leq 1}$.

7.2. The Segre embedding. We extend the previous result to $d > 2$, writing $(n^d)$ for $(n, \ldots, n)$ as before. The specialization $H_{(n^d)} := H_{(n^d)}(1, t)$ of our generating function from (8) is the Hilbert series of the image of the Segre embedding:

$$\mathbb{P}(\mathbb{C}^n) \times \cdots \times \mathbb{P}(\mathbb{C}^n) \hookrightarrow \mathbb{P}(\mathbb{C}^n \otimes \cdots \otimes \mathbb{C}^n), \quad ([v^{(1)}], \ldots, [v^{(d)}]) \mapsto [v^{(1)} \otimes \cdots \otimes v^{(d)}]$$

(see [14; 21]). That is, $H_{(n^d)}$ is the Hilbert series of the coordinate ring of the simple tensors. (In the case $d = 2$, the set of simple tensors in $\mathbb{C}^n \otimes \mathbb{C}^n$ is identified with $D_n^{\leq 1}$, coinciding with the previous subsection.)

To sketch this generalization of the $d = 2$ case, which we presented in detail above, we let $m$ range over all multi-indices $(m_1, \ldots, m_d) \in [n] \times \cdots \times [n]$. Now consider a simple tensor $v^{(1)} \otimes \cdots \otimes v^{(d)}$. We can expand this tensor in the standard basis as

$$\sum_m \left( \frac{v^{(1)}_{m_1} \cdots v^{(d)}_{m_d}}{w_m} \right) e_{m_1} \otimes \cdots \otimes e_{m_d},$$

where $v^{(k)}_{\ell}$ is the $\ell$-th coordinate of the vector $v^{(k)}$, the $w_m$ are coordinate functions. For any two multi-indices $m$ and $m'$, we see that the quadratic $w_m w_{m'}$ is invariant under the exchange of indices componentwise between $m$ and $m'$. Intuitively, then, we can again mod out by the determinantal ideal generated by all $2 \times 2$ minors, just as we did in the $d = 2$ case above.

The upshot is that a basis for the coordinate ring of the simple tensors is given by those monomials $w_{m_1} \cdots w_{m_d}$, such that the set $\{m_i\}$ of multi-indices forms a chain. We therefore have the generalization of (18), and we conclude that $H_{(n^d)}$ is the Hilbert series of the simple tensors in $\mathbb{C}^n \times \cdots \times \mathbb{C}^n$.

7.3. Representation theory. Returning to the $d = 2$ case, the coordinate ring $\mathbb{C}[D_n^{\leq 1}]$ is an infinite-dimensional vector space known as the first Wallach representation of the Lie algebra $u(n,n)$ of the indefinite unitary group (see [10]). We will show how the action on $\mathbb{C}[D_n^{\leq 1}]$ corresponds to manipulating the two compositions in our EMD$_2$ setting.

Consider the polynomial ring $\mathbb{C}[x,y] := \mathbb{C}[x_1, \ldots, x_n, y_1, \ldots, y_n]$. On one hand, $\mathbb{C}[x,y]$ admits an action of $G = \text{GL}_1(\mathbb{C})$, the multiplicative group of nonzero complex numbers, via

$$(g \cdot f)(x, y) = f(g^{-1}x, gy)$$

for $g \in G$ and $f \in \mathbb{C}[x, y]$. Note that the invariants under the $G$-action are those polynomials in which the degree of each term is the same with respect to $x$ as it is with respect to $y$; in other words, $\mathbb{C}[x, y]^G$ is generated by the monomials $x_i y_j$. (This is a special case of the first fundamental theorem of invariant theory; see [13, Section 5.2.1].) But since the kernel of the ring homomorphism $w_{ij} \mapsto x_i y_j$ is precisely the determinantal ideal $\mathcal{I}$, we have $\mathbb{C}[x, y]^G \simeq \mathbb{C}[D_n^{\leq 1}]$. (This is a special case of the second fundamental theorem of invariant theory; see [13, Lemma 5.2.4].)

As a result of Howe duality in Type A — the delicate details of which are expounded in [16; 17] — the space $\mathbb{C}[x, y]$ is also a module under the action of the Lie algebra $u(n,n)$ by differential operators. Upon complexification, this gives rise to an action by the Lie algebra $\text{gl}_{2n}$ of $2n \times 2n$ complex matrices. In
particular, the invariant subring \( C[G] \) is the irreducible, infinite-dimensional \( gl_{2n} \)-module with highest weight \((-1, \ldots, -1, 0, \ldots, 0)\) in standard coordinates, where there are \( n \) entries of \(-1\).

Explicitly, \( gl_{2n} \) acts via differential operators on \( C[G] \), of the following four forms (see [8, Section 2], for the action in full detail):

1. \( x_i \frac{\partial}{\partial x_j} \) (Euler operators; technically the action includes the extra term \( +\delta_{ij} \));
2. \( y_i \frac{\partial}{\partial y_j} \) (Euler operators);
3. \( \frac{\partial^2}{\partial x_i \partial y_j} \) (“raising operators”);
4. \( x_i y_j \) (“lowering operators”).

Note that all these operators preserve the difference between the degree with respect to \( x \) and the degree with respect to \( y \). Therefore the \( gl_{2n} \)-action preserves \( C[G] \), which we observed is generated by the elements \( x_i y_j \).

This \( gl_{2n} \)-action can be described in terms of our EMD\(^s_2\) setting in this paper. First, observe that any degree-\( s \) monic monomial in \( C[G] \) corresponds uniquely to an ordered pair of compositions \((\mu, \nu) \in C(s, n) \times C(s, n)\), via

\[
(\mu, \nu) \leftrightarrow x^{\mu} y^{\nu} := x_1^{\mu(1)} \cdots x_n^{\mu(n)} y_1^{\nu(1)} \cdots y_n^{\nu(n)}. 
\]

Now we can see how each type (1)–(4) of differential operator has an interpretation in the EMD\(^s_2\) context. Up to scaling by coefficients, we observe the following:

1. The Euler operator \( x_i \frac{\partial}{\partial x_j} \) corresponds to moving 1 unit in \( \mu \) from bin \( j \) to bin \( i \), since the exponent of \( x_j \) decreases by 1 and the exponent of \( x_i \) increases by 1.
2. The Euler operator \( y_i \frac{\partial}{\partial y_j} \) corresponds to moving 1 unit in \( \nu \) from bin \( j \) to bin \( i \).
3. The raising operator \( \frac{\partial^2}{\partial x_i \partial y_j} \) corresponds to removing 1 unit from each composition: from bin \( i \) in \( \mu \) and from bin \( j \) in \( \nu \).
4. The lowering operator \( x_i y_j \) corresponds to adding 1 unit to each composition: to bin \( i \) in \( \mu \) and to bin \( j \) in \( \nu \).

It will be interesting to study further whether this connection to representation theory might be exploited in applications of EMD\(^s_2\).

8. Proof of Proposition 2

The methods in this paper depended heavily upon the fact that we need to consider only those arrays \( J \) whose support is a chain. This followed from the statement in Proposition 2 — yet to be proved — that our cost array \( C \) has the Monge property. Before proving this here, we state three useful lemmas, the first of which is proved in [2; 22]:

**Lemma 8.** An \( n \times \cdots \times n \) array \( A \) has the Monge property if and only if every two-dimensional plane of \( A \) has the Monge property.
To make this explicit, we choose any two distinct indices \( i, j \) from \([1, \ldots, d]\), and then fix the remaining \( d - 2 \) coordinates at the values \( \tilde{m}_1, \ldots, \tilde{m}_{i-1}, \tilde{m}_{i+1}, \ldots, \tilde{m}_{j-1}, \tilde{m}_{j+1}, \ldots, \tilde{m}_d \in [n] \). Then we will write \( \vec{m}_{i,j}^{k,\ell} := (\tilde{m}_1, \ldots, \tilde{m}_{i-1}, k, \tilde{m}_{i+1}, \ldots, \tilde{m}_{j-1}, \ell, \tilde{m}_{j+1}, \ldots, \tilde{m}_d) \). In other words, \( \vec{m}_{i,j}^{k,\ell} \) is the vector in which the \( i \)-th coordinate is \( k \), the \( j \)-th coordinate is \( \ell \), and the remaining coordinates are the fixed values \( \tilde{m}_1, \ldots, \tilde{m}_d \). Now we can naturally define the two-dimensional subarray \( A_{i,j}^{k,\ell} \) in which

\[
A_{i,j}^{k,\ell}(k, \ell) := A(\vec{m}_{i,j}^{k,\ell}).
\] (19)

Then Lemma 8 states that \( A \) has the Monge property if and only if \( A_{i,j}^{k,\ell} \) has the Monge property for every choice of distinct \( i \) and \( j \).

This reduction to the two-dimensional case is extremely useful because of the following characterization of two-dimensional Monge arrays, proved in [22]:

**Lemma 9.** Let \( A \) be an \( n \times n \) array. Then \( A \) has the Monge property if and only if

\[
A(k, \ell) + A(k + 1, \ell + 1) \leq A(k + 1, \ell) + A(k, \ell + 1)
\]

for all \( k, \ell \in [n - 1] \).

In other words, choose a position \((k, \ell)\) and then consider the \( 2 \times 2 \) subarray consisting of \( A(k, \ell) \) and its three neighbors to the east, south, and southeast. The condition displayed in the lemma means that the sum of the upper-left and lower-right entries must never be greater than the sum of the lower-left and upper-right entries.

We will need one final lemma, specific to the cost function \( C \) in this paper. Recall from Proposition 1 that if we let \( \tilde{m} \) denote a vector \( m \) with its coordinates rearranged in ascending order, then

\[
C(m) = -\tilde{m}_1 - \cdots - \tilde{m}_{\lfloor (d+1)/2 \rfloor} + \tilde{m}_{\lfloor (d+1)/2 \rfloor + 1} + \cdots + \tilde{m}_d \quad (\text{d even}),
\]

or

\[
C(m) = -\tilde{m}_1 - \cdots - \tilde{m}_{\lfloor (d+1)/2 \rfloor - 1} + \tilde{m}_{\lfloor (d+1)/2 \rfloor + 1} + \cdots + \tilde{m}_d \quad (\text{d odd}).
\]

The index \( \lfloor (d+1)/2 \rfloor \) gave a kind of “median” of the coordinates in \( m \); from now on, however, we will work instead with

\[
M := \left\lfloor \frac{d+1}{2} \right\rfloor + 1 = \left\lfloor \frac{d+2}{2} \right\rfloor.
\]

Intuitively, this index \( M \) gives the next-greatest coordinate after the “median.” The picture is the following, where the vertical lines divide the coordinates into two equal sets (with one leftover coordinate in the middle if \( d \) is odd):

\[
d \text{ even: } \quad m = (\tilde{m}_1, \ldots, \tilde{m}_{M-1}, | \tilde{m}_M, \ldots, \tilde{m}_d),
\]

\[
d \text{ odd: } \quad m = (\tilde{m}_1, \ldots, | \tilde{m}_{M-1}, \tilde{m}_M, \ldots, \tilde{m}_d).
\]

With this indexing in mind, we state our final lemma, which records the effect on \( C(m) \) of adding 1 to a single coordinate \( m_i \). Recall from earlier that \( e(i) \) denotes the vector whose coordinates are all 0 except for a 1 in the \( i \)-th component.

**Lemma 10.** Adding 1 to a single coordinate \( m_i \) of \( m \) has one of three effects on \( C(m) \): it either increases by 1, decreases by 1, or remains the same. The effect depends on the value of \( m_i \) relative to the other coordinates of \( m \):

1. \( C(m + e(i)) = C(m) + 1 \) if \( m_i \geq \tilde{m}_M \).
Case 2 Case 4

To show that this condition holds true, we need to examine six possible cases, indicated in Table 4.

Lemma 10. Now let

\[ C(m + e(i)) = C(m) \]

We prove each of the three cases; the reader may find it helpful to keep an eye on the two possible “pictures” of \( \tilde{m} \) displayed before this lemma, along with the two possible sums for \( C(m) \) displayed just before that.

1. Assume \( m_i \geq \tilde{m}_M \). Then \( m_i + 1 > \tilde{m}_M \), and so in the sum defining \( C(m) \), we must have positive \( m_i \) replaced by positive \((m_i + 1)\). Hence \( C(m) \) has increased by 1.

2. (a) Assume \( d \) is even and \( m_i < \tilde{m}_M \). Then \( m_i + 1 \leq \tilde{m}_M \), and so in the sum defining \( C(m) \), we must have negative \( m_i \) replaced by negative \((m_i + 1)\). Hence \( C(m) \) has decreased by 1.

(b) Assume \( d \) is odd and \( m_i < \tilde{m}_{M-1} \). Then \( m_i + 1 \leq \tilde{m}_{M-1} \), and so we must have negative \( m_i \) replaced by negative \((m_i + 1)\). Hence \( C(m) \) has decreased by 1.

3. Assume \( d \) is odd and \( m_i = \tilde{m}_{M-1} < \tilde{m}_M \); note that \( \tilde{m}_{M-1} \) does not appear in the sum defining \( C(m) \). Then \( \tilde{m}_{M-2} < m_i + 1 \leq \tilde{m}_M \), and so \( m_i + 1 \) still does not appear in the sum defining \( C(m + e(i)) \). Hence \( C(m) \) remains unchanged. □

We are now ready for the proof, in which we show that an arbitrary two-dimensional subarray of \( C \) has the Monge property.

Proof of Proposition 2. Let \( i, j \) be two distinct indices in \( \{1, \ldots, d\} \). Fix the remaining coordinates \( \tilde{m}_1, \ldots, \tilde{m}_d \) as above, and let \( C^{i,j} \) be the corresponding two-dimensional subarray of \( C \) defined in (19). Now let \( m_i, m_j \in [n - 1] \). By Lemmas 8 and 9, it will suffice to show that

\[
C^{i,j}(m_i, m_j) + C^{i,j}(m_i + 1, m_j + 1) \leq C^{i,j}(m_i + 1, m_j) + C^{i,j}(m_i, m_j + 1).
\]

But this condition can be rewritten as the following, where we simply write \( \tilde{m} \) for \( \tilde{m}^{i,j} \):

\[
C(\tilde{m}) + C(\tilde{m} + e(i) + e(j)) \leq C(\tilde{m} + e(i)) + C(\tilde{m} + e(j)). \tag{20}
\]

To show that this condition holds true, we need to examine six possible cases, indicated in Table 4.

We now examine each of these cases. All simplifications are directly justified by the results in Lemma 10.

<table>
<thead>
<tr>
<th>Case</th>
<th>Condition</th>
<th>( C(\tilde{m} + e(i)) = C(\tilde{m}) + 1 )</th>
<th>( C(\tilde{m} + e(i)) = C(\tilde{m}) - 1 )</th>
<th>( C(\tilde{m} + e(i)) = C(\tilde{m}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( C(\tilde{m} + e(j)) = C(\tilde{m}) + 1 )</td>
<td>Case 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>( C(\tilde{m} + e(j)) = C(\tilde{m}) )</td>
<td>Case 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>( C(\tilde{m} + e(j)) = C(\tilde{m}) )</td>
<td>Case 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>( C(\tilde{m} + e(j)) = C(\tilde{m}) )</td>
<td>Case 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>( C(\tilde{m} + e(j)) = C(\tilde{m}) )</td>
<td>Case 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>( C(\tilde{m} + e(j)) = C(\tilde{m}) )</td>
<td>Case 6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Six possible cases, depending on whether adding 1 to \( m_i \) and \( m_j \) (independently) causes \( C \) to increase, decrease, or remain the same.
Case 1. In this case, the right-hand side of (20) is $2 \cdot C(\bar{m}) + 2$. For the left-hand side, we know in general that $C(\bar{m} + e(i) + e(j)) = C(\bar{m} + e(i) + e(j))$, which by Lemma 10 can be no greater than $C(\bar{m}) + 2$. Hence the inequality in (20) must hold.

Case 2. In this case, the right-hand side of (20) is $2 \cdot C(\bar{m})$. As for the second term on the left-hand side, by Lemma 10, we must have $m_i \geq \tilde{m}_M$; meanwhile, $m_j$ is strictly less than either $\tilde{m}_M$ (if $d$ is even) or $\tilde{m}_{M-1}$ (if $d$ is odd), and so neither inequality is affected by adding 1 to $m_i$. Therefore we have

$$C(\bar{m} + e(i) + e(j)) = C((\bar{m} + e(i)) + e(j))$$
$$= C((\bar{m} + e(i)) - 1$$
$$= C(\bar{m}) + 1 - 1$$
$$= C(\bar{m}).$$

Hence we have an equality in (20).

Case 3. Similar to Case 2, the two additions are independent of each other. The right-hand side of (20) is $2 \cdot C(\bar{m}) + 1$. In this case, we must have $d$ odd; also, $m_i \geq \tilde{m}_M$, along with $m_j = \tilde{m}_{M-1} < \tilde{m}_M$. Then

$$C(\bar{m} + e(i) + e(j)) = C((\bar{m} + e(i)) + e(j))$$
$$= C((\bar{m} + e(i))$$
$$= C(\bar{m}) + 1.$$

Again we obtain an equality in (20).

Case 4. The right-hand side of (20) is $2 \cdot C(\bar{m}) - 2$. If $d$ is even, then both $m_i$ and $m_j$ are strictly less than $\tilde{m}_M$, and if $d$ is odd, then both are strictly less than $\tilde{m}_{M-1}$. Either way, after adding 1 to $m_i$, the same inequality still holds for $m_j$, and so again we have

$$C(\bar{m} + e(i) + e(j)) = C((\bar{m} + e(i)) + e(j))$$
$$= C((\bar{m} + e(i)) - 1$$
$$= C(\bar{m}) - 1 - 1$$
$$= C(\bar{m}) - 2,$$

and we get an equality in (20).

Case 5. The right-hand side of (20) is $2 \cdot C(\bar{m}) - 1$. In this case, $d$ must be odd, with

$$m_i < \tilde{m}_{M-1} = m_j < \tilde{m}_M.$$

After adding 1 to $m_j$, we still have $m_i$ less than the $(M-1)$-th component in the new rearranged vector, and so the effects of the two additions are independent. We obtain

$$C(\bar{m} + e(i) + e(j)) = C((\bar{m} + e(j)) + e(i))$$
$$= C((\bar{m} + e(j)) - 1$$
$$= C(\bar{m}) - 1$$

and so we have an equality in (20).
Case 6. This is the slightly surprising case, in which the two additions are not independent of each other. The right-hand side of (20) is \(2 \cdot C(\vec{m})\), and we know that \(d\) must be odd, with \(m_i = m_j = \tilde{m}_{M-1} < \tilde{m}_M\). After adding 1 to \(m_i\), we obtain a vector \(\vec{m}'\) in which \(m'_i = m_j\) is now strictly less than \(\tilde{m}'_{M-1}\), and so now adding 1 to \(m_j\) results in an overall decrease by 1. Hence we have

\[
C(\vec{m} + e(i) + e(j)) = C((\vec{m} + e(i)) + e(j)) \\
= C((\vec{m} + e(i)) - 1 \\
= C(\vec{m}) - 1.
\]

Hence the left-hand side of (20) is less than the right-hand side, and the condition is still satisfied.

We have exhausted all possible cases, and so since (20) holds in each of them, the two-dimensional array \(C_{i,j}\) has the Monge property. Since \(i\) and \(j\) were arbitrary, every two-dimensional subarray of \(C\) has the Monge property, and so by Lemma 8, we conclude that \(C\) itself has the Monge property. □

9. Proof of Proposition 6

We prove here the recursive definition for our generating function \(H_n\) from Section 5.1. Recall that the formal definition is

\[
H_n(z, t) := \sum_{s=0}^{\infty} \left( \sum_{\mu \in C(s, n_1) \times \cdots \times C(s, n_d)} z^{\text{EMD}_d(\mu)} \right) t^s,
\]

and Proposition 6 states the recursion as follows, where the sum is over all nonempty subsets \(A \subseteq [d]z\):

\[
H_n = \sum_A (-1)^{|A|-1} \cdot \frac{H_{n-e(A)}}{1 - zC(n)}
\]

where \(H_{(1^d)} = 1/(1-t)\), and \(H_{n-e(A)} = 0\) if \(n-e(A)\) contains a 0.

Proof of Proposition 6. Each \(\mu\) corresponds to a unique monomial

\[
\mu \leftrightarrow \prod_m w_m^{J_\mu(m)},
\]

where \(\mu \leftrightarrow J_\mu\) is the RSK correspondence in (5). The variables \(w_m\) are indexed by multi-indices \(m \in [n_1] \times \cdots \times [n_d]\). Note that the degree of this monomial equals \(s\) (the sum of the entries of \(J_\mu\)). Now making the substitution

\[
w_m \mapsto z^{C(m)} t,
\]

the above correspondence gives us the map

\[
\mu \leftrightarrow \prod_m w_m^{J_\mu(m)} \mapsto z^{\text{EMD}_d(\mu)} t^s.
\]

Therefore the generating function \(H_n\) is just the image of the formal sum \(H_n^*\) of all monomials of the form (21), under the substitution (22); as \(s\) ranges over all nonnegative integers, there is one monomial in \(H_n^*\) for each possible \(\mu \in C(s, n_1) \times \cdots \times C(s, n_d)\).
Since \( m \preceq n \) for all \( m \), every array \( J_\mu \) is allowed to contain \( n \) in its support without violating the chain condition. This means that every monomial in \( H_n^* \) is allowed to contain the variable \( w_n \), and so we may factor out the sum of all possible powers of \( w_n \), rewriting as

\[
H_n^* = \sum_\mu \left( \prod_m w_m^{J_\mu (m)} \right) = \sum_{r=0}^{\infty} w_n^r \cdot f ( w_{m \neq n} ) = \frac{f ( w_{m \neq n} )}{1 - w_n},
\]

where \( f \) is an infinite formal sum of monomials in the variables \( w_m \) where \( m \neq n \). Now we focus on rewriting this numerator \( f \). Suppose we subtract 1 from exactly one of the coordinates of \( n \); the possible results are \( n - e(i) \) for \( i = 1, \ldots, d \). Now, on one hand, any monomial in \( f \) containing the variable \( w_{n - e(i)} \) appears in \( H_{n - e(i)}^* \). But on the other hand, note that all of these \( n - e(i) \) are mutually incomparable under the product order, and so at most one of them can be in the support of some \( J_\mu \), because of the chain condition. Therefore any monomial in \( f \) contains at most one of the variables \( w_{n - e(i)} \). But the sum \( \sum_{i=1}^d H_{n - e(i)}^* \) still overcounts the monomials appearing in \( f \), since the same monomial may appear in several distinct summands.

In other words, we want \( f \) to be the formal sum of the union (without multiplicity) of the monomials which appear in the summands \( H_{n-e(i)}^* \). We can achieve this by using the inclusion–exclusion principle: subtract those monomials which appear in at least 2 of the summands, then add back the monomials which appear in at least 3 of the summands, then subtract those appearing in at least 4 summands, and so on, until we arrive at those monomials appearing in all \( d \) of the summands. We can write this inclusion–exclusion as an alternating sum over nonempty subsets \( A \subseteq [d] \), adding when \( |A| \) is odd and subtracting when \( |A| \) is even:

\[
f = \sum_A (-1)^{|A|-1} \cdot H_{n-e(A)}^*.
\]

Finally, applying the substitution (22), we obtain

\[
H_n = H_n^* \bigg|_{w_m = z^{C(m)} t} = \sum_A (-1)^{|A|-1} \cdot H_{n-e(A)}^* \bigg|_{w_m = z^{C(m)} t} = \sum_A (-1)^{|A|-1} \cdot H_{n-e(A)}^* \bigg|_{1 - w_n = 1 - z^{C(n)} t},
\]

proving the recursion.

As for the base case \( H_{1, \ldots, 1} = 1/(1 - t) \), there is only one element in \( C(s, 1) \), and so since every \( n_i = 1 \), the inside sum in (7) has only one term; moreover, this unique \( \mu \) is just \( d \) copies of the same trivial composition of \( s \) into 1 part, meaning that \( \text{EMD}_d^s (\mu) = 0 \). Hence \( H_{1, \ldots, 1}(z, t) = \sum z^0 t^s = \sum_t t^s \), whose closed form is \( 1/(1 - t) \). Likewise, since \( C(s, 0) \) is empty, we must have \( H = 0 \) if any of the \( n_i \) become 0.

\[\Box\]

Acknowledgements

The author would like to thank Rebecca Bourn and Jeb Willenbring for the conversations about their original paper [5]. Jeb’s observations about the connections to representation theory were especially vital to Section 7. The author is also grateful for the many helpful comments from both referees.

References


Received 2020-10-26. Revised 2021-04-05. Accepted 2021-05-18.

WILLIAM Q. ERICKSON: wqe@uwm.edu
Department of Mathematical Sciences, University of Wisconsin–Milwaukee, Milwaukee, WI, United States
We relate scattering amplitudes in particle physics to maximum likelihood estimation for discrete models in algebraic statistics. The scattering potential plays the role of the log-likelihood function, and its critical points are solutions to rational function equations. We study the ML degree of low-rank tensor models in statistics, and we revisit physical theories proposed by Arkani-Hamed, Cachazo and their collaborators. Recent advances in numerical algebraic geometry are employed to compute and certify critical points. We also discuss positive models and how to compute their amplitudes.

1. Introduction

Likelihood equations are equations among rational functions that arise in various contexts, notably in high energy physics [2; 3] and in algebraic statistics [24; 28]. We establish a new link between these two fields. This is interesting for both sides, and may lead to unexpected advances in nonlinear algebra [23]. Specifically, we develop the connection between maximum likelihood estimation [13; 18] and the geometric theory of scattering amplitudes [9; 11]. Our goal is the practical solution of likelihood equations with certified numerical methods [7; 8].

On the statistics side, a discrete model is a subvariety $X$ of the real projective space $\mathbb{P}^n$, which is assumed to intersect the simplex $\Delta_n$ of positive points. The homogeneous coordinates $p = (p_0 : p_1 : \cdots : p_n)$ are interpreted as unknown probabilities for the $n+1$ states, subject to the constraint that $p$ lies in the model $X$. When collecting data, we write $s_i$ for the number of times the $i$-th state was observed. The data vector $s = (s_0, s_1, \ldots, s_n)$ is also viewed modulo scaling, i.e., $s$ lies in $\Delta_n \subset \mathbb{P}^n$. We are interested in the log-likelihood function

$$s_0 \cdot \log(p_0) + s_1 \cdot \log(p_1) + \cdots + s_n \cdot \log(p_n) - (s_0 + s_1 + \cdots + s_n) \cdot \log(p_0 + p_1 + \cdots + p_n).$$

This is a well-defined function on $\Delta_n \subset \mathbb{P}^n$. The aim of likelihood inference in data analysis is to maximize (1) over all points $p$ in the model $X \cap \Delta_n$. In algebraic statistics, we care about all complex critical points. Their number, for generic $s$, is the maximum likelihood (ML) degree of the model $X$. If $X$ is smooth then the ML degree equals the Euler characteristic of the open variety $X^o$, which is the complement of the divisor in $X$ defined by $p_0 p_1 \cdots p_n (\sum_{i=0}^n p_i) = 0$. Computing ML degrees and identifying critical points is an active area of research [25].

The situation is similar in the study of potentials and associated amplitudes in physics. Here the role of the data vector $s$ is played by the vector of Mandelstam invariants, which is constrained to lie in the


Keywords: likelihood equations, scattering amplitudes, numerical nonlinear algebra.
kinematic space. This mirrors the constraint that the coefficients in (1) sum to zero. In recent physical theories [4; 9; 11], the variety $X^o$ is the configuration space of $m$ points in general position in $\mathbb{P}^{k-1}$, up to projective transformations. This is modeled by the Grassmannian $\text{Gr}(k, m) \subset \mathbb{P}^{\binom{m}{k} - 1}$, modulo the action of the torus $(\mathbb{C}^*)^m$. Let $\text{Gr}(k, m)^o$ be the open Grassmannian where all Plücker coordinates are nonzero. We work in the $(k-1)(m-k-1)$-dimensional manifold $X^o = \text{Gr}(k, m)^o / (\mathbb{C}^*)^m$. The ML degree of $X^o$ is the number of critical points on $X^o$ of the potential function, for generic $s$. A scattering amplitude is the sum of a certain rational function over all critical points. This is a global residue [14; 15], so it evaluates to a rational function in the Mandelstam invariants $s$.

The present article is organized as follows. Section 2 develops the promised connection for the Grassmannian of lines ($k = 2$). Here $X^o$ is the moduli space $\mathcal{M}_{0, m}$ of $m$ marked points in $\mathbb{P}^1$. This prominent space is here recast as a statistical model. The ML degree of that model is $(m - 3)!$ and all critical points are real, thanks to Varchenko’s theorem [24, Theorem 1.5]. Our computational results for $m \leq 13$ are found in Table 1. This is extended to arbitrary linear statistical models in Section 3. We show in that setting how the software HomotopyContinuation.jl [7; 8] is used to find and certify all critical points of (1). A key idea is to refrain from clearing denominators and work with rational functions directly.

Section 4 concerns higher Grassmannians ($k \geq 3$) and their associated likelihood equations. We focus on the case $m = 8$, $k = 3$, where the amplitudes literature [11; 12] reports the ML degree 188112. We interpret this CEGM theory as a nonlinear statistical model with $n = 47$. Our method computes and certifies all 188112 critical points in a few minutes, for random Mandelstam invariants with $s \geq 0$ in (1), and we show that most of them are real.

In Section 5 we apply our approach to a class of models that is important in statistics, namely conditional independence of identically distributed random variables. This corresponds to symmetric tensors of low rank, so here $X$ is a Veronese secant variety. We determine the ML degree in several new cases, well beyond the degree 12 for tossing coins in the running example of [17]. This opens up a new chapter in likelihood inference for tensors.

In Section 6 we finally turn to amplitudes. We build on the theory of stringy canonical forms due to Arkani-Hamed, He and Lam [5]. Definition 12 introduces a statistical version of positive geometries [4; 6]. The biadjoint theory amplitudes in [5] are limits of marginal likelihood integrals. They can be computed combinatorially from Newton polytopes, or as global residues, by summing the reciprocal toric Hessian of the function (1) over its critical points.

2. Points on the Line

We begin with a first direct connection between algebraic statistics and particle physics. The $m$-particle CHY scattering equations [9] will be presented as likelihood equations for a linear statistical model on the moduli space $\mathcal{M}_{0, m}$. We introduce these rational function equations, and we solve them using state-of-the-art tools from numerical algebraic geometry [7; 8; 27].

We consider $m \geq 4$ points in $\mathbb{P}^1$ whose homogeneous coordinates are the columns of

$$
\begin{bmatrix}
0 & 1 & 1 & \cdots & 1 & 1 & 1 \\
-1 & 0 & x_1 & x_2 & \cdots & x_{m-4} & x_{m-3} & 1
\end{bmatrix}
$$

(2)
We write $q_{ij}$ for the $2 \times 2$ minor given by the $i$-th and the $j$-th column of this $2 \times m$-matrix. The moduli space $\mathcal{M}_{0,m} = \text{Gr}(2, m)^{\circ} / (\mathbb{C}^*)^n$ is the set of points for which these minors are nonzero. This is the complement of a hyperplane arrangement in $\mathbb{C}^{m-3}$. The corresponding real arrangement in $\mathbb{R}^{m-3}$ has $(m-3)!$ bounded regions, given by the possible orderings of $x_1, x_2, \ldots, x_{m-3}$ in $[0, 1]$. These regions are simplices and they define a triangulation of the cube $[0, 1]^{m-3}$. One of them is the positive region

$$
\mathcal{M}_{0,m}^+ = \{0 < x_1 < x_2 < \cdots < x_{m-3} < 1\}.
$$

We now define a statistical model $X$ on $n+1 = m(m-3)/2$ states. The states are the pairs $(i, j)$ where $2 \leq i < j \leq m$ and $(i, j) \neq (2, m)$. The parameter vector $(x_1, \ldots, x_{m-3})$ is assumed to lie in $\mathcal{M}_{0,m}^+$. The probability of observing the state $(i, j)$ is $p_{ij} = \alpha_{ij} q_{ij}$, where

$$
\alpha_{im} = \frac{1}{m-3}, \quad \alpha_{ij} = \frac{1}{(m-3)^2} \quad \text{and} \quad \alpha_{2j} = \frac{2m-2j-1}{(m-3)^2} \quad \text{for} \quad 3 \leq i < j \leq m-1. \quad (3)
$$

These positive constants are chosen so that the sum of the $n+1$ linear expressions $p_{ij}$ equals 1.

Suppose we collect data. For each of the $n+1$ states $(i, j)$ as above, we record the number $s_{ij}$ of observations of that state. The aim of statistical inference is to find the point $\hat{x} = (\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_{m-3})$ in the parameter space $\mathcal{M}_{0,m}^+$ that best explains the data. Adopting the classical frequentist framework, this is done by maximizing the log-likelihood function

$$
L(x) = \sum_{(i, j)} s_{ij} \log(p_{ij}(x)) = \sum_{(i, j)} s_{ij} \log(q_{ij}(x)) + \text{const.} \quad (4)
$$

We write Crit$(L)$ for the set of critical points of $L$, i.e., the solutions of the likelihood equations

$$
\frac{\partial L}{\partial x_1} = \frac{\partial L}{\partial x_2} = \cdots = \frac{\partial L}{\partial x_{m-3}} = 0. \quad (5)
$$

This is a system of $m-3$ rational function equations in the $m-3$ unknowns $x_1, \ldots, x_{m-3}$.

**Proposition 1.** If all $s_{ij}$ are positive then (5) has precisely $(m-3)!$ complex solutions. All solutions are real, and there is one solution for each of the orderings of the $m-3$ coordinates.

**Proof.** This result is known in the physics literature due to Cachazo, Mizera and Zhang [10]. We here derive it from Varchenko’s theorem in algebraic statistics [13, Theorem 13]. That theorem states that the likelihood equations of a linear space $X$ have only real solutions, and there is one solution in each bounded region of the arrangement in the real linear space $X_{\mathbb{R}}$ defined by the $n+1$ hyperplanes $\{p_i = 0\}$. For the CHY model, we identify $X_{\mathbb{R}}$ with the parameter space $\mathbb{R}^{m-3}$, where the hyperplanes are $\{x_i = 0\}, \{x_i = 1\}$ and $\{x_i = x_j\}$. Every point with $x_i < 0$ or $x_i > 1$ for some $i$ can be moved to infinity without crossing a hyperplane. This implies that the bounded regions are the simplices $\{0 < x_{\pi_1} < \cdots < x_{\pi_{m-3}} < 1\}$, where $\pi$ runs over all $(m-3)!$ permutations of the set $\{1, 2, \ldots, m-3\}$. \qed
Example 2 \((m = 6, n = 8)\). We consider a linear model \(X\) on nine states 23, 24, \ldots, 56. Their probabilities, which sum to 1, are linear functions of three model parameters \(x_1, x_2, x_3\):

\[
\begin{align*}
p_{23} &= \frac{5x_1}{9}, & p_{24} &= \frac{x_2}{3}, & p_{25} &= \frac{x_3}{9}, \\
p_{34} &= \frac{(x_2 - x_1)}{9}, & p_{35} &= \frac{(x_3 - x_1)}{9}, & p_{45} &= \frac{(x_3 - x_2)}{9}, \\
p_{36} &= \frac{(1 - x_1)}{3}, & p_{46} &= \frac{(1 - x_2)}{3}, & p_{56} &= \frac{(1 - x_3)}{3}.
\end{align*}
\]

This maps the tetrahedron \(M_{0,6}^+ = \{0 < x_1 < x_2 < x_3 < 1\}\) into the probability simplex \(\Delta_8\). Suppose we collect data with sample size 170, and the resulting data vector has coordinates

\[
s_{23} = 25, \quad s_{24} = 23, \quad s_{25} = 16, \quad s_{34} = 12, \quad s_{35} = 22, \quad s_{45} = 16, \quad s_{36} = 14, \quad s_{46} = 15, \quad s_{56} = 27. \tag{7}
\]

We must solve an optimization problem on \(M_{0,6}^+\), namely to maximize the function

\[
L = s_{23} \log(p_{23}) + s_{24} \log(p_{24}) + s_{25} \log(p_{25}) + s_{34} \log(p_{34}) + s_{35} \log(p_{35}) + s_{45} \log(p_{45})
+ s_{36} \log(p_{36}) + s_{46} \log(p_{46}) + s_{56} \log(p_{56}). \tag{8}
\]

The set \(\text{Crit}(L)\) has one point in each bounded region of the arrangement of nine planes \(\{p_{ij} = 0\}\) in \(\mathbb{R}^3\). The six bounded regions lie in the cube \([0, 1]^3\). They correspond to the orderings of the values \(x_1, x_2, x_3\). For instance, for the data in (7), the six critical points are

\[
\hat{x}_1 = 0.240043275929170, \quad \hat{x}_2 = 0.508172206739870, \quad \hat{x}_3 = 0.777005866817260; \\
x_1 = 0.223437550855307, \quad x_2 = 0.843543048681696, \quad x_3 = 0.51870638908326; \\
x_1 = 0.481967726451097, \quad x_2 = 0.235545240880672, \quad x_3 = 0.781115679885971; \\
x_1 = 0.618277926209287, \quad x_2 = 0.8519744545945199, \quad x_3 = 0.155992558374125; \\
x_1 = 0.861996060709608, \quad x_2 = 0.217605043343923, \quad x_3 = 0.453238947004789; \\
x_1 = 0.863192417250353, \quad x_2 = 0.578669456252017, \quad x_3 = 0.157960116395912.
\]

The first triple is the maximum likelihood estimate. The learned distribution in the model is

\[
\begin{align*}
\hat{p}_{23} &= 0.13336, & \hat{p}_{24} &= 0.16939, & \hat{p}_{25} &= 0.08633, & \hat{p}_{34} &= 0.02979, & \hat{p}_{35} &= 0.05966, \\
\hat{p}_{36} &= 0.25332, & \hat{p}_{45} &= 0.02987, & \hat{p}_{46} &= 0.16394, & \hat{p}_{56} &= 0.07433. \tag{9}
\end{align*}
\]

We shall see that this computation can be done for much larger values of \(m\) and \(n\).

We now turn to physics. In quantum field theory, the \(s_{ij}\) are known as Mandelstam invariants. One writes them in a symmetric \(m \times m\)-matrix with zeros on the diagonal, so we have \(s_{ii} = 0\) and \(s_{ij} = s_{ji}\). Momentum conservation means that the row sums are zero, i.e., \(\sum_{j=1}^{m} s_{ij} = 0\) for \(i = 1, \ldots, m\). These equations define the kinematic space, which has dimension \(n + 1 = \binom{m}{2} - m\). On that space, the \(m\) Mandelstam invariants \(s_{12}, s_{13}, \ldots, s_{1m}\) and \(s_{2m}\) can be written uniquely in terms of our counts \(s_{ij}\) in the statistical model above. For instance, for \(m = 6\), the kinematic space is parametrized by the nine counts...
The scattering potential in the CHY model coincides with the log-likelihood function $L$, up to the additive constant in (4). Hence the scattering equations are the likelihood equations.

We now come to the punchline of this section: current off-the-shelf software from numerical algebraic geometry is highly efficient and reliable in solving our equations. For our computations we used the julia package HomotopyContinuation.jl, due to Breiding and Timme [7], including the recent certification feature [8] which rests on interval arithmetic.

In Table 1 we present the timings we obtained for solving the scattering equations (5) when the number of particles is $m = 10, 11, 12, 13$. Recall that the solutions are the critical points of $L$ in the moduli space $\mathcal{M}_{0,m}$. In later sections we apply these methods for solving likelihood equations coming from other statistical models, including higher Grassmannians.

The first two columns in Table 1 show the number $n + 1 = m(m - 3)/2$ of states in the statistical model and the ML degree $(m - 3)!$. The last three columns show computation times. The most relevant among these is $t_R$. This is the time in seconds for computing all $(m - 3)!$ real critical points for a given system of Mandelstam invariants $s_{ij} > 0$. For instance, for $m = 12$, it takes less than one minute to compute all $(12 - 3)! = 362880$ solutions.

### Table 1. Computing and certifying solutions to CHY scattering equations with the method in Section 3. Here $t_C, t_R, t_{\text{cert}}$ denote timings (in seconds) that are explained in Example 3.

<table>
<thead>
<tr>
<th>$m$</th>
<th>$n + 1$</th>
<th>$(m-3)!$</th>
<th>$t_C$</th>
<th>$t_R$</th>
<th>$t_{\text{cert}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>35</td>
<td>5040</td>
<td>0.75</td>
<td>0.28</td>
<td>0.5</td>
</tr>
<tr>
<td>11</td>
<td>44</td>
<td>40320</td>
<td>13.4</td>
<td>3.4</td>
<td>4.0</td>
</tr>
<tr>
<td>12</td>
<td>54</td>
<td>362880</td>
<td>124.6</td>
<td>43.7</td>
<td>45.0</td>
</tr>
<tr>
<td>13</td>
<td>65</td>
<td>3628800</td>
<td>2141.5</td>
<td>578.2</td>
<td>1178.0</td>
</tr>
</tbody>
</table>

3. **Linear models and how to compute**

We here explain our methodology for solving the likelihood equations. For ease of illustration we consider linear statistical models, with the understanding that the computations are analogous for nonlinear models. The scope of that becomes visible in the next two sections.

Fix affine-linear polynomials $p_0(x), p_1(x), \ldots, p_n(x)$ with real coefficients in $d$ unknowns $x = (x_1, x_2, \ldots, x_d)$. We assume that $p_0(x) + p_1(x) + \cdots + p_n(x) = 1$ and that the convex polytope

$$\Theta = \{x \in \mathbb{R}^d : p_i(x) \geq 0\}$$
has dimension \( d \). The model is the \( d \)-dimensional linear space \( X \) in \( \mathbb{P}^n \) parametrized by

\[ x \mapsto (p_0(x) : \cdots : p_n(x)). \]

Given any positive real data vector \( s = (s_0, s_1, \ldots, s_n) \), we wish to find all critical points of the log-likelihood function \( L \) in (1).

By Varchenko’s theorem, all complex critical points are real, and there is one critical point in each bounded region of the arrangement of \( n + 1 \) hyperplanes \( \{ p_i(x) = 0 \} \) in \( \mathbb{R}^d \). One of these bounded regions is the polytope \( \Theta \), so this contains a unique critical point \( \hat{x} \). Its image \( \hat{p} = p(\hat{x}) \) in \( \Delta_n \) is the distribution in the model \( X \) that best explains the data \( s \).

The software \texttt{HomotopyContinuation.jl} [7] is very user-friendly. We will show how to compute all critical points with version 2.3.1. We start by generating a random linear model:

\begin{verbatim}
@var x[1:d]
c = rand(n+1); c = c/sum(c)
p = [randn(d)'*x + c[i] for i = 1:n]
p = push!(p,1-sum(p))
\end{verbatim}

The array \( p \) contains \( n + 1 \) affine polynomials in the unknowns \( x \). Their constant terms are the positive reals in \( c \) that sum to 1. The polytope \( \Theta \) has dimension \( d \) since \( 0 \in \text{int}(\Theta) \). The next step is to construct the log-likelihood function and compute its derivatives. Using the logarithm function in \texttt{HomotopyContinuation.jl}, this can be done in two lines of code:

\begin{verbatim}
@var s[0:n]
L = sum([s[i]*log(p[i]) for i = 1:n+1])
F = System(differentiate(L,x), parameters = s)
\end{verbatim}

Here \( F \) represents the rational map

\[ F : \mathbb{C}^d \times \mathbb{C}^{n+1} \to \mathbb{C}^d, \quad (x, s) \mapsto \left( \frac{\partial L}{\partial x_1}, \ldots, \frac{\partial L}{\partial x_d} \right). \]

We choose a random complex data vector \( s^* \in \mathbb{C}^{n+1} \), and we solve the system \( F(x; s^*) = 0 \) as follows:

\begin{verbatim}
monodromy_result = monodromy_solve(F)
s_star = parameters(monodromy_result)
\end{verbatim}

This uses the \textit{monodromy method} for solving a generic instance of a parametrized family [16]. We stress that we do not turn rational functions into polynomials by clearing denominators. Working directly with the rational functions allows for cheaper evaluation of \( F \) and it avoids spurious solutions in the hyperplanes \( \{ p_i(x) = 0 \} \). Once we have the solutions for \( s^* \in \mathbb{C}^{n+1} \), we can find the solutions for any data vector \( (s^*)' \in \mathbb{R}^{n+1}_{>0} \) via a (straight line) \textit{coefficient parameter homotopy}. Here the vector \( s \) moves from \( s^* \) to \( (s^*)' \) along a straight line in \( \mathbb{C}^{n+1} \).

As the \textit{start parameter values} \( s^* \) move to the \textit{target parameter values} \( (s^*)' \), the solutions of \( F(x; s^*) = 0 \) move towards the solutions of \( F(x; (s^*)') = 0 \). We can track them numerically. For details, see [27,
Chapter 7. The coefficient parameter homotopy is implemented in the `solve` function. The following code solves \( F(x; (s^*)') = 0 \) for random \((s^*)' \in \mathbb{R}_{>0}^{n+1}\):

```plaintext
cstartsols = solutions(monomodromy_result)
s_star_prime = rand(length(s))
cp_result = solve(F, startsols; start_parameters = s_star, target_parameters = s_star_prime)
```

The solutions computed via monodromy are stored in `startsols`. These can be used as starting points in the coefficient parameter homotopy for solving any new instance \( F(x; (s^*)') = 0 \) of our equations. Hence, the monodromy computation happens only once for a given model.

Finally, we certify the solutions found by the coefficient parameter homotopy using the certification technique described recently in [8]. Each solution that has been certified is guaranteed to be an approximate solution, in a suitable sense, to our system of equations:

```plaintext
cert = certify(F, solutions(cp_result), s_star_prime)
```

In our discussion we described a workflow consisting of three steps: monodromy, coefficient parameter homotopy, and certification. These steps are easy to run, and they can be applied to any statistical model and hence to any system of scattering equations in physics. A nice feature of linear models, like CHY in Section 2, is that the method can solve the likelihood equations using real arithmetic only. This allows us to reduce the computation time.

We now explain the real arithmetic idea. In general, one uses complex start values \(s^*\) to avoid the discriminant locus of the family \( F(x; s) = 0 \). For linear models, this discriminant is an affine transformation of the entropic discriminant [26]. It is known from [21, Theorem 6.2] that the entropic discriminant is a sum of squares. This implies that the real locus of our discriminant has codimension \(\geq 2\). One also finds that this locus is disjoint from \(\mathbb{R}_{>0}^{n+1}\).

As the data vector \(s\) varies continuously in \(\mathbb{R}_{>0}^{n+1}\), the solutions to \( F(x; s) = 0 \) move in distinct bounded regions in \(\mathbb{R}^d\). Therefore, once we have solved \( F(x; s^*) = 0 \) for some \(s^* \in \mathbb{R}_{>0}^{n+1}\), we can solve \( F(x; (s^*)') = 0 \) for any \((s^*)' \in \mathbb{R}_{>0}^{n+1}\) via a straight line coefficient parameter homotopy that uses only real arithmetic. In particular, for computing the MLE, we only need to track one solution, namely that in the distinguished polytope \(\mathcal{O}\).

**Example 3** (scattering equations on \(\mathcal{M}_{0,m}\)). Section 2 addressed a linear model from physics [9; 11] with \(d = m - 3, n = m(m - 3)/2 - 1\) and ML degree \((m - 3)!\). Our computations for Table 1 used the workflow described above. The columns \(t_c\) and \(t_{cert}\) show the computation times (in seconds) for the coefficient parameter homotopy from \(s^* \in \mathbb{C}^{n+1}\) to \((s^*)' \in \mathbb{R}_{>0}^{n+1}\) and for the certification respectively. The column \(t_R\) shows the time for path tracking over the reals, from \(s^* \in \mathbb{R}_{>0}^{n+1}\) to \((s^*)' \in \mathbb{R}_{>0}^{n+1}\). In each run, all \((m - 3)!\) solutions were certified. The time for the monodromy step is not reported, as it is an off-line step which happens only once. For instance, for \(m = 12\), the off-line step takes about 14 minutes. All computations were run on a 16 GB MacBook Pro with an Intel Core i7 processor working at 2.6 GHz.

**Example 4** (random linear models). We examined random models for various \((n, d)\). Unlike in Section 2, the \(p_i(x)\) are now dense. The number of bounded regions equals \(\binom{n}{d}\). This is the ML degree; see
Table 2. Solving the likelihood equations for random linear models. Here \( \binom{n}{d} \) is the ML degree, \( t_C \) and \( t_R \) are the timings for solving, and \( t_{\text{cert}} \) is the timing for certifying the solutions.

<table>
<thead>
<tr>
<th>((n, d))</th>
<th>( \binom{n}{d} )</th>
<th>( t_C )</th>
<th>( t_R )</th>
<th>( t_{\text{cert}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>((12, 6))</td>
<td>924</td>
<td>0.25</td>
<td>0.09</td>
<td>0.15</td>
</tr>
<tr>
<td>((13, 6))</td>
<td>1716</td>
<td>0.46</td>
<td>0.13</td>
<td>0.27</td>
</tr>
<tr>
<td>((14, 7))</td>
<td>3432</td>
<td>1.34</td>
<td>0.44</td>
<td>0.87</td>
</tr>
<tr>
<td>((15, 7))</td>
<td>6435</td>
<td>2.12</td>
<td>0.87</td>
<td>1.46</td>
</tr>
<tr>
<td>((16, 8))</td>
<td>12870</td>
<td>5.06</td>
<td>2.00</td>
<td>2.91</td>
</tr>
<tr>
<td>((17, 8))</td>
<td>24310</td>
<td>6.25</td>
<td>3.60</td>
<td>7.25</td>
</tr>
</tbody>
</table>

[17, equation (8)]. Using the same computer as in Example 3, we obtained the results in Table 2, for various central binomial coefficients. Again, we do not report the timings for the off-line step, which happens once per pair \((n, d)\). All models in Table 2 were solved easily using the default settings in \texttt{HomotopyContinuation.jl}. Larger values of \((n, d)\) are more challenging. The straightforward approach we presented above ran into numerical difficulties. The monodromy loop sometimes failed to find a full set of starting solutions, and a few paths got lost in the coefficient parameter homotopy. Solving larger problems reliably will require a more clever approach or more conservative settings.

Examples 3 and 4 lead to the following conclusion. The special combinatorial structure of the CHY scattering equations allows us to solve large instances with a fairly naive method. Things are different for generic linear models. We encountered numerical issues for the default settings when the ML degree exceeds 20000. The same dichotomy occurs for the models studied in the next two sections. Low degree and sparsity render the equations from physics especially suitable for reliable and certified computations with \texttt{HomotopyContinuation.jl}.

4. Higher Grassmannians

Let \( \text{Gr}(k, m) \) denote the Grassmannian in its Plücker embedding in \( \mathbb{P}^{\binom{m}{k} - 1} \), with Plücker coordinates \( p_I \) indexed by increasing sequences \( I = (1 \leq i_1 < i_2 < \cdots < i_k \leq m) \). We write \( \text{Gr}(k, m)^o \) for the open part where all \( p_I \) are nonzero and \( X^o \) for its quotient modulo \( (\mathbb{C}^*)^m \). We represent each point in \( X^o \) by a \( k \times m \) matrix that has been normalized and contains \((k - 1)(m - k - 1) = \dim(X^o)\) unknowns. There are different conventions for setting this up. For \( k = 3 \), we place \( 2m - 8 \) unknowns \( x_1, \ldots, x_{m-4} \) and \( y_1, \ldots, y_{m-4} \) in the matrix as follows:

\[
\begin{bmatrix}
0 & 0 & 1 & 1 & 1 & 1 & \cdots & 1 \\
0 & -1 & 0 & 1 & x_1 & x_2 & x_3 & \cdots & x_{m-4} \\
1 & 0 & 0 & 1 & y_1 & y_2 & y_3 & \cdots & y_{m-4}
\end{bmatrix}.
\]  

(11)

This ensures that \( m \) special minors \( p_I \) are equal to 1. These are the minors indexed by

\[
I = 123, 124, \ldots, 12m, 134, 234.
\]

(12)
Table 3. Computation times for solving the CEGM scattering equations.

<table>
<thead>
<tr>
<th>$m$</th>
<th>$n + 1$</th>
<th>ML degree</th>
<th>$t_C$</th>
<th>$t_{\text{cert}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>14</td>
<td>26</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>7</td>
<td>28</td>
<td>1272</td>
<td>0.35</td>
<td>0.19</td>
</tr>
<tr>
<td>8</td>
<td>48</td>
<td>188112</td>
<td>70.03</td>
<td>47.71</td>
</tr>
</tbody>
</table>

The following result is known in the literature on scattering amplitudes; see [5, Section 7.1; 12, Section 3; 11, Appendix C]. Our computations furnish an independent verification.

**Proposition 5.** The ML degree of the models $X^o$ for $m = 6, 7, 8$ equals 26, 1272 and 188112.

**Sketch of proof.** The certification with HomotopyContinuation.jl furnishes a solid proof of the lower bound. The proof is an identity in interval arithmetic [8]. The upper bound requires more work. We can either use the degenerations known as soft limits [12], or Thomas Lam’s approach (mentioned in [12, Section 1]) that rests on finite fields and the Weil conjectures, or the trace test method in numerical algebraic geometry. It would be desirable to find a general formula and theoretical understanding for the Euler characteristic of $X^o = \text{Gr}(k, m)^o$. □

In the development of algebraic statistics there was an earlier attempt to view the Grassmannian $\text{Gr}(k, m)$ as a discrete statistical model. It has dimension $k(m - k)$, it has $n + 1 = \binom{m}{k}$ states, and the Plücker coordinates are the probabilities. We refer to [17, Section 5] where the numbers 4 and 22 were reported for the ML degrees of $\text{Gr}(2, 4)$ and $\text{Gr}(2, 5)$. That model is different from the one studied here, where the dimension is $(k - 1)(m - k - 1)$, the number of states is $n + 1 = \binom{m}{k} - m$, and $\text{Gr}(2, m)$ has ML degree $(m - 3)!$. In light of the ubiquity and importance of the moduli space $\mathcal{M}_{0, m}$, we have concluded that the physical model $X^o$ above is the better way to think about the Grassmannian in the setting of algebraic statistics.

In what follows we work in the set-up for $k = 3$ as in [11; 12]. The task is to compute the set $\text{Crit}(L)$ of critical points of the scattering potential $L = \sum_I s_I \log(p_I)$. We assign positive reals to the $m^3 - m$ Mandelstam invariants $s_I$ where $I$ is any triple not listed in (12). The $m$ remaining Mandelstam invariants $s_I$ from (12) are determined from the kinematic relations

$$\sum_{j,k} s_{ijk} = 0 \quad \text{for } i = 1, \ldots, m.$$  

Here $(s_{ijk})$ is a symmetric tensor with $s_{ijk} = 0$ unless $i, j, k$ are distinct; see [11, equation (1.6)]. Rewriting the kinematic equations, we obtain formulas that are analogous to (10). However, the $m$ Mandelstam invariants $s_I$ from (12) do not matter for us, since $\log(p_I) = 0$, so they do not appear in the scattering potential $L$. For the other $n + 1$ indices $I$, the polynomials $p_I$ are bilinear in the unknowns $x_i, y_j$. In conclusion, our task is to solve a system of $2m - 8$ rational function equations in $2m - 8$ unknowns, namely $\partial L/\partial x_i = \partial L/\partial y_i = 0$ for $i = 1, \ldots, m - 4$.

We use the techniques from Section 3 to solve these equations for $m = 6, 7, 8$. The results are reported in Table 3 using the same notation as in the previous sections. For $m = 6$, we confirmed that all 26
solutions are real (see [11, Appendix C]). In the case \( m = 7 \), all 1272 solutions are computed in a fraction of a second. For concreteness, let us consider the data

\[
\begin{align*}
  s_{135} &= 45, & s_{235} &= 597, & s_{145} &= 473, & s_{245} &= 745, & s_{345} &= 29, & s_{136} &= 296, & s_{236} &= 503, \\
  s_{146} &= 725, & s_{246} &= 402, & s_{346} &= 132, & s_{156} &= 557, & s_{256} &= 649, & s_{356} &= 461, & s_{456} &= 246, \\
  s_{137} &= 636, & s_{237} &= 662, & s_{147} &= 37, & s_{247} &= 945, & s_{347} &= 87, & s_{157} &= 613, & s_{257} &= 819, \\
  s_{357} &= 889, & s_{457} &= 473, & s_{167} &= 665, & s_{267} &= 57, & s_{367} &= 340, & s_{467} &= 621, & s_{567} &= 562. 
\end{align*}
\]

These are the \( n + 1 = 28 \) Mandelstam invariants not in (12). For these data, we found 1272 solutions in 0.35 seconds, and we certified them in 0.19 seconds. Precisely 1210 of the solutions are real. To the best of our knowledge, no complete set of solutions to the scattering equations for \( m = 7 \) with general \( s_{ijk} \) has been reported in the literature so far.

Using HomotopyContinuation.jl we can also solve the likelihood equations for \( m = 8 \). This works in the order of minutes. But there are challenges for this large nonlinear model. While our earlier models showed the power of solve as a blackbox routine, here the situation is more delicate. It may happen that not all 188112 paths are tracked successfully in the coefficient parameter homotopy. For an example, fix the \( n + 1 = 48 \) Mandelstam invariants

\[
\begin{align*}
  s_{135} &= 632, & s_{235} &= 5076, & s_{145} &= 6368, & s_{245} &= 619, & s_{345} &= 8083, & s_{136} &= 5762, \\
  s_{236} &= 2099, & s_{146} &= 7767, & s_{246} &= 9208, & s_{346} &= 4889, & s_{156} &= 4412, & s_{256} &= 1024, \\
  s_{356} &= 5988, & s_{456} &= 924, & s_{137} &= 3430, & s_{237} &= 1017, & s_{147} &= 6235, & s_{247} &= 8010, \\
  s_{347} &= 9867, & s_{157} &= 2364, & s_{257} &= 9661, & s_{357} &= 7008, & s_{457} &= 4706, & s_{167} &= 2892, \\
  s_{267} &= 7670, & s_{367} &= 5769, & s_{467} &= 3188, & s_{567} &= 9696, & s_{138} &= 6264, & s_{238} &= 5878, \\
  s_{148} &= 1442, & s_{248} &= 1501, & s_{348} &= 4225, & s_{158} &= 579, & s_{258} &= 7524, & s_{358} &= 394, \\
  s_{458} &= 878, & s_{168} &= 7684, & s_{268} &= 5985, & s_{368} &= 9306, & s_{468} &= 8429, & s_{568} &= 648, \\
  s_{178} &= 697, & s_{278} &= 8414, & s_{378} &= 3151, & s_{478} &= 369, & s_{578} &= 3176, & s_{678} &= 8649. 
\end{align*}
\]

Starting with the output \texttt{startsols} from the off-line phase, the command \texttt{solve} finds 188109 distinct solutions in 70 seconds. The remaining three solutions are found by a few extra minutes of monodromy loops. The 188109 earlier solutions in \texttt{cp_result} serve as seeds:

\[
R = \text{monodromy\_solve}(F,\text{solutions(cp\_result)},s\_star\_prime)
\]

When running the off-line step for any new statistical model, it is very helpful to know the ML degree ahead of time. In our situation, with knowledge of Proposition 5, we can use the option \texttt{target\_solutions\_count = 188112} in the command \texttt{monodromy\_solve}, both for off-line and for on-line. This interrupts the monodromy loop when all solutions are found.

All in all, the on-line phase for a given vector of Mandelstam invariants takes no more than a few minutes. This includes the coefficient parameter homotopy, the on-line monodromy phase described above, and the certification step that furnishes the proof of correctness.
Remark 6. A notable feature of the $k = 2$ model in Section 2 is that all critical points of the log-likelihood function are real (Proposition 1). The same holds for $k = 3$, $m = 6$. For $k = 3$, $m \geq 7$, this is no longer true. In fact, in these cases, it is not known whether all solutions can be real. However, we observed experimentally that most of the solutions are real. For instance, for $k = 3$, $m = 8$, with the data in (14), precisely 149408 out of 188112 critical points are real. We do not know whether the 48 Mandelstam invariants $s_{ijk}$ can be chosen so that all 188112 complex solutions are real.

Remark 7. It would be interesting to investigate the likelihood geometry of positroid cells in $\text{Gr}(k, m)$, taken modulo the $(\mathbb{C}^*)^m$ action as in [6]. The software HomotopyContinuation.jl will be useful for finding the ML degrees of such models. For these computations, one replaces the matrices in (2) and (11) with the network parametrization of positroid cells [6; 29].

5. Low rank tensors

In this section we return to algebraic statistics. We apply our methods to the model of conditional independence for identically distributed random variables. This corresponds to symmetric tensors of low rank. We here study their ML degree and likelihood equations.

We consider symmetric tensors of format $m \times m \times \cdots \times m$ where the number of factors is $\ell$. Our model $X$ is the variety of symmetric tensors of rank $\leq k$, or equivalently, the $k$-th secant variety of the $\ell$-th Veronese embedding of $\mathbb{P}^{m-1}$. The dimension of the model equals $\dim(X) = km - 1$. We follow the set-up in (1), with the number of states $n + 1 = (m + \ell - 1)$. The state space is the set $\Omega_{m, \ell}$ of sequences $I = (i_1, i_2, \ldots, i_m) \in \mathbb{N}^m$ with $i_1 + i_2 + \cdots + i_m = \ell$.

The parameter space for our statistical model is the polytope $\Theta = (\Delta_{m-1})^k \times \Delta_{k-1}$, where the points $x_i$ in the $i$-th simplex $\Delta_{m-1}$ are distributions on the $i$-th random variable with $m$ states, and points $y$ in the simplex $\Delta_{k-1}$ specify the mixture parameters. Hence $x = (x_{i,j})$ is a nonnegative $k \times m$ matrix whose rows sum to 1, and $y$ is a nonnegative vector in $\mathbb{R}^k$ whose entries sum to 1. The probability of observing the state $I = (i_1, i_2, \ldots, i_m)$ equals

$$p_I(x, y) = \frac{\ell!}{i_1! i_2! \cdots i_m!} \sum_{j=1}^{k} y_{j}^{i_{j,1}} x_{j,2}^{i_{j,2}} \cdots x_{j,m}^{i_{j,m}}.$$  \hspace{1cm} (15)

The resulting natural parametrization of the conditional independence model is the map

$$\Theta \rightarrow \Delta_n, \quad (x, y) \mapsto (p_I(x, y))_{I \in \Omega_{m, \ell}}.$$  \hspace{1cm} (16)

This polynomial map is $k!$-to-1, due to label swapping, which amounts to permuting rows of $x$ and entries of $y$. The variety $X$ is the image in $\mathbb{P}^n$ of the complexification of the map (16).

Fix counts $s_I \in \mathbb{N}$ for $I \in \Omega_{m, \ell}$. Statisticians aim to maximize the log-likelihood function

$$L = \sum_{I \in \Omega_{m, \ell}} s_I \cdot \log(p_I(x, y)).$$

In this formula we incorporate the substitutions $x_{j,m} = 1 - \sum_{i=1}^{m-1} x_{j,i}$ and $y_k = 1 - \sum_{j=1}^{k-1} y_j$. 
\[ \begin{array}{c|cccccc} k & \ell = 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ \hline 2 & 12 & 39 & 82 & 158 & 268 & 427 & 634 \\ 3 & 1 & 1 & 111 & 645 & \geq 2121 & & \\
\end{array} \quad \begin{array}{c|cccc} k & \ell = 3 & 4 & 5 \\ \hline 2 & 121 & 1449 & 8727 \\ 3 & 646 & \geq 154492 \\ \end{array} \]

Table 4. Experimentally obtained ML degrees for symmetric tensors of rank \( k \) and order \( \ell \). The size is \( m = 2 \) (left) or \( m = 3 \) (right). Multiply by \( k! \) for the number of solutions to (17).

We shall compute all complex critical points of \( L \) by solving the likelihood equations

\[
\frac{\partial L}{\partial x_{i,j}} = \frac{\partial L}{\partial y_i} = 0 \quad \text{for } i = 1, 2, \ldots, k \text{ and } j = 1, 2, \ldots, m - 1.
\]

This is a system of \( km - 1 \) rational function equations in \( km - 1 \) unknowns. We denote the corresponding rational map by \( F(x, y, s) : \mathbb{C}^{km-1} \times \mathbb{C}^{n+1} \rightarrow \mathbb{C}^{km-1} \). The ML degree of the model \( X \) is the number of complex solutions to the system (17) divided by \( k! = 1 \cdot 2 \cdots k \). The maximum likelihood parameter \((\hat{x}, \hat{y})\) is one of the real solutions in the polytope \( \Theta \).

**Example 8.** Two small instances were studied in [17]. The case \( k = m = 2, \ell = 4 \) is featured in [17, Section 1] where a gambler tosses one of two biased coins four times, and \( X \subset \mathbb{P}^4 \) is the hypersurface given by a \( 3 \times 3 \) Hankel determinant. This has ML degree 12, so (17) has 24 solutions. A data vector \( s \) with three local maxima in \( \Delta_4 \) is listed in [17, Example 10]. In the table at the end of [17, Section 5] we learn that the model with \( k = m = 2, \ell = 5 \) has ML degree 39, so (17) has 78 solutions. At that time, over 15 years ago, symbolic computing with \texttt{Singular} was the method of choice, and finding 78 solutions was not that easy.

Using the numerical methods presented in Section 3, we solved the likelihood equations for \( m = 2, 3 \) and \( k = 2, 3 \). For various \( \ell \), we ran many iterations of the monodromy loop\(^1\) to count the number of solutions to (17). Dividing that number by \( k! \) gives an integer, and that is the ML degree for the model. A subsequent run of the certification feature in \texttt{HomotopyContinuation.jl} furnishes a proof that the proposed number is a lower bound on the ML degree. However, our method does not give a proof that this is also an upper bound.

**Remark 9** (a view from nonlinear algebra). Points in the ambient space \( \mathbb{P}^n \) for our models in Table 4 correspond to binary forms and ternary forms. For example, the entry 111 on the left is the ML degree for the \( 4 \times 4 \) Hankel determinant which defines binary sextics of rank 3. The entry 646 on the right concerns plane cubic curves of rank 3. This is the hypersurface in \( \mathbb{P}^9 \) defined by the Aronhold invariant, shown in equation (9.15) and Example 11.12 in [23]. We solved the likelihood equations (17) in the naive way, by computing all \( 646 \times 3! = 3876 \) zeros of the rational functions. Further computational progress is surely possible. But, just like in Example 4, this will require exploiting the special structure of the problem at hand.

\(^1\)The optional argument \texttt{group_action} of \texttt{monodromy_solve} can be used to speed up the computations.
A next goal is the likelihood geometry of $4 \times 4 \times 4$ tensors. For a geometer, these are cubic surfaces in $\mathbb{P}^3$, with parameters $m = 4$, $\ell = 3$, $n = 19$. In the book cover of [24], this means that DiaNA now juggles three dice, each labeled $A$, $C$, $G$, $T$. We studied this model for cubic surfaces of rank $k = 2$. Our computations suggest that the ML degree equals 6483.

We next present an explicit numerical example, for the model of plane cubics of rank 2.

Example 10 ($m = \ell = 3, k = 2, n = 9$). Consider the data vector $s \in \mathbb{N}^{10}$ with coordinates

\[
\begin{align*}
s_{300} &= 8263, & s_{210} &= 4935, & s_{201} &= 8990, & s_{120} &= 7238, & s_{111} &= 5034, \\
s_{102} &= 5106, & s_{030} &= 5181, & s_{021} &= 6843, & s_{012} &= 5282, & s_{003} &= 9501.
\end{align*}
\]

The log-likelihood function $L$ has 242 complex critical points, so there are 121 critical points in the secant variety $X \subset \mathbb{P}^9$. Precisely eight of them lie in the actual model $X \cap \Delta_0$. These come from 16 critical points in $\Theta = \Delta_2 \times \Delta_2 \times \Delta_1$. The maximum likelihood estimate equals

\[
\begin{align*}
\hat{p}_{300} &= 0.0661, & \hat{p}_{210} &= 0.1585, & \hat{p}_{201} &= 0.0937, & \hat{p}_{120} &= 0.1269, & \hat{p}_{111} &= 0.1542, \\
\hat{p}_{102} &= 0.0711, & \hat{p}_{030} &= 0.0340, & \hat{p}_{021} &= 0.0658, & \hat{p}_{012} &= 0.0883, & \hat{p}_{003} &= 0.1414.
\end{align*}
\]

The 121 critical points in $X$ are $3 \times 3 \times 3$ tensors of complex rank 2. Among these 121 tensors, 47 are real. We found that 20 have real rank 2, so each has two real preimages in $\mathbb{R}^5$. The other 27 have real rank 3. They come from complex conjugate pairs of parameters $(x, y)$.

We now offer some pertinent remarks on numerical algebraic geometry. Our object of interest is the rational map $F : \mathbb{C}^{km-1} \times \mathbb{C}^{n+1} \to \mathbb{C}^{km-1}$ defined by the gradient of $L$. To find all solutions of $F(x, y; s^*) = 0$ for general complex data $s^* \in \mathbb{C}^{n+1}$, it is necessary that the monodromy action on $F(x, y; s^*)^{-1}(0)$ is transitive. This happens if and only if the incidence variety $\{(x, y, s) : F(x, y, s) = 0\}$ in the total space $\mathbb{C}^{km-1} \times \mathbb{C}^{n+1}$ is irreducible [16, Section 2]. However, for our parametrized tensor models, this incidence variety is reducible.

Example 11 ($m = k = 2, \ell = 4$). For any $s \in \mathbb{C}^5$, we consider the solutions to the critical equations $F = (\partial L / \partial x_{11}, \partial L / \partial x_{21}, \partial L / \partial y_1) = 0$ in the open subset where the denominators are nonzero. The incidence variety $Y$ is the closure of this set in $\mathbb{C}^3 \times \mathbb{C}^5$. In a Gröbner basis approach, this would be computed by clearing denominators in $F$ and then saturating the denominators. To appreciate the complexity of this, note that the three numerators have degree 25, with 2025 terms, 2418 terms and 2439 terms respectively. This is why we do not clear denominators.

We see that $Y$ is reducible because all terms of $\partial L / \partial x_{11}$ are multiples of $y_1$. Points in the locus $\{y_1 = 0\}$ parametrize tensors of rank one. This gives an extraneous component of $Y$. Interestingly, $Y$ has dimension 6, because a rank 1 tensor arises from a line of parameter values, given by $x_{11} = x_{21}$ and $y_1$ arbitrary. The fibers of the map $Y \to \mathbb{C}^5$ contain a line and 24 isolated points that represent $24/2! = 12$ rank-2 tensors. These are the critical points we are interested in. The corresponding 5-dimensional component of $Y$ parametrizes the likelihood correspondence, i.e., the irreducible variety in $\mathbb{P}^4 \times \mathbb{P}^4$ from [18, Definition 1.5].
In summary, one drawback of our approach in this paper is the presence of extraneous components in the incidence variety. From a numerical point of view, this makes the monodromy procedure more challenging. The phenomenon of path jumping may bring us to other components, leading to the computation of spurious solutions. For computing the ML degree of our tensor models, we are only interested in critical points in the regular locus of $X$. These are tensors of complex rank exactly $k$. They live on a component of $Y = \{F(x, y, s) = 0\} \subset \mathbb{C}^{km-1} \times \mathbb{C}^{n+1}$, called the dominant component in [16, Remark 2.2]. We can compute all solutions on that component by making sure that our seed lies on it.

6. Positive models and their amplitudes

The physical theory of scattering amplitudes is concerned with evaluating certain integrals of rational functions. In our statistical setting, these correspond to marginal likelihood integrals

$$
\int_\Theta p_0(x)^{s_0} p_1(x)^{s_1} \cdots p_n(x)^{s_n} \mu(x) \, dx.
$$

Such integrals arise in Bayesian statistics. In that paradigm one integrates the likelihood function over the parameter space $\Theta$ where the kernel is given by a measure $\mu(x)$, known as the prior belief. In general, it is a difficult problem to evaluate the integral (18) exactly and reliably. See [22] for an approach in the context of conditional independence as in Section 5.

It is a classical theme in mathematical statistics to connect Bayesian inference with the optimization problem (MLE) we explored in the previous sections. In this section we present new ideas for advancing that theme. These are inspired by positive geometries from Feynman diagrams and scattering amplitudes. We build on the theory of stringy canonical forms [5].

**Definition 12.** A discrete statistical model $X$ is called positive if it has a parametrization by positive rational functions $p_i(x)$ that sum to 1, where the parameter space is the orthant $\Theta = \mathbb{R}^d_{>0}$. A positive rational function is the ratio of two polynomials with positive coefficients.

Many familiar models in statistics are positive. To begin with, the probability simplex $\Delta_n$ of all distributions on $n + 1$ states is a positive model, thanks to the parametrization

$$
p : \mathbb{R}^n_{>0} \rightarrow \Delta_n, \quad x \mapsto \frac{1}{1 + x_1 + x_2 + \cdots + x_n}(1, x_1, x_2, \ldots, x_n).
$$

Next are the two families in [24, Section 1.2]. In a toric model, $p_i(x)$ is a monomial with a positive coefficient divided by the sum of these $n + 1$ monomials [1; 18]. Every linear model $X$ is a positive model, since $X \cap \Delta_n$ is a polytope whose vertices have nonnegative coordinates. For instance, a positive parametrization $y \mapsto p(y)$ for Example 2 is found by replacing

$$
x_1 = \frac{y_1}{1 + y_1 + y_2 + y_3}, \quad x_2 = \frac{y_1 + y_2}{1 + y_1 + y_2 + y_3} \quad \text{and} \quad x_3 = \frac{y_1 + y_2 + y_3}{1 + y_1 + y_2 + y_3}.
$$

Every model $X$ with ML degree one is a positive model, by the parametrization in [18, Corollary 3.12]. Mixtures of positive models are positive models, by using (19) for the mixture parameters. In particular, all discrete conditional independence models [28, Chapter 4] are positive models. In the setting of Section 5,
we use (19) to positively parametrize the factors in \( \Theta = (\Delta_{m-1})^k \times \Delta_{k-1} \), and we then compose this with the positive polynomials in (15).

Fix a positive model \( X \). We factor the numerator and denominator of each \( p_i(x) \) into positive polynomials, we write \( q_1(x), \ldots, q_e(x) \) for all the factors that occur, and we augment this list by \( x_1, \ldots, x_d \). We now rewrite the marginal likelihood integral (18) in the form seen in [5, (1.3)]. To this end, we set \( \varepsilon = 1 \) and \( \mu(x) = 1 \) for now. Then the integral (18) becomes

\[
\varepsilon^d \int_{\mathbb{R}^d} [x_1^{u_1} \cdots x_d^{u_d} q_1(x)^{-v_1} q_2(x)^{-v_2} \cdots q_e(x)^{-v_e}] \varepsilon \frac{dx_1}{x_1} \cdots \frac{dx_d}{x_d},
\]

where \( u_i, v_j \) are certain \( \mathbb{Z} \)-linear combinations of \( s_0, \ldots, s_n \). The log-likelihood function equals

\[
L = \sum_{i=1}^d u_i \log(x_i) - \sum_{j=1}^e v_j \log(q_j(x)).
\]

The set \( \text{Crit}(L) \subset \mathbb{C}^d \) of all critical points of \( L \) can be computed reliably using the methods in this paper. We define the \textit{toric Hessian} of \( L \) to be the symmetric \( d \times d \)-matrix \( H_L(x) \) whose entries are the rational functions \( \theta_i \theta_j L \), where \( \theta_i = x_i \partial_{x_i} \) is the \( i \)-th Euler operator.

In their recent work [5], Arkani-Hamed, He and Lam define the \textit{amplitude} of \( L \) to be the limit of the integral (21) as \( \varepsilon \) tends to zero. This corresponds to taking the field theory limit of integrals arising in string theory, where \( \varepsilon = \alpha' \) represents the inverse of the string tension. Given data \( s \) such that all \( v_j \) are positive, they consider the polytope \( P = \sum_{j=1}^e v_j \text{New}(q_j) \) and assume that \( u = (u_1, \ldots, u_d) \) lies in \( P \).

**Theorem 13.** The amplitude of a positive model \( X \) is a rational function in the data \( s_0, s_1, \ldots, s_n \). It equals the volume of the dual polytope \( (P - u)^* \) and it can be computed as

\[
\text{amplitude}(X) = (-1)^d \sum_{\xi \in \text{Crit}(L)} \det(H_L(\xi))^{-1}.
\]

**Sketch of proof.** This is our interpretation of the results in [5]. The amplitude depends only on the Newton polytopes \( \text{New}(q_j) \) and not on the specific positive coefficients of \( q_j(x) \). The hypothesis that \( u \) is in the interior of \( P \) ensures that (21) converges [5, Section 4.1]. The volume formula appears in [5, (2.5)] for \( s = 1 \) and in [5, (4.15)] for \( s \geq 2 \). The critical equations of \( L \) are the saddle point equations for the marginal likelihood integral (21) when \( \varepsilon \to \infty \). These equations appear in [5, Section 7.1]. They encode the pushforward formula for canonical forms of positive geometries. The toric Hessian is a convenient tool for writing the Jacobian of the system [5, (7.3)], and hence for computing the integral in [5, (7.5)]. □

**Example 14** \((d = n = 2)\). For the model \( X = \mathbb{P}^2 \), parametrized by (19), the integral (21) is

\[
\varepsilon^2 \int_0^\infty \int_0^\infty \left[ \frac{x_1^{s_1} x_2^{s_2}}{(1 + x_1 + x_2)^{s_0 + s_1 + s_2}} \right] \varepsilon \frac{dx_1}{x_1} \frac{dx_2}{x_2}.
\]

The log-likelihood function \( L = s_1 \log(x_1) + s_2 \log(x_2) - (s_0 + s_1 + s_2) \log(1 + x_1 + x_2) \) has only one critical point, namely \((\hat{x}_1, \hat{x}_2) = (1/s_0)(s_1, s_2)\). Substituting this into \( 1/\det(H_L(x)) \), we get

\[
\text{amplitude}(X) = \frac{1}{s_0 s_1} + \frac{1}{s_0 s_2} + \frac{1}{s_1 s_2} = \text{area}((P - (s_1, s_2))^*).
\]
Here, $P$ is the unit triangle $\text{conv}((0,0), (0,1), (1,0))$ scaled by the sample size $s_0 + s_1 + s_2$.

The special case $e = 1$ in Theorem 13 corresponds to the class of toric models in statistics; see [18, Section 3; 24, Section 1.2]. Any polynomial $q(x) = \sum_{j=0}^{n} c_j x^a_j$ with positive coefficients $c_j > 0$ defines a toric model $X$, by setting $p_j(x) = c_j x^a_j/q(x)$ for $j = 0, \ldots, n$. The ML degree of $X$ depends in subtle ways on the coefficients $c_j$. This was observed in [5, Section 7.1] and studied in detail in [1]. Both sources contain many open problems. For instance, it is conjectured in [5, Section 7.1] that the number $(m - 3)!$ from Section 2 is the minimal ML degree among all toric models supported on the associahedron. The diffeomorphism referred to in [5, Claim 4] is the familiar toric moment map [23, Theorem 8.24].

Up to a nonzero constant factor, the amplitude of the toric model $X$ equals the defining polynomial of the adjoint hypersurface of $P$, in the sense of Wachspress geometry [20]. Here we also need to divide by the product of the linear forms given by the facets of $P$. We learned this from unpublished lecture notes by Christian Gaetz which connect [3] with [20].

**Example 15** (measuring the dual of a square). The toric model for $q(x) = 1 + x_1 + x_2 + x_1 x_2$ is the independence model for two binary random variables, with data $s = (s_{ij})_{0 \leq i, j \leq 1}$. Here $n = 3$, $d = 2$, and $X$ is the Segre quadric in $\mathbb{P}^3$. The marginal likelihood integral in (21) is

$$
e^2 \int_{0}^{\infty} \int_{0}^{\infty} \left[ \frac{x_1^{s_{10} + s_{11}} x_2^{s_{01} + s_{11}}}{((1 + x_1)(1 + x_2))^{s_{00} + s_{10} + s_{11}}} \right]^e \frac{dx_1}{x_1} \frac{dx_2}{x_2}. $$

The limit for $\varepsilon \to 0$ is the amplitude. Its denominator is the product of the row and column sums of the contingency table $s$. The adjoint is the square of the sample size. Hence,

$$\text{amplitude}(X) = \frac{(s_{00} + s_{01} + s_{10} + s_{11})^2}{(s_{00} + s_{01})(s_{10} + s_{11})(s_{00} + s_{10})(s_{01} + s_{11})}.$$ 

Here $P$ is the square $[0,1]^2$ times the sample size. This is translated by $u = (s_{10} + s_{11}, s_{01} + s_{11})$. The normalized area of the dual quadrilateral $(P - u)^*$ equals the amplitude. Note that the assumption $u \in P$ from Theorem 13 is naturally satisfied in the statistical setting.

In earlier sections we showed that HomotopyContinuation.jl is fast for computing the critical set $\text{Crit}(L)$ of the log-likelihood function $L$. And it comes with certification. We use this to compute the sum (22) and hence to evaluate amplitudes for positive models. While the meaning of these amplitudes for Bayesian statistics is not clear yet, there is considerable interest in such computations among particle physicists. We next illustrate this for the CHY and CEGM models in Sections 2 and 4. We follow the set-up in [5, Section 6.2].

Let us begin with the $k = 2$ model in Section 3, with positive reparametrization as in (20).

**Example 16** ($k = 2$, $m = 6$). We compute the amplitude for the CHY model in Example 2. In terms of the positive parameters $y_1, y_2, y_3$ from (20), the log-likelihood function in (8) is

$$L = s_{23} \log(y_1) + s_{34} \log(y_2) + s_{45} \log(y_3) + s_{24} \log(y_1 + y_2) + s_{25} \log(y_1 + y_2 + y_3) + s_{35} \log(y_2 + y_3)$$

$$+ s_{36} \log(1 + y_2 + y_3) + s_{46} \log(1 + y_3) - \left( \sum_{(i,j)} s_{ij} \right) \log(1 + y_1 + y_2 + y_3).$$
The toric Hessian $H_L(y)$ is a symmetric 3 × 3-matrix whose entries are rational functions. The sum of the values of $-\det(H_L(y))^{-1}$ at the six critical points of $L$ is the amplitude

$$
\frac{1}{s_{12}s_{34}s_{56}} + \frac{1}{s_{12}s_{56}s_{123}} + \frac{1}{s_{23}s_{56}s_{123}} + \frac{1}{s_{23}s_{34}s_{234}} + \frac{1}{s_{34}s_{56}s_{234}} + \frac{1}{s_{16}s_{23}s_{45}} + \frac{1}{s_{12}s_{34}s_{45}} + \frac{1}{s_{12}s_{45}s_{123}} + \frac{1}{s_{12}s_{56}s_{345}} + \frac{1}{s_{16}s_{45}s_{345}} + \frac{1}{s_{16}s_{45}s_{345}} + \frac{1}{s_{23}s_{45}s_{123}}. (23)
$$

Here we abbreviate $s_{ijk} = s_{ij} + s_{ik} + s_{jk}$. The 14 terms in this sum correspond to the planar trivalent trees with six labeled leaves, and hence to the vertices of the associahedron in $\mathbb{R}^3$.

For a numerical example take the data in (7) and (10). The unique positive critical point $(\hat{y}_1, \hat{y}_2, \hat{y}_3) = (1.076\ldots, 1.202\ldots, 1.205\ldots)$ maps to the MLE in (9). The amplitude (23) equals

$$
\frac{16074421}{56770632000} = 0.00028314676856\ldots
$$

Using the abbreviation $y_{i,j} = \sum_{i \leq k \leq j} y_k$, the associated integral (21) has the form

$$
\varepsilon^3 \int_{\mathbb{R}^3_{>0}} \left[ \prod_{i,j} \frac{y_1^{s_{23}} y_2^{s_{34}} y_3^{s_{45}}}{y_{1,2}^{s_{24}} y_{1,3}^{s_{25}} y_{2,3}^{s_{35}} (1 + y_{2,3})^{-s_{36}} (1 + y_{1,3})^{-s_{46}} (1 + \sum_{i,j} s_{ij})} \right] \varepsilon^3 dy_1 dy_2 dy_3.
$$

The theory in [5] requires the hypotheses

$$
s_{23} \geq 0, \quad s_{34} \geq 0, \quad s_{45} \geq 0, \quad s_{24} \leq 0, \quad s_{25} \leq 0, \quad s_{35} \leq 0, \quad s_{36} \leq 0, \quad s_{46} \leq 0, \quad \sum_{i,j} s_{ij} \geq 0.
$$

If this holds then the leading order ($\varepsilon \to 0$) of the integral equals the volume of $(P - (s_{23}, s_{34}, s_{45}))^*$, where $P$ is the associahedron

$$
c_{24} \text{New}(y_{1,2}) + c_{25} \text{New}(y_{1,3}) + c_{35} \text{New}(y_{2,3}) + c_{36} \text{New}(1 + y_{2,3}) + c_{46} \text{New}(1 + y_{1,3}) + \sum_{i,j} s_{ij} \text{New}(1 + y_{1,3}).
$$

Here $c_{ij} = -s_{ij}$. The hypothesis fails for (7), but summing over Crit($L$) always works.

We now reiterate the punchline from Section 2 for amplitudes: current off-the-shelf software from numerical algebraic geometry is highly efficient and reliable for computing amplitudes by evaluating the sum (22). For our computations we used HomotopyContinuation.jl [7; 8]. We carried this out for models with $k = 2$ and $k = 3$. If $v_1, \ldots, v_e > 0$ then the amplitude measures the volume of the dual polytope in Theorem 13.

For $k = 2$, our computations validate known formulas involving planar trees like (23). For $k = 3$, $m \leq 7$, Cachazo et al. [11] describe formulas in terms of rays of the positive tropical Grassmannian, but in general there is still plenty of room for further discovery.

**Example 17** ($k = 2$). We used the positive parametrization [5, (1.5)] of $\mathcal{M}_{0,m}^+$ to verify (22) for the CHY model. Fix integer values for the Mandelstam invariants such that the hypotheses on $u$ and $v$ in Theorem 13 are satisfied. We compute the volume of $(P - u)^*$ in two ways. First the exact rational number is obtained using Polymake.jl [19]. Secondly, summing over the computed critical points as in (22) gives a floating point approximation. The cases we checked are $m = 5, 6, \ldots, 10$. Using double
precision arithmetic, the numerical evaluation of (22) agrees with the volume up to at least 12 significant digits in all cases. Computing the Hessian determinant and summing over the 5040 solutions for \( m = 10 \) takes about 20 seconds. The computation time for finding these solutions appears in Table 1.

**Example 18** \((k = 3)\). For \( m = 7 \), we compute the amplitude of the CEGM model for the data in (13). Our code finds the numerical value \( 3.5930250842 \cdot 10^{-19} \). This equals

\[
338162975644111011121443846899947726824222223033915555493151512849490472904959110279
\]

\[
94116511751278934751720147702872529781445322700367349835265330333349828096628760994174234572501760000
\]

This rational number is computed with a formula from [11, Section 4] which was kindly shared with us by Nick Early. In our study of amplitudes for CEGM models, we used the positive parametrization obtained from (11) by recursively setting \( x_0 = y_0 = 1 \) and

\[
x_\ell = x_{\ell - 1} + z_\ell, \quad y_\ell = y_{\ell - 1} + z_\ell(1 + w_1 + \cdots + w_\ell), \quad \ell = 1, \ldots, m - 4.
\]

Since this parametrization augments the degree of the equations, it is better to first solve the scattering equations using the formulation (11) and then compute the \((z, w)\) coordinates of the solutions \((x, y)\) via (24). Computing the sum (22) over the 1272 solutions takes about 11 seconds. Like Example 16, this illustrates the validity of (22) when the assumptions on \( u, v \) in Theorem 13 are violated. For \( m = 8 \), we obtain the numerical approximation \( 1.3609103649662523 \cdot 10^{-34} \) for the amplitude of the CEGM model with data (14).

We conclude with a summary of what has been accomplished in this paper. A connection has been made between algebraic statistics and the study of scattering amplitudes in physics. Positive models play the role of positive geometries. We showed how to solve the likelihood equations with certified numerical methods, and how to use this for evaluating amplitudes. Our case study offers a new tool kit for statistics and physics, based on nonlinear algebra.

Here is what we did not do: we did not prove new theorems in pure mathematics. We did not achieve notable methodological progress in statistics or theoretical advances in physics. The contribution of this work lies in building a bridge. Others may now cross that bridge, and use our tool kit to gain insights on the numerous fascinating problems that remain open.

**Acknowledgements**

We are very grateful to Sascha Timme for his help with the software HomotopyContinuation.jl, and to Pieter Bomans and Taylor Brysiewicz for discussions. We thank Freddy Cachazo and Nick Early for patient tutoring and inspiring conversations. Many thanks also to Kathlén Kohn, Thomas Lam and two anonymous referees for their helpful suggestions and comments.

**References**


Received 2020-12-09. Revised 2021-04-29. Accepted 2021-05-18.

BERND STURMFELS: bernd@mis.mpg.de
*Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany*

and

Department of Mathematics, University of California, Berkeley, Berkeley, CA, United States

SIMON TELEN: simon.telen@mis.mpg.de
*Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany*
TORIC INVARIANT THEORY FOR MAXIMUM LIKELIHOOD ESTIMATION IN LOG-LINEAR MODELS

CARLOS AMÉNDOA, KATHLÉN KOHN, PHILIPP REICHENBACH AND ANNA SEIGAL

We establish connections between invariant theory and maximum likelihood estimation for discrete statistical models. We show that norm minimization over a torus orbit is equivalent to maximum likelihood estimation in log-linear models. We use notions of stability under a torus action to characterize the existence of the maximum likelihood estimate, and discuss connections to scaling algorithms.

1. Introduction

Fruitful, sometimes unexpected, connections between algebra and statistics are constantly being discovered in the field of algebraic statistics. In this paper we unveil a connection between toric invariant theory and maximum likelihood estimation for log-linear models. Log-linear models are widespread in statistics and play a fundamental role in categorical data analysis, with a wide range of applications [5]. They consist of discrete probability distributions whose coordinatewise logarithm lies in a fixed linear space and include, for example, independence models and discrete graphical models [21]. There is a long history of the study of log-linear models in statistics, with an emphasis on understanding their maximum likelihood inference [13]. This concerns the existence of the maximum likelihood estimate (MLE), which maximizes the likelihood function given sample data, and statistical procedures for its computation.

Log-linear models play a prominent role in algebraic statistics [27]. The key link to algebra is that the Zariski closure of a log-linear model is a toric variety, defined by a monomial parametrization. Toric varieties have a foundational place among the algebraic varieties studied in algebraic geometry [7].

In our companion work [1], we establish a connection between finding the MLE and norm minimization along an orbit under a group action. We focus there on the setting of Gaussian group models, centered multivariate Gaussian models whose concentration matrices are of the form $g^Tg$, where $g$ lies in a group. In this paper, we study the connection between invariant theory and maximum likelihood estimation in the setting of discrete exponential families. We find remarkable similarities and differences between the discrete and Gaussian settings.

The paper is organized as follows. We introduce maximum likelihood estimation and our toric invariant theory setting in the expository Sections 2 and 3. Our main results are in Section 4. We give a characterization of MLE existence in terms of the existence of a vector of minimal norm in an orbit under a torus action (see Theorem 4.3), and an explicit way to compute the MLE from such a vector (see Theorem 4.7). We provide an alternative characterization in terms of null cones in Propositions 4.4


Keywords: maximum likelihood estimation, log-linear models, graphical models, torus actions, null cone, scaling algorithms.
and 4.5. We compare iterative proportional scaling (IPS), a classical method to find the MLE for log-linear models, with approaches to norm minimization in Section 5. We conclude the paper with a comparison with the multivariate Gaussian setting of [1] in Section 6 and outline a possible generalization for future research.

2. Maximum likelihood estimation

In this section we describe our statistical set-up: maximum likelihood estimation for discrete probability distributions. A distribution on \( m \) states is determined by its probability mass function \( p \), where \( p_j \) is the probability that the \( j \)-th state occurs. Such a probability mass function is a point in the \((m-1)\)-dimensional probability simplex:

\[
\Delta_{m-1} = \left\{ p \in \mathbb{R}^m \mid p_j \geq 0 \text{ for all } j \text{ and } \sum_{j=1}^m p_j = 1 \right\}.
\]

A statistical model \( \mathcal{M} \) of distributions with \( m \) states is a subset of \( \Delta_{m-1} \).

The data for a discrete distribution is a vector of counts \( u \in \mathbb{Z}_{\geq 0}^m \), where the coordinate \( u_j \) is the number of times that the \( j \)-th state occurs, and \( n = u_+ := \sum_{j=1}^m u_j \) is the total number of observations. The corresponding empirical distribution is \( \tilde{u} = \frac{1}{n} u \in \Delta_{m-1} \).

Maximum likelihood estimation in \( \mathcal{M} \) given data \( u \) finds a point in the model most likely to give rise to the observed data. That is, an MLE given \( u \) is a maximizer \( \hat{p} \) of the likelihood function over the model \( \mathcal{M} \).

The likelihood function is

\[
L_u(p) = p_1^{u_1} \cdots p_m^{u_m}.
\]

(1)

For example, if the model fills \( \Delta_{m-1} \), the likelihood is maximized uniquely at \( \hat{p} = \tilde{u} \).

An MLE given \( u \) is, equivalently, a point in \( \mathcal{M} \) that minimizes the Kullback-Leibler (KL) divergence to the empirical distribution \( \tilde{u} \). The KL divergence from \( q \in \mathbb{R}_{\geq 0}^m \) to \( p \in \mathbb{R}_{\geq 0}^m \) is

\[
\text{KL}(p \parallel q) = \sum_{j=1}^m p_j \log \frac{p_j}{q_j}.
\]

Although the KL divergence is not a metric, for \( p, q \in \Delta_{m-1} \) it satisfies \( \text{KL}(p \parallel q) \geq 0 \), and \( \text{KL}(p \parallel q) = 0 \) if and only if \( p = q \). The logarithm of the likelihood given \( u \) can be written, up to additive constant, as

\[
\ell_u(p) = -n \sum_{j=1}^m \tilde{u}_j \log \frac{\tilde{u}_j}{p_j} = -n \text{KL}(\tilde{u} \parallel p).
\]

We see that maximizing the log-likelihood is equivalent to minimizing the KL divergence to the empirical distribution.

3. Toric invariant theory

In this section we describe the invariant theory of a torus action that we will use. We begin by introducing notions of stability under a group action.
3A. Stability. Invariant theory studies actions of a group $G$ and notions of stability with respect to this action. In this article we work with linear actions on a complex vector space. Such a linear action assigns to a group element $g \in G$ an invertible matrix in $\text{GL}_m(\mathbb{C})$. The group element $g \in G$ acts on $\mathbb{C}^m$ by left multiplication with the matrix. The action of group element $g$ on vector $v$ is denoted by $g \cdot v$. For a vector $v \in \mathbb{C}^m$, we define the capacity to be

$$\text{cap}(v) := \inf_{g \in G} \|g \cdot v\|^2.$$  

Here, and throughout the paper, $\| \cdot \|$ denotes the Euclidean norm for vectors and the Frobenius norm for matrices. We now define the four notions of stability for such an action.

**Definition 3.1.** Let $v \in \mathbb{C}^m$. We denote the orbit of $v$ by $G \cdot v$, the orbit closure with respect to the Euclidean topology by $\overline{G \cdot v}$ and the stabilizer $\{g \in G : g \cdot v = v\}$ by $G_v$. We say $v$ is

(a) **unstable**, if $0 \in \overline{G \cdot v}$, i.e., $\text{cap}(v) = 0$;
(b) **semistable**, if $0 \not\in \overline{G \cdot v}$, i.e., $\text{cap}(v) > 0$;
(c) **polystable**, if $v \neq 0$ and $G \cdot v$ is closed;
(d) **stable**, if $v$ is polystable and $G_v$ is finite.

The set of unstable points is called the **null cone** of the group action.

3B. A torus action. We consider the $d$-dimensional complex torus, denoted $\text{GT}_d(\mathbb{C})$ or $\text{GT}_d$. We sometimes consider the real analogue, denoted $\text{GT}_d(\mathbb{R})$. We consider the action of $\text{GT}_d$ on a complex projective space $\mathbb{P}^{m-1}_\mathbb{C}$, encoded by a $d \times m$ matrix of integers $A = (a_{ij})$. The torus element $\lambda = (\lambda_1, \ldots, \lambda_d)$ acts on a point $v$ in $\mathbb{P}^{m-1}_\mathbb{C}$ by multiplication by the diagonal matrix

$$
\begin{bmatrix}
\lambda_1^{a_{11}} \lambda_2^{a_{21}} \cdots \lambda_d^{a_{d1}} \\
\lambda_1^{a_{12}} \lambda_2^{a_{22}} \cdots \lambda_d^{a_{d2}} \\
\vdots \\
\lambda_1^{a_{1m}} \lambda_2^{a_{2m}} \cdots \lambda_d^{a_{dm}}
\end{bmatrix},
$$

(2)
i.e., it acts on the coordinates of the point $v$ via $v_j \mapsto \lambda_1^{a_{1j}} \cdots \lambda_d^{a_{dj}} v_j$.

**Remark 3.2.** An action of a group $G$ on $\mathbb{C}^m$ induces an action on the polynomial ring $\mathbb{C}[x_1, \ldots, x_m]$ by $g \cdot f(x) := f(g^{-1} \cdot x)$, where $x = (x_1, \ldots, x_m)^T$. For the action of the torus $\text{GT}_d$ given by matrix $A$, the map on indeterminates is $x_j \mapsto \lambda_1^{-a_{1j}} \cdots \lambda_d^{-a_{dj}} x_j$.

A **linearization** of the action of $\text{GT}_d$ on $\mathbb{P}^{m-1}$ is a corresponding action on the underlying $m$-dimensional vector space $\mathbb{C}^m$. It is given by a character of the torus, $b \in \mathbb{Z}^d$. For the linearization given by matrix $A \in \mathbb{Z}^{d \times m}$ and vector $b \in \mathbb{Z}^d$, the torus element $\lambda$ acts on the vector $v$ in $\mathbb{C}^m$ via

$$v_j \mapsto \lambda_1^{a_{1j} - b_1} \cdots \lambda_d^{a_{dj} - b_d} v_j.$$  

(3)

**Remark 3.3.** The name linearization comes from the setting of an algebraic group acting on a complex variety $X$, as follows. We fix a line bundle over $X$, i.e., a map $p : L \to X$, with certain properties, whose
fibers are copies of \( \mathbb{C} \). Given a group action on \( X \), a linearization is an action on \( L \) that agrees with the original action under projection under \( p \), and that is a linear action on each fiber [10, Chapter 7]. For example, the following projection map is a line bundle,

\[
p : \{(x, v) \in \mathbb{P}_\mathbb{C}^{m-1} \times \mathbb{C}^m \mid v \in \ell_x \} \to \mathbb{P}_\mathbb{C}^{m-1},
\]

where the fiber over each \( x \in \mathbb{P}_\mathbb{C}^{m-1} \) corresponds to the line \( \ell_x \) in \( \mathbb{C}^m \) that the point represents. That way, a linearization lifts an action on \( \mathbb{P}_\mathbb{C}^{m-1} \) to an action on \( \mathbb{C}^m \).

We now consider the notions of stability in Definition 3.1 for this torus action. They specialize to give polyhedral conditions. The convex hull of the columns \( a_j \in \mathbb{Z}^d \) of the matrix \( A \) is the polytope

\[
P(A) := \text{conv}\{a_1, \ldots, a_m\} \subseteq \mathbb{R}^d.
\]

The set \( P(A) \) consists of vectors in \( \mathbb{R}^d \) of the form \( Au \) for some \( u \in \Delta_{m-1} \). We define subpolytopes that depend on an indexing set \( J \subseteq [m] \)

\[
P_J(A) := \text{conv}\{a_j \mid j \in J\}.
\]

For \( v \in \mathbb{C}^m \), let \( \text{supp}(v) := \{j \mid v_j \neq 0\} \subseteq [m] \). We abbreviate \( P_{\text{supp}(v)}(A) \) to \( P_v(A) \), i.e., we define

\[
P_v(A) := \text{conv}\{a_j \mid j \in \text{supp}(v)\}.
\]

For a polytope \( P \subseteq \mathbb{R}^d \), we denote its interior by \( \text{int}(P) \) and its relative interior by \( \text{relint}(P) \).

**Theorem 3.4 (Hilbert–Mumford criterion for a torus).** Let \( v \in \mathbb{C}^m \) and consider the action of the complex torus \( \text{GT}_d \) on \( \mathbb{C}^m \) given by matrix \( A \in \mathbb{Z}^{d \times m} \) with linearization \( b \in \mathbb{Z}^d \). We have

(a) \( v \) unstable \( \iff \) \( b \notin P_v(A) \),

(b) \( v \) semistable \( \iff \) \( b \in P_v(A) \),

(c) \( v \) polystable \( \iff \) \( b \in \text{relint}(P_v(A)) \),

(d) \( v \) stable \( \iff \) \( b \in \text{int}(P_v(A)) \).

We give an elementary proof of Theorem 3.4 in Appendix A. Alternative proofs can be found in [10, Theorem 9.2] or [29, Theorem 1.5.1].

**3C. The moment map.** We introduce the moment map and state the Kempf–Ness theorem, for a torus action. As before, \( \text{GT}_d \) denotes the \( d \)-dimensional complex torus, and we consider its action on \( \mathbb{C}^m \) via the matrix \( A \in \mathbb{Z}^{d \times m} \). We first consider the trivial linearization \( b = 0 \) and later a general linearization \( b \in \mathbb{Z}^d \).

Fix \( v \in \mathbb{C}^m \) and consider a torus element \( \lambda = (\lambda_1, \ldots, \lambda_d) \) in \( \text{GT}_d \). The \( j \)-th coordinate of the vector \( \lambda \cdot v \in \mathbb{C}^m \) is

\[
(\lambda \cdot v)_j = \lambda_1^{a_1 j} \cdots \lambda_d^{a_d j} v_j.
\]
Next, we consider the map that sends $\lambda$ to the squared norm of $\lambda \cdot v$:

$$\gamma_v : \mathbb{G}_T^d \to \mathbb{R}, \quad \lambda \mapsto \|\lambda \cdot v\|^2 = \sum_{j=1}^{m} |\lambda_1|^{2a_{1j}} \cdots |\lambda_d|^{2a_{dj}} |v_j|^2.$$  

The infimum of $\gamma_v$ over $\lambda \in \mathbb{G}_T^d$ is the capacity of $v$.

More generally, for an algebraic group $G$ acting linearly on a space $V$ we can consider the map $\gamma_v : G \to \mathbb{R}$ that sends $g \mapsto \|g \cdot v\|^2$, for fixed $v \in V$. The derivative is a map $D_I \gamma_v : T_I G \to \mathbb{R}$, where $T_I G$ is the tangent space to $G$ at $I$. The moment map $\mu$ assigns to $v \in V$ the derivative of the map $\gamma_v$ at $I \in G$.

For the group $\mathbb{G}_T^d$, the tangent space at $I$ is equal to $\mathbb{C}^d$, and the derivative is a map $\mathbb{C}^d \to \mathbb{R}$. Recall that the map $f : \mathbb{C} \to \mathbb{C}$, $z \mapsto |z|^2$ is not complex differentiable. We identify $\mathbb{C}$ with $\mathbb{R}^2$, writing $z = x + iy$. Then the differential of $f$ is given in terms of $x$ and $y$, and their tangent directions $\hat{x}$ and $\hat{y}$, as $\tilde{z} \mapsto 2x \hat{x} + 2y \hat{y}$. In particular, the differential at $z = 1$ is the map $\tilde{z} \mapsto 2\Re(\tilde{z})$, where $\Re(\cdot)$ denotes the real part of a complex scalar. Extending to the multivariate setting, we obtain the derivative map

$$D_I \gamma_v : \mathbb{C}^d \to \mathbb{R}, \quad \lambda \mapsto \sum_{j=1}^{d} \left( \sum_{i=1}^{m} 2a_{ij} |v_j|^2 \right) \Re(\tilde{\lambda}_i) = 2\Re\left( \sum_{i=1}^{d} (Av^{(2)})_i \tilde{\lambda}_i \right),$$

where $v^{(2)}$ is the vector with $j$-th coordinate $|v_j|^2$. We can identify $\mathbb{C}^d$ with the space of $\mathbb{R}$-linear functionals $\text{Hom}(\mathbb{C}^d, \mathbb{R})$, by associating to $u \in \mathbb{C}^d$ the map $w \mapsto \Re(\sum_{i=1}^{d} u_i w_i)$. Under this identification, the linear map $D_I \gamma_v$ corresponds to the vector $2Av^{(2)} \in \mathbb{C}^d$. Hence the moment map, for linearization $b = 0$, is

$$\mu : \mathbb{C}^m \to \mathbb{C}^d, \quad v \mapsto 2Av^{(2)}.$$  

For a general linearization $b \in \mathbb{Z}^d$, we replace the columns $a_j$ of $A$ by $a_j - b$. This replaces the vector $Av^{(2)}$ by $Av^{(2)} - b\|v\|^2$, where $\|v\|^2 = \sum_{j=1}^{m} |v_j|^2$. We obtain

$$\mu : \mathbb{C}^m \to \mathbb{C}^d, \quad v \mapsto 2(Av^{(2)} - \|v\|^2 b).$$

The Kempf–Ness theorem relates points of minimal norm in an orbit, or orbit closure, to the vanishing of the moment map. It was first proven in [19]. Nowadays, several statements are referred to as Kempf–Ness theorem; see [1, Section 2] for a summary. For our torus action, we obtain the following.

**Theorem 3.5** (Kempf–Ness theorem for a torus). Consider the torus action of $\mathbb{G}_T^d$ given by matrix $A \in \mathbb{Z}^{d \times m}$ with linearization $b \in \mathbb{Z}^d$. A vector is semistable (resp. polystable) if and only if there is a nonzero $v$ in its orbit closure (resp. orbit) with $Av^{(2)} = \|v\|^2 b$. This $v$ is a vector of minimal norm in the orbit closure (resp. orbit).

We give two proofs of Theorem 3.5 in Appendix B. The first proof is a translation from the original paper of Kempf and Ness [19]; the second proof uses Theorem 3.4.

**3D. The null cone.** The set of unstable points under a group action on a vector space is the null cone, see Definition 3.1. In many settings of interest, the null cone is a Zariski closed set, the vanishing locus of all homogeneous invariants of positive degree. It is a classical object of interest, studied by Hilbert [18].
The irreducible components of the null cone are the linear spaces. We describe it in terms of the standard basis vectors in \( \mathbb{C}^m \), denoted \( e_1, \ldots, e_m \). The linear space spanned by \( \{ e_j \mid j \in J \} \) is denoted \( \langle e_j \mid j \in J \rangle \).

A vector \( b \) in \( P(A) \) can be written as a convex combination of the \( m \) columns of \( A \). We consider the maximal subpolytope of \( P(A) \) that does not contain \( b \), as well as the minimal subpolytope of \( P(A) \) that contains \( b \). Both minimality and maximality are taken with respect to inclusion in the set \([m]\).

In Proposition 3.6, we see the connection between irreducible components of the null cone and maximal subpolytopes of \( P(A) \) not containing \( b \), see Figure 1. Then, in Proposition 3.7, we see that minimal subpolytopes containing \( b \) give set-theoretic defining equations for the null cone, see Figure 2.

**Proposition 3.6.** Consider the action of \( \text{GT}_d \) on \( \mathbb{C}^m \) given by matrix \( A \in \mathbb{Z}^{d \times m} \) with linearization \( b \in \mathbb{Z}^d \). The irreducible components of the null cone are the linear spaces \( \langle e_j \mid j \in J \rangle \), where \( P_J(A) \) is a maximal subpolytope of \( P(A) \) with \( b \notin P_J(A) \).

**Proof.** Assume that a point \( v \in \mathbb{C}^m \) lies in a linear space \( \langle e_j \mid j \in J \rangle \) where \( b \notin P_J(A) \). Then \( \text{supp}(v) \subseteq J \), hence \( b \notin P_v(A) \), and \( v \) is unstable by Theorem 3.4(a). Conversely, assume that \( v \in \mathbb{C}^m \) is not contained in any linear space \( \langle e_j \mid j \in J \rangle \) as in the statement. Since the \( P_J(A) \) are maximal with \( b \notin P_J(A) \), we have \( b \in P_v(A) \) and \( v \) is semistable.

**Proposition 3.7.** Consider the action of \( \text{GT}_d \) on \( \mathbb{C}^m \) given by matrix \( A \in \mathbb{Z}^{d \times m} \) with linearization \( b \in \mathbb{Z}^d \). A vector \( v \in \mathbb{C}^m \) is in the null cone if and only if all products \( \prod_{j \in J} v_j \) vanish, where \( J \subseteq [m] \) indexes a minimal subpolytope of \( P(A) \) containing \( b \).

**Proof.** Denote \( v_J := \prod_{j \in J} v_j \). If some \( v_J \) is nonzero, i.e., \( J \subseteq \text{supp}(v) \), then \( b \in P_J(A) \) implies \( b \in P_v(A) \), hence \( v \) is semistable by Theorem 3.4(b). Conversely, if \( v \) is semistable then \( b \in P_v(A) \). By minimality of the minimal subpolytopes \( P_J(A) \) containing \( b \) we have, for some \( J \) in the statement, the containment \( P_J(A) \subseteq P_v(A) \), i.e., \( J \subseteq \text{supp}(v) \), hence \( v_J \neq 0 \).
The null cone is defined by the vanishing of all homogeneous invariants of positive degree. The monomials from Proposition 3.7 give the square-free part of the generators of the null cone. We describe how to take powers of the indeterminates appearing in the monomials, in order to turn them into invariants.

Let \( J \subseteq [m] \) index a minimal subpolytope of \( P(A) \) containing \( b \). Then 0 can be written as a strictly positive convex combination of \( \{(a_j - b) \mid j \in J\} \). Since the entries of the matrix \( A \) and the vector \( b \) are integers, the convex combination is rational. Multiplying by the lowest common denominator gives a positive integer linear combination

\[
\sum_{j \in J} r_j (a_j - b) = 0, \quad r_j \in \mathbb{Z}_{>0}.
\]

The monomials \( \prod_{j \in J} x_j^{r_j} \) are invariants under the group action, since

\[
\lambda \cdot \left( \prod_{j \in J} x_j^{r_j} \right) = \prod_{j \in J} (\lambda^{-(a_j - b)} x_j)^{r_j} = \prod_{j \in J} x_j^{r_j} \cdot \lambda^{-\sum_{j \in J} r_j (a_j - b)} = \prod_{j \in J} x_j^{r_j},
\]

where the first equality follows from Remark 3.2 and the last equality follows from (4).

\section*{4. Main Results}

We begin this section by introducing maximum likelihood estimation in log-linear models, pointing out connections to torus actions. We relate stability under a torus action to maximum likelihood estimation for log-linear models in Section 4B, and describe how to compute the MLE from the moment map in Section 4C.

\subsection*{4A. Log-linear models} A log-linear model consists of distributions whose logarithms lie in a fixed linear space. The log-linear model corresponding to a matrix \( A \in \mathbb{Z}^{d \times m} \) is

\[
\mathcal{M}_A = \{ p \in \Delta_{m-1} \mid \log p \in \text{rowspan}(A) \}.
\]

The coordinatewise logarithm \( \log p \) applies to \( p \) with strictly positive entries, and we therefore have \( \mathcal{M}_A \subseteq \text{relint}(\Delta_{m-1}) \). A parametrization of the model \( \mathcal{M}_A \) is given by

\[
\phi^A : \mathbb{R}_{>0}^d \rightarrow \Delta_{m-1}, \quad \theta \mapsto \left( \frac{1}{Z(\theta)} \prod_{i=1}^d \theta_i^{a_{ij}} \right)_{1 \leq j \leq m}.
\]
where $Z$ is a normalization factor. We observe a first connection between the statistical model and a torus action: the map $\phi^A$ is, up to normalization, the action (3) of the real positive torus element $\theta$ on the all-ones vector $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^m$ with trivial linearization $b = 0$.

For the log-linear model $\mathcal{M}_A$, we assume that the vector $\mathbf{1}$ is in the row span of $A$; this is a common assumption for statistical, as well as algebraic, reasons. First, such log-linear models are equivalent to discrete exponential families [27, Section 6.2]. Second, the assumption means the uniform distribution $\frac{1}{m}\mathbf{1}$ is in the model. Moreover, consider the Zariski closure of $\mathcal{M}_A$ in $\mathbb{C}^m$, defined by the ideal

$$I_A = \{ p^v - p^w \mid v, w \in \mathbb{Z}_{\geq 0}^m \text{ such that } Av = Aw \} \quad (7)$$

in the ring $\mathbb{C}[p_1, \ldots, p_m]$, where $p^v := \prod_{j=1}^m p_j^{v_j}$ for $v \in \mathbb{Z}_{\geq 0}^m$. If $\mathbf{1} \in \text{rowspan}(A)$, this becomes a homogeneous ideal: indeed, if $r^T A = \mathbf{1}$ for some $r \in \mathbb{R}^d$, then multiplying $Av = Aw$ by this vector results in $\mathbf{1}v = \mathbf{1}w$.

**Example 4.1.** A probability distribution on two ternary random variables is a $3 \times 3$ matrix $p = (p_{ij})$ of nonnegative entries that sum to one. A distribution lies in the independence model if

$$p_{ij} = p_{i+} p_{+j} \quad \text{for all } 1 \leq i, j \leq 3,$$

where $p_{i+}$ is the sum of the $i$-th row of $p$, and $p_{+j}$ the sum of the $j$-th column. We view the independence model as a discrete exponential family, hence require that the entries of $p$ are strictly positive. The independence model on a pair of ternary random variables is the log-linear model $\mathcal{M}_A$ where

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \in \mathbb{Z}^{6 \times 9}.$$  

A distribution $p \in \mathcal{M}_A$ has nine states, i.e., $\mathcal{M}_A \subseteq \Delta_8$. We identify $\mathbb{R}^9$ with $\mathbb{R}^{3 \times 3}$ to view the nine state random variable as a pair of ternary random variables.

The action of $\text{GT}_6(\mathbb{R})$ on $\mathbb{R}^{3 \times 3}$ given by (2), is as follows. The torus element

$$\begin{bmatrix} v_1 & v_2 & v_3 & v_4 & v_5 \end{bmatrix} = (\lambda_1 \lambda_2 \lambda_3 \mu_1 \mu_2 \mu_3)$$

acts on a matrix $x \in \mathbb{R}^{3 \times 3}$ by multiplying the entry $x_{ij}$ by $\prod_{k=1}^6 v_k^{a_k}$ where $a$ is the column of $A$ with index $(i, j)$. This is the left-right action of $\text{GT}_3 \times \text{GT}_3$ on the space of $3 \times 3$ matrices; it sends $x_{ij} \mapsto \lambda_i \mu_j x_{ij}$. In particular, the orbit of the all-ones matrix consists of all rank one matrices with all entries nonzero. The intersection of this orbit with $\Delta_8$ gives the independence model on a pair of ternary variables.

We now consider maximum likelihood estimation for log-linear models. Recall that an observed vector of counts $u \in \mathbb{Z}_{\geq 0}^m$ defines an empirical distribution $\tilde{u} \in \Delta_{m-1}$. The vector $Au$ is a vector of sufficient statistics for the model $\mathcal{M}_A$. A maximum likelihood estimate is a point $q \in \mathcal{M}_A$ such that

$$Aq = Au,$$  

(8)
see, e.g., [11, Proposition 2.1.5] or [27, Corollary 7.3.9].

Since the model $\mathcal{M}_A$ is not closed, the MLE may not exist. Birch [3] was the first to rigorously study MLE existence in the context of multiway tables, where he observed that $u$ having all entries strictly positive is a sufficient condition for the MLE to exist and derived condition (8), sometimes known as Birch’s Theorem, see [23, Theorem 1.10]. The fact that some entries could still be 0 without affecting MLE existence was not fully understood until the work of Haberman, who gave the first characterization of MLE existence in her paper [17]. A modern necessary and sufficient condition is the following.

**Proposition 4.2** [27, Theorem 8.2.1]. Let $A \in \mathbb{Z}^{d \times m}$ be such that $\mathbb{1} \in \text{rowspan}(A)$ and let $\mathcal{M}_A$ be the corresponding log-linear model. Suppose we observe a vector of counts $u \in \mathbb{Z}^m_{\geq 0}$. Then the MLE given $u$ exists in $\mathcal{M}_A$ if and only if $A\bar{u}$ lies in the relative interior of the polytope $P(A)$.

In particular, we see that, indeed, if $u$ has all entries positive, the MLE always exists. However, if $u$ has some entries zero, the MLE may or may not exist.

Following [21, Section 4.2.3], we define the *extended log-linear model* $\overline{\mathcal{M}}_A$ to be the closure of $\mathcal{M}_A$ in the Euclidean topology on $\mathbb{R}^m$. The extended model allows for distributions that have some zero coordinates. The MLE always exists for the extended model, because it is compact and the likelihood function is continuous. If the MLE given $u$ does not exist in $\mathcal{M}_A$, we refer to the MLE given $u$ in the extended model $\overline{\mathcal{M}}_A$ as the *extended MLE* given $u$.

Since the likelihood function (1) is strictly concave for log-linear models, the MLE is unique if it exists, and similarly for the extended MLE.

### 4B. Relating stability to the MLE

We now describe connections between existence of the MLE for a log-linear model and stability under a torus action.

**Theorem 4.3.** Consider a vector of counts $u \in \mathbb{Z}^m_{\geq 0}$ with sample size $u_+ = n$, matrix $A \in \mathbb{Z}^{d \times m}$ with $\mathbb{1} \in \mathbb{C}^m$ in the rowspan, and vector $b = Au \in \mathbb{Z}^d$. The stability under the action of the complex torus $\mathbb{G}_T^d$ given by matrix $nA$ with linearization $b$ is related to ML estimation in $\mathcal{M}_A$ as follows:

(a) $\mathbb{1}$ unstable does not happen,
(b) $\mathbb{1}$ semistable $\iff$ extended MLE exists and is unique,
(c) $\mathbb{1}$ polystable $\iff$ MLE exists and is unique,
(d) $\mathbb{1}$ stable does not happen.

**Proof.** We refer to the conditions for the different notions of stability, coming from the Hilbert–Mumford criterion in Theorem 3.4. By Proposition 4.2, the MLE of $u$ exists if and only if $b$ lies in the relative interior of the polytope $P(nA)$, which is the condition for polystability in Theorem 3.4.

It remains to see that the cases of unstable and stable do not occur. The all-ones vector $\mathbb{1}$ can never be unstable with respect to the action in Theorem 4.3, because $b = Au$ is in the polytope $P(nA)$. Finally, the stable case also cannot arise, due to the assumption that the vector $\mathbb{1}$ lies in the row span of $A$, as follows. Writing $\mathbb{1}$ as a linear combination of the rows, i.e., $r^T A = \mathbb{1}$, we have that all vectors $a_j$ lie on the hyperplane $r_1 x_1 + \cdots + r_d x_d = 1$ and the polytope $P(A)$ has empty interior in $\mathbb{R}^d$. \qed
Figure 3. The face $F_b(A)$ of $P(A)$, for four choices of $b \in \mathbb{Z}^2$. For example, the leftmost picture displays the face $P_J(A)$ for $J = \{1, 2, 3, 4\}$. The vectors $a_i$ outside of the face are in the intersection of all the irreducible components of the null cone, see Proposition 4.5. For the corresponding components of the null cone, see Figure 1.

We remark that we could take any other vector of full support in Theorem 4.3. The theorem shows that MLE existence can be tested by checking polystability under the group action. We now give alternative characterizations that involve semistability, which has advantages over polystability. The semistability of $v$ can be checked by evaluating generators of the null cone at $v$. If all generators vanish then $v$ is unstable, otherwise it is semistable.

**Proposition 4.4.** For a vector of counts $u \in \mathbb{Z}^m_{\geq 0}$ with $u_+ = n$ and $A \in \mathbb{Z}^{d \times m}$, the MLE given $u$ exists if and only if there is some $b \in \mathbb{Z}^d$ of the form $b = Av$ for $v \in \mathbb{R}^m_{>0}$, such that $u$ is semistable for the torus action given by matrix $nA$ with linearization $b$.

**Proof.** We first assume that the MLE given $u$ exists. Since the vector $Au$ lies in the polytope $P_u(nA)$, the vector $u \in \mathbb{Z}^m_{\geq 0}$ is semistable for the action given by matrix $nA$ with linearization $Au$, by Theorem 3.4(b). Moreover, since $Au$ is in the relative interior of $P(nA)$, by Proposition 4.2, the vector $Au$ is of the form $A v$ for some $v \in \mathbb{R}^m_{>0}$.

Conversely, if the MLE given $u$ does not exist, then $Au$ lies on the boundary of the polytope $P(nA)$. Hence the whole polytope $P_u(nA)$ is contained in the boundary. Thus, for every $b \in \mathbb{Z}^d$ of the form $b = Av$ for $v \in \mathbb{R}^m_{\geq 0}$, we have $b \notin P_u(nA)$. Then $u$ is unstable under the torus action given by matrix $nA$ with linearization $b$, by Theorem 3.4(a). \qed

To test MLE existence with Proposition 4.4, we need to test null cone membership for multiple linearizations. We now discuss a different approach, involving one null cone. For a vector $b \in P(A)$ we denote by $F_b(A)$ the minimal face of the polytope $P(A)$ that contains $b$; see Figure 3.

**Proposition 4.5.** Consider a vector of counts $u \in \mathbb{Z}^m_{\geq 0}$ with $u_+ = n$ and $A \in \mathbb{Z}^{d \times m}$. The intersection of the irreducible components of the null cone for the torus action given by matrix $nA$ with linearization $b = Au$ is $\langle e_j \mid j \notin F_b(nA) \rangle$.

In particular, the MLE given $u$ exists in $M_A$ if and only if the intersection of the irreducible components of the null cone is $\{0\}$.

**Proof.** Define $A' := nA$ and consider the polytope $P(A')$, the convex hull of $a'_j := na_j$. We consider the null cone under the torus action given by matrix $A'$ with linearization $b$. A linear space $\langle e_j \mid j \in J \rangle$ is in the null cone if and only if $b \notin P_J(A')$, by Proposition 3.6.
We will show that \( e_j \) is contained in every irreducible component of the null cone if and only if \( a'_j \notin F_b(A') \). From this, the second paragraph of the statement follows because the MLE given \( u \) exists if and only if \( b = Au \) is in the relative interior of the polytope \( P(A') \), i.e., \( b \) does not lie on a proper face, and \( F_b(A') = P(A') \).

Consider an index \( j \) with \( a'_j \notin F_b(A') \). All possible expressions for \( b = Av \) for some \( v \geq 0 \) have \( v_j = 0 \), since \( F_b(A') \) is a face of \( P(A') \). Let \( J' \subseteq [m] \) be such that \( b \notin P_j(A') \), i.e., \( \langle e_j \mid j' \in J' \rangle \) is in the null cone. Taking \( J' = J \cup \{j\} \), the polytope \( P_{J'}(A') \) still does not contain \( b \). Hence, \( e_j \) lies in an irreducible component of the null cone that contains \( \langle e_j \mid j' \in J' \rangle \); so \( e_j \) lies in every irreducible component.

Conversely, consider an index \( j \) with \( a'_j \in F_b(A') \). We show that there exists an irreducible component of the null cone that does not contain \( e_j \). For each facet \( F \subseteq F_b(A') \), let \( v_F \) be a vector with \( \text{supp}(v_F) = \{k \mid a'_k \in F\} \), and take \( w_F \) with \( \text{supp}(w_F) = \text{supp}(v_F) \cup \{j\} \). The union of \( P_{w_F}(A') \) over facets \( F \subseteq F_b(A') \) is the whole polytope \( F_b(A') \), so \( b \in P_{w_F}(A') \) for some facet \( F \). By the minimality of \( F_b(A') \), we have \( b \notin P_{v_F}(A') \). Hence \( \langle e_k \mid a'_k \in F \rangle \) is contained in an irreducible component of the null cone but, since \( b \in P_{w_F}(A') \), the irreducible component does not contain \( e_j \).

Example 4.6. We illustrate Proposition 4.5 for the log-linear model \( \mathcal{M}_A \), where

\[
A = \begin{pmatrix}
p_{000} & p_{001} & p_{010} & p_{011} & p_{100} & p_{101} & p_{110} & p_{111} \\
p_{00+} & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
p_{01+} & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
p_{10+} & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
p_{11+} & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
p_{+00} & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
p_{+01} & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
p_{+10} & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
p_{+11} & 1 & 1 & 1 & 1 & 1 & 1 & 1 
\end{pmatrix}
\]

This is the graphical model on three binary random variables \( X_i \) given by the path graph 1—2—3, defined by the conditional independence relation \( X_1 \perp\!
\!
\!\perp X_3 \mid X_2 \). To identify the graphical model with \( \mathcal{M}_A \), we identify \( \mathbb{R}^8 \) with \( \mathbb{R}^{2 \times 2 \times 2} \) and label the columns of \( A \) by entries \( p_{ijk} \). The sufficient statistics of the model are the eight marginals \( p_{i+j} := p_{ij0} + p_{ij1} \) and \( p_{i+i} := p_{0ij} + p_{1ij} \), where \( (i, j) \in \{0, 1\}^2 \).

We compute the irreducible components of the null cone for the torus action given by matrix \( nA \) with linearization \( Au \), for various \( u \in \mathbb{Z}^8 \). The null cone is the zero locus of those monomials in the ring \( \mathbb{C}[x_1, \ldots, x_8] \) such that the supports of their exponent vectors index minimal subpolytopes of \( P(nA) \) that contain \( b \), as in Proposition 3.7.

Let \( u = [1 \ 0 \ 1 \ 0 \ 0 \ 1 \ 0 \ 1]^{\top} \). Then \( b = 1 \in \mathbb{R}^8 \) and the null cone is the vanishing locus of \( x_1x_3x_6x_8, x_1x_4x_7, x_2x_3x_5x_8, \) and \( x_2x_4x_5x_7 \). The irreducible components only intersect at \( \{0\} \), hence the MLE given \( u \) exists in \( \mathcal{M}_A \).

Let \( u = [1 \ 1 \ 1 \ 1 \ 0 \ 1 \ 1 \ 0]^{\top} \). Then \( b = [2 \ 2 \ 1 \ 1 \ 1 \ 2 \ 2 \ 1] \) and the null cone is the vanishing locus of \( x_1x_2x_3x_4x_6x_7, x_2^2x_3x_4x_5x_7, x_1x_2x_3^2x_6x_8, \) and \( x_2^2x_3^2x_5x_8 \). The irreducible components only intersect at \( \{0\} \), so the MLE given \( u \) exists in \( \mathcal{M}_A \).
When \( u = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 1 & 0 \end{bmatrix}^T \), the null cone is the vanishing locus of \( x_1x_3x_6 \) and \( x_2x_3x_5 \). The irreducible components intersect at \((e_4, e_7, e_8)\), hence the MLE given \( u \) does not exist in \( \mathcal{M}_A \). We can also see this from Proposition 4.2, as follows. The vector \( b = Au = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 1 & 0 \end{bmatrix}^T \) has some zero entries. Since \( P(A) \) only contains nonnegative points, \( b \) must lie on the boundary of \( P(A) \).

### 4C. The moment map gives the MLE

In Theorem 4.3, we compared two optimization problems: finding the MLE in a log-linear model, and norm minimization in an orbit under a related torus action. More specifically, we have seen that one problem attains its optimum if and only if the other one does. We now describe how these two optima are related via the moment map. For this, we consider two possible closures of the log-linear model \( \mathcal{M}_A \). The Euclidean closure of the model is the extended log-linear model \( \overline{\mathcal{M}}_A \). We also consider the smallest Zariski closed subset of \( \Delta_{m-1} \) containing \( \mathcal{M}_A \), denoted \( \overline{\mathcal{M}}_A^Z \), i.e., the Zariski closure of \( \mathcal{M}_A \) in \( \mathbb{R}^m \) intersected with the simplex \( \Delta_{m-1} \). Its defining ideal is given in (7). In the proof of the following theorem, we use the result that these two closures are equal [15, Theorem 3.2].

**Theorem 4.7.** Let \( u \in \mathbb{Z}^m_{\geq 0} \) be a vector of counts with \( u_+ = n \). Consider a matrix \( A \in \mathbb{Z}^{d \times m} \) with \( \mathbb{1} \in \text{rowspan}(A) \), and let \( b = Au \in \mathbb{Z}^d \). Consider the orbit closure of \( \mathbb{1} \) under the torus action of \( GT_d \) given by matrix \( nA \) with linearization \( b \). Let \( q \in \mathbb{C}^m \) be a point in the orbit closure where the moment map vanishes. Then the extended MLE given \( u \) for the model \( \mathcal{M}_A \) has \( j \)-th entry

\[
\frac{|q_j|^2}{\|q\|^2}. \tag{9}
\]

If \( \mathbb{1} \) is polystable, then this vector is the MLE.

**Proof.** At a point \( q \in \mathbb{C}^m \) where the moment map vanishes, we have \( nAq^{(2)} = \|q\|^2b \) by Theorem 3.5. Consider the vector \( q' \) with \( j \)-th entry as in (9). We show that \( q' \) is the extended MLE given \( u \) in \( \mathcal{M}_A \). Since \( q' \in \Delta_{m-1} \) and \( Aq' = A\frac{u}{n} \), it remains to show that \( q' \in \overline{\mathcal{M}}_A \). Using the equality \( \overline{\mathcal{M}}_A^Z = \overline{\mathcal{M}}_A \), it suffices to show that \( q' \) satisfies the equations in (7). Since \( q \) lies in the orbit closure of \( \mathbb{1} \), it satisfies the equations in (7), where \( A \) is replaced by the matrix with \((i, j)\) entry \( nai_j - bj \). That is, it satisfies \( q^v - q^w = 0 \) for all \( v, w \in \mathbb{Z}^m_{\geq 0} \) with \( nAv - bw_+ = nAw - bw_+ \). We show that this covers all pairs of vectors \( v, w \) with \( Av = Aw \). Indeed, if \( Av = Aw \) then \( v_+ = w_+ \), because \( \mathbb{1} \) is in the row span of \( A \). Hence \( q \in \overline{\mathcal{M}}_A \). Now we conclude that \( q' \) also satisfies the equations in (7), as follows. For each equation \( q^v = q^w \), we can take norms on both sides and square both sides. The equality \( v_+ = w_+ \) then shows that \( (q')^v = (q')^w \), since the denominator is present on both sides with equal power \( \|q\|^{2v_+} \).

In the polystable case, the vector \( q \) is in the orbit of \( \mathbb{1} \), hence has all entries positive. Thus the entries of \( q' \) are also positive, so \( q' \) is the MLE given \( u \) in \( \mathcal{M}_A \).

**Example 4.8.** Consider the log-linear model \( \mathcal{M}_A \), and vector of counts \( u \), where

\[
A = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 1 & 2 \end{bmatrix}, \quad u = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad b = Au = \begin{bmatrix} 5 \\ 3 \end{bmatrix}.
\]

This model is the plane conic \( x_2^2 = x_1x_3 \). The existence of the MLE given \( u \) in \( \mathcal{M}_A \) can be characterized by the torus action given by matrix \( nA \) with linearization \( b \), by Theorem 4.3, where \( n = u_+ = 4 \). Since \( b \)
is a positive combination of the columns of $A$, the vector $\mathbb{1}$ is polystable under this action and the MLE given $u$ exists. The MLE relates to a point of minimal 2-norm in the orbit of $\mathbb{1}$ under the torus action given by matrix $nA$ with linearization $b$, by Theorem 4.7. We show how to obtain the MLE from a point $q$ of minimal norm in the orbit of $\mathbb{1}$.

Since $q$ lies in the orbit of $\mathbb{1}$, its entries are $q_j = \lambda_1^{n_{a_{1j}} - b_1} \lambda_2^{n_{a_{2j}} - b_2}$, where $\lambda_i$ are nonzero complex numbers, i.e., $q = [\lambda^3 \lambda^{-1} \lambda^{-5}]^T$ where $\lambda = \lambda_1/\lambda_2$. Moreover, the moment map vanishes at $q$, so we have $nAq = \|q\|^2b$. Combining these, gives the condition $3v^2 - v - 5 = 0$, where $v = |\lambda|^8$, and we obtain that the MLE is

$$\hat{p} = \frac{1}{v^2 + v + 1} \begin{bmatrix} v^2 \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{31 + \sqrt{61}}{4\sqrt{61} + 52} \\ \frac{3 + 3\sqrt{61}}{4\sqrt{61} + 52} \\ \frac{9}{2\sqrt{61} + 26} \end{bmatrix} \sim \begin{bmatrix} 0.4662 \\ 0.3175 \\ 0.2162 \end{bmatrix}.$$ 

Theorem 4.7 shows that the MLE can be obtained from norm minimization on an orbit. It suggests the possibility of using algorithms from invariant theory to compute the MLE, as we describe in Section 5. In the next example, we motivate the study of these algorithms, returning to the independence model.

The independence model on a pair of discrete random variables is a log-linear model, as we saw in a special case in Example 4.1. In the following example, we apply Theorem 4.7 to obtain the MLE for the independence model from the point of minimal 2-norm in the orbit of $\mathbb{1} \otimes \mathbb{1}$ under a torus action.

Example 4.9. The independence model on a pair of random variables, each with $m$ states, is the log-linear model $\mathcal{M}_A$, where

$$A = \begin{bmatrix} I_m \otimes \mathbb{1} \\ \mathbb{1} \otimes I_m \end{bmatrix} \in \mathbb{Z}^{2m \times m^2}. \quad (10)$$

The first $m$ rows are $I_m \otimes \mathbb{1}$ and second $m$ rows are $\mathbb{1} \otimes I_m$, where $I_m$ is the $m \times m$ identity matrix, and $\mathbb{1}$ is the all-ones vector of length $m$. The Kronecker product $A_1 \otimes A_2$ of two matrices $A_k \in \mathbb{R}^{m_k \times n_k}$ is a matrix of size $m_1m_2 \times n_1n_2$. We index its rows by $(i_1, i_2)$ where $i_k$ ranges from 1 to $m_k$, and its columns by $(j_1, j_2)$, where $j$ ranges from 1 to $n_k$. The entry of $A_1 \otimes A_2$ at index $((i_1, i_2), (j_1, j_2))$ is $(A_1)_{i_1j_1} (A_2)_{i_2j_2}$.

See Example 4.1 for (10) in the case $m = 3$.

The model is the orbit of the all-ones matrix under the left-right action of $\text{GT}_m(\mathbb{R}) \times \text{GT}_m(\mathbb{R})$ on the space of $m \times m$ matrices, after restricting to positive entries that sum to one. Equivalently, the model is the orbit under the action in (2) of the torus $\text{GT}_2m(\mathbb{R})$ on $\mathbb{C}^{m \times m}$ given by the matrix $A$ in (10), again after restricting to positive entries that sum to one. Equivalently, the model consists of all rank one matrices with positive entries summing to one.

Given a data matrix $u \in \mathbb{Z}^{m \times m}$, we consider the orbit of the all-ones matrix $\mathbb{1} \otimes \mathbb{1} \in \mathbb{C}^{m \times m}$, under the action of $\text{GT}_2m(\mathbb{C})$ given by the matrix $nA$ with linearization $b$, where $A$ is (10), the sample size is
\[ n = u_{++}, \text{ and } b = Au \in \mathbb{Z}^{2m}. \] We seek a matrix in the orbit closure of \( \mathbb{1} \otimes \mathbb{1} \) at which the infimum norm is attained. By Kempf–Ness, such matrices are those at which the moment map vanishes. The vanishing of the moment map at \( q \in \mathbb{C}^{m \times m} \) gives, by Theorem 3.5,

\[
\begin{bmatrix}
|q_{1+}|^2 \\
\vdots \\
|q_{m+}|^2 \\
|q_{+1}|^2 \\
\vdots \\
|q_{+m}|^2
\end{bmatrix} = \|q\|^2 \begin{bmatrix}
u_{1+} \\
\vdots \\
u_{m+} \\
u_{+1} \\
\vdots \\
u_{+m}
\end{bmatrix}.
\]

We relate the MLE to the matrix \( p \) with entries \( p_{ij} = |q_{ij}|^2/\|q\|^2 \). The matrix \( p \) has nonnegative entries summing to one, and the same row and column sums as the empirical distribution \( \bar{u} \). It remains to show that \( p \) is in \( M_A \). For \( g = \left[ \lambda_1 \ldots \lambda_m \mu_1 \ldots \mu_m \right] \) in \( \text{GT}_{2m}(\mathbb{C}) \), we have

\[
(\cdot \cdot (\mathbb{1} \otimes \mathbb{1}))_{ij} = \lambda_i^n \mu_j^n \left( \prod_{k=1}^{m} \lambda_k^{-b_k} \right) \left( \prod_{l=1}^{m} \mu_l^{-b_{m+l}} \right).
\]

Hence \( p \) is a scalar multiple of the matrix with \((i, j)\) entry \( |\lambda_i|^{2n} |\mu_j|^{2n} \), and all such matrices have rank at most one. The latter is a closed condition, so any nonzero \( p \) obtained from the orbit closure of \( \mathbb{1} \otimes \mathbb{1} \) has rank one. Hence \( p \) lies in the closure of the independence model. If the orbit is closed, all entries are positive and it is the MLE. Otherwise, it is the extended MLE; see Theorem 4.3.

5. Scaling algorithms

We saw in Theorem 4.7 that the MLE in a log-linear model can be obtained from a point of minimal norm in an orbit. This connects two problems:

1. norm minimization in a complex torus orbit,
2. maximum likelihood estimation in a log-linear model.

Algorithms exist for both problems: the former can be approached with convex optimization methods, and the latter with an algorithm called iterative proportional scaling. In fact, both families of algorithms can be thought of as generalizations of two sides of a classical scaling algorithm due to Sinkhorn [25]. We explain these different generalizations, and how Theorem 4.7 completes the circle of algorithms, see Figure 4.

5A. Sinkhorn scaling. The classical scaling algorithm of Sinkhorn [25] scales a square matrix \( M \) with positive entries to a doubly stochastic matrix. That is, one finds diagonal matrices \( D_1 \) and \( D_2 \) such that \( D_1 M D_2 \) has all row sums and all column sums equal to one. The doubly stochastic matrix is obtained by alternately scaling the row and column marginals to one. A natural extension is to scale the matrix \( M \) to other fixed row sums and column sums [26]. Both versions of Sinkhorn scaling are depicted on the left of Figure 4. These algorithms involve the left-right action of a pair of tori \( \text{GT}_{m_1} \times \text{GT}_{m_2} \) on an \( m_1 \times m_2 \) matrix: the algorithms iterate between updates via the left torus and via the right torus.
Figure 4. Overview of different scaling algorithms. The historical progression is from left to right, starting with two different sides to Sinkhorn scaling.

Alternating scaling of the rows and columns of a matrix to fixed marginals is an instance of a scaling algorithm which, in the statistics literature, goes back to Deming and Stephan in [9]. For the independence model on two variables, the algorithm finds the MLE by alternating between scaling the row sums and the column sums to match the marginals of the empirical distribution. Given an observed matrix of counts \( u \in \mathbb{Z}_{\geq 0}^{m \times m} \) with sample size \( u_{++} = n \), and initialized at the uniform distribution, the algorithm has two steps. The \((i, j)\) entry for the two steps is

\[
\frac{1}{m^2} \mapsto \frac{1}{m} \cdot \frac{u_{i+}}{n} \mapsto \frac{u_{i+}}{n} \cdot \frac{u_{+j}}{n}.
\]

(11)

If all entries are positive, the output is the MLE to the independence model given \( u \), otherwise it is the extended MLE. This is the first example of iterative proportional scaling (IPS), which we describe in the next subsection.

5B. Iterative proportional scaling. In the previous section, we saw that alternating scaling of a matrix to fixed row and column sums gives the MLE to the independence model, when initialized at the uniform distribution. This is scaling under a product of tori \( \mathbb{G} T_m \times \mathbb{G} T_m \). We saw in Examples 4.1 and 4.9 how the independence model fits into the framework of log-linear models. In terms of the group action, this replaces the left-right action of a pair of tori \( \mathbb{G} T_m \times \mathbb{G} T_m \) with the action of a single torus \( \mathbb{G} T_{2m} \), acting via (2), where \( A \) is the matrix in (10).

In this section, we explain how Sinkhorn scaling extends to algorithms for maximum likelihood estimation for a general log-linear model, the bottom arrow of Figure 4.

Alternating between matching row and column sums can be extended to hierarchical models, which summarize data by contingency tables [12], by iteratively updating the various marginals. The approach was extended to more general log-linear models by Darroch and Ratcliff in [8].

For the log-linear model \( \mathcal{M}_A \), the MLE \( \hat{p} \) must satisfy the equation \( A \hat{p} = A \hat{u} \) from Birch’s theorem, where \( \hat{u} = \frac{u}{n} \) is the empirical distribution. IPS finds the extended MLE in \( \mathcal{M}_A \) given an empirical distribution \( \hat{u} \in \Delta_{m-1} \). We define IPS for a log-linear model given by a matrix \( A \in \mathbb{Z}_{\geq 0}^{m \times m} \) whose column
We modify \( A \) where \( \alpha \) (obtained by componentwise division of \( A \)) and the rows of \( A \) are all equal. Starting at the uniform distribution \( p^{(0)} = \frac{1}{m} \hat{1} \), we iterate until the \( k \)-th update \( p^{(k)} \) has sufficient statistics \( b^{(k)} = A p^{(k)} \) close to the target sufficient statistics \( b = A \hat{u} \), i.e., until (8) holds approximately. The update step is

\[
p_j^{(k+1)} = \prod_{i=1}^{d} \left( \frac{(A \hat{u})_i}{(A p^{(k)})_i} \right)^{a_{ij}/\alpha} p_j^{(k)}, \tag{12}
\]

where \( \alpha \) is the common column sum of \( A \); see [27, Algorithm 7.3.11]. This is the action of a torus element (obtained by componentwise division of \( A \hat{u} \) by \( A p^{(k)} \) and then componentwise exponentiation by \( 1/\alpha \)) on the vector \( p^{(k)} \). Here the torus action is given by the matrix \( A \) with linearization \( b = 0 \), see (3).

We can view maximum likelihood estimation as a norm minimization problem in a different way to Theorem 4.7, by interpreting IPS as minimizing KL divergence.

**Proposition 5.1.** Consider the log-linear model \( M_A \) where \( A \in \mathbb{Z}^{d \times m} \) has \( \hat{1} \) in its row span. Then there exists a matrix \( \tilde{A} \in \mathbb{Q}_{\geq 0}^{(d+1) \times m} \), with all column sums equal, such that \( M_A = M_{\tilde{A}} \), iterative proportional scaling in (12) with matrix \( \tilde{A} \) converges, and at each update step the KL divergence to the MLE decreases.

**Proof.** The proof of convergence of IPS is given in [8, Theorem 1] in the case where the entries of \( A \) are real and nonnegative with each column of \( A \) summing to one. There, the authors show that each step of IPS decreases the KL divergence \( KL(\hat{p} \| p^{(k)}) \) from the \( k \)-th iterate \( p^{(k)} \) to the MLE \( \hat{p} \). Since replacing \( A \) by \( \frac{1}{\alpha} A \) does not change the update step (12), the KL divergence also decreases for any matrix with real and nonnegative entries and all column sums equal.

We explain how this covers log-linear models defined by integer matrices with \( \hat{1} \) in the row span. We modify \( A \) without changing its row span, i.e., without changing the model \( M_A \). First, we add a sufficiently large positive integer to every entry of \( A \). For a general choice of integer, this does not change rowspan(\( A \)) since it adds a multiple of the vector \( \hat{1} \), which belongs to rowspan(\( A \)), to every row. Second, let \( \alpha \) be the maximum of the column sums \( a_{+,i} \). Add another row to the matrix, with entries \( \alpha - a_{+,i} \). The extra row is a linear combination of \( \hat{1} \) and the rows of \( A \), so the augmented matrix has the same row span as \( A \). The column sums of the augmented matrix \( \tilde{A} \) are all \( \alpha \).

**Remark 5.2.** We saw in Section 2 that \( \hat{p} = \arg\min_{p \in \mathcal{M}} KL(\hat{u} \| p) \). Here, we use KL divergence differently, measuring the KL divergence from iterate \( p^{(k)} \) to the MLE, \( KL(\hat{p} \| p^{(k)}) \).

Curiously, when IPS for log-linear models in (12) is applied to the independence model, we do not recover the classical IPS with Sinkhorn updates, because the column sums of the integer matrix \( A \) for the independence model in (10) are \( \alpha = 2 \), hence there is a square root in the update step. If, instead, we did IPS with the same matrix \( A \) but \( \alpha = 1 \) in (12) we would recover the two steps in (11) in a single step. This leads naturally to the question of which exponents \( \alpha \) achieve convergence, and how the choice of \( \alpha \) affects the convergence rate. This is the essence of an open problem in algebraic statistics, see [11, Section 7.3].

**5C. Norm minimization.** In this section, we explain how Sinkhorn scaling generalizes to norm minimization in invariant theory; see the top arrow of Figure 4.

The condition that a matrix can be scaled to a doubly stochastic matrix is dual to testing membership in the null cone under a group action, as follows. We consider pairs of diagonal matrices \( (D_1, D_2) \) of
determinant one that act on square matrices $M$ via $M \mapsto D_1 M D_2$. A matrix does not lie in the null cone under this action if and only if its orbit closure contains a matrix $M$ such that the matrix with $(i, j)$ entry $|m_{ij}|^2$ is a nonzero scalar multiple of a doubly stochastic matrix [14, Corollary 3.6]. This is an instance of Kempf–Ness. Norm minimization on the orbit of a square matrix either converges to zero, or to a nonzero matrix $M$ at which the moment map vanishes. The condition $\mu(M) = 0$ translates to the matrix with entries $|m_{ij}|^2$ being a scalar multiple of a doubly stochastic matrix. So we see that norm minimization scales to a nonzero multiple of a doubly stochastic matrix, if such a matrix exists in the orbit.

Norm minimization on an orbit can be considered for a wide range of groups and their actions. If a group $G$ can be expressed as a product of groups, then the alternating minimization idea from Sinkhorn’s algorithm generalizes. An important example of this is operator scaling, which solves the scaling problem for the left-right action of $\text{SL}_{m_1}(\mathbb{C}) \times \text{SL}_{m_2}(\mathbb{C})$ on the space of matrix tuples $(\mathbb{C}^{m_1 \times m_2})^n$. We discuss connections between operator scaling and statistics in our companion paper [1].

We consider the norm minimization problem for the action of the torus $\text{GT}_d(\mathbb{C})$ given by the matrix $nA$ with linearization $b$. We take $b = Au$, where $u$ is a vector of counts. By Kempf–Ness, Theorem 3.5, a vector is not in the null cone if and only if there is a nonzero vector $q$ in its orbit closure satisfying $Aq^{(2)} = \|q\|^2 b$. Hence, the problem of scaling a vector by acting with the torus to such a nonzero vector $q$ is dual to testing membership in the null cone under the torus action. This duality generalizes the discussion of doubly stochastic matrices above.

Since the vector $1$ is semistable, see Theorem 4.3, norm minimization converges to such a nonzero vector $q$. This is a convex optimization problem, as follows. Consider the action of $\text{GT}_d(\mathbb{C})$ given by matrix $A' = nA - b \otimes 1 \in \mathbb{Z}^{d \times m}$. For a torus element $(\lambda_1, \ldots, \lambda_d)$, the coordinate change $y_i := \log |\lambda_i|^2$ gives

$$\text{cap}(1) = \inf_{\lambda \in \text{GT}_d(\mathbb{C})} \|\lambda \cdot 1\|^2 = \inf_{\lambda \in \text{GT}_d(\mathbb{C})} \sum_{j=1}^m \prod_{i=1}^d |\lambda_i|^{2a'_{ij}} = \inf_{y \in \mathbb{R}^d} \sum_{j=1}^m \exp(y, a'_j).$$

Convexity then follows from the fact that each exponential function is convex and a sum of convex functions is convex. This minimization problem is known as geometric programming. Hence, common algorithms from the vast literature on convex optimization can be used to compute the capacity and find the MLE, e.g., interior point methods [6] or ellipsoid methods.

5D. Comparison of algorithms. We have seen in the previous two subsections that IPS and norm minimization are generalizations of Sinkhorn scaling that have emerged in different communities. Theorem 4.7 closes the cycle of algorithms from different communities, by showing how to obtain the (extended) MLE from a complex point of minimal norm in an orbit (or orbit closure); see Figure 4.

This bridges several differences between IPS and norm minimization. We summarize these differences here. First, when computing the capacity, the norm is minimized along a complex orbit closure (see Theorem 4.7), whereas every step in IPS involves real numbers. Secondly, the torus action given by matrix $nA$ that is used for computing the capacity is linearized by $b = Au$ (see Theorem 4.7), whereas IPS uses the action given by matrix $A$ with trivial linearization $b = 0$. Finally, the objective functions differ: the capacity is defined in terms of the Euclidean norm, which does not appear in IPS; instead
IPS minimizes KL divergence (see Proposition 5.1). In the following example we see that, while IPS decreases the KL divergence to the MLE, it may increase the Euclidean norm.

**Example 5.3.** Consider the matrix $A$ and vector of counts $u$ from Example 4.8. We can use IPS to compute the MLE in $\mathcal{M}_A$. We start at the uniform distribution $p^{(0)} = \frac{1}{3} \mathbb{1}$ and do update steps as in (12) with matrix $A$. These IPS steps converge by Proposition 5.1, since the matrix $A$ has real nonnegative entries and all column sums equal. We obtain

$$p^{(1)} = \left[ \frac{5}{12}, \frac{\sqrt{15}}{12}, \frac{1}{4} \right]^T.$$  

Note that the sum of the entries of $p^{(1)}$ is strictly less than one. The KL divergence from the uniform distribution to the MLE is $\text{KL}(\hat{p} \parallel p^{(0)}) \sim 0.047$, and after the first update it is $\text{KL}(\hat{p} \parallel p^{(1)}) \sim 0.016$. However, we have $\| p^{(1)} \|^2 = \frac{49}{144}$, which exceeds $\| p^{(0)} \|^2 = \frac{1}{3}$. 

### 6. Comparison with multivariate Gaussian models

We highlight similarities and differences with the multivariate Gaussian setting studied in [1]. For this, we compare results from this paper with the related results in [1]. We start by comparing the two statistical settings.

In the discrete setting, a model is given as a subset of the $(m-1)$-dimensional probability simplex $\Delta_{m-1} \subseteq \mathbb{R}^m$. In comparison, in the multivariate Gaussian setting, a model is given by a set of concentration matrices $\Psi$ in the cone of positive definite matrices. For a discrete model $\mathcal{M} \subseteq \Delta_{m-1}$ the data is a vector of counts $u \in \mathbb{Z}_{\geq 0}^m$ with $u_+ = n$ the total numbers of observations. The log-likelihood given $u$ at $p \in \mathcal{M}$ is $\sum_{j=1}^m u_j \log(p_j)$. In comparison, for a Gaussian model the data is summarized by the sample covariance matrix

$$S_Y = \frac{1}{n} \sum_{i=1}^n Y_i Y_i^T$$

and the log-likelihood given a tuple of samples $Y \in (\mathbb{R}^m)^n$ is $\log \det(\Psi) - \text{tr}(\Psi S_Y)$.

#### 6A. Stability.

In both papers we link notions of stability under group actions to maximum likelihood estimation of certain statistical models: for log-linear models in Theorem 4.3 and for Gaussian group models in [1, Section 3]. For log-linear models it is enough to consider actions of complex tori on $\mathbb{C}^m$. In contrast, in [1] we work with actions of (reductive algebraic) groups over $\mathbb{R}$ or $\mathbb{C}$, depending on whether we consider multivariate Gaussian distributions on $\mathbb{R}^m$ or $\mathbb{C}^m$. In the log-linear case we study stability of the all-ones vector, while in [1] we consider the stability notions for the observed tuple of samples.

For log-linear models, the log-likelihood is always bounded from above and the all-ones vector cannot be unstable. In contrast, in the Gaussian setting a tuple of samples is unstable if and only if the log-likelihood is not bounded from above. In both cases, semistability is equivalent to the log-likelihood being bounded from above and polystability is equivalent to the existence of an MLE. In the log-linear case, the MLE is unique if it exists, while for Gaussian group models there may be infinitely many. In
fact, the existence of a unique MLE for Gaussian group models relates to stability of a tuple of samples. In contrast, for log-linear models the all-ones vector is never stable.

6B. MLE computation. An important similarity between the log-linear and Gaussian settings is that norm minimizers under the respective group actions give an MLE (if it exists), see Theorem 4.7 and [1, Section 3]. For log-linear models, we compute real MLEs from complex torus orbits. However, for Gaussian group models, we compute the MLE over $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ from orbits over the same field $\mathbb{K}$. If the all-ones vector is semistable but not polystable, Theorem 4.7 yields the extended MLE. However, in the Gaussian case, if a tuple of samples $Y$ is semistable but not polystable there is no corresponding notion of extended MLE.

6C. Scaling. From the point of view of scaling algorithms, Sinkhorn’s algorithm is a common origin to both the log-linear and the Gaussian settings. As we described in Section 5, Sinkhorn scaling to target marginals is IPS for the independence model and this extends to IPS for a general log-linear model. On the Gaussian side, Sinkhorn scaling generalizes to alternating minimization procedures for computing MLEs of matrix normal models and tensor normal models. This algorithm is used both in invariant theory for norm minimization and in statistics to compute the MLE; see [1]. In contrast, for log-linear models the algorithms from invariant theory and statistics are not the same; see Figure 4.

We conclude this paper by pointing out that log-linear models and the Gaussian group models in [1] are examples of exponential transformation families. Hence, it is an interesting and natural question to ask whether there is a unifying concept that links invariant theory to maximum likelihood estimation for exponential families.

Appendix A: Hilbert–Mumford for a torus action

Almost all the results in this paper use the polyhedral characterization of stability under a torus action, given by the Hilbert–Mumford criterion (Theorem 3.4). In this appendix we present a proof of Theorem 3.4, in what we hope is an elementary and accessible style.

Remark A.1 (disregarding the linearization). The setting of Theorem 3.4 is a torus action of $\mathbb{G}_T d$ on $\mathbb{C}^m$ given by a matrix $A \in \mathbb{Z}^{d \times m}$ with linearization $b \in \mathbb{Z}^d$. For our statistical connections, it is important to separate the role of $A$ (which is fixed by the model) from that of $b$ (which depends on the data). However, we can remove the need for a linearization by altering the matrix $A$, as follows. The action of $\mathbb{G}_T d$ on $\mathbb{C}^m$ given by matrix $A \in \mathbb{Z}^{d \times m}$ with linearization $b \in \mathbb{Z}^d$ is the action given by matrix $A' \in \mathbb{Z}^{d \times m}$ with linearization $0 \in \mathbb{Z}^d$, where the matrix $A'$ has $j$-th column $a_j - b$, see (3). Hence we assume without loss of generality that the linearization is zero for proving Theorems 3.4 and 3.5. The effect of the linearization on the moment map is outlined in Section 3C.

The classical statement of the Hilbert–Mumford criterion, see [22, page 53], concerns one parameter subgroups. For the group $\mathbb{G}_T d$, a one parameter subgroup is given by a map

$$\sigma : \mathbb{C}^\times \to \mathbb{G}_T d, \quad \lambda \mapsto (\lambda^{\sigma_1}, \ldots, \lambda^{\sigma_d}),$$

(13)
We consider \( \lim_{\lambda \to 0} \sigma(\lambda) \cdot v \). The \( j \)-th entry of the limiting vector is zero for \( j \notin \text{supp}(v) \). For \( j \in \text{supp}(v) \), we have three possibilities:

\[
\left( \lim_{\lambda \to 0} \sigma(\lambda) \cdot v \right)_j = \begin{cases} 
0 & \text{if } \langle \sigma, a_j \rangle > 0, \\
 v_j & \text{if } \langle \sigma, a_j \rangle = 0, \\
 \infty & \text{if } \langle \sigma, a_j \rangle < 0.
\end{cases}
\] (14)

The classical statement of the Hilbert Mumford criterion for a torus action is as follows.

**Theorem A.2.** Consider the action of \( \mathbb{G}_m \) on \( \mathbb{C}^m \) via the matrix \( A \in \mathbb{Z}^{d \times m} \). Given a nonzero \( v \in \mathbb{C}^m \), with zero in its orbit closure, there exists a one-parameter subgroup of \( \mathbb{G}_m \) that scales \( v \) to zero in the limit.

We give a proof of Theorem A.2 following [28]. Other references for the statement of the theorem include [30, Proposition 5.3; 4, Lemma 3.4] for a torus, and [4, Theorem 4.1; 22, page 53; 30, Theorem 5.2] for a general reductive group.

**Proof of Theorem A.2.** We seek a one parameter subgroup \( \sigma : \mathbb{C}^* \to \mathbb{G}_m \) such that \( \lim_{\lambda \to 0} \sigma(\lambda) \cdot v \) is zero. From the form of a one parameter subgroup from (13) and the limiting behavior from (14), we see that this is equivalent to showing that

there exists \( \sigma \in \mathbb{Z}^d \) such that \( \langle \sigma, a_j \rangle > 0 \) for all \( j \in \text{supp}(v) \). (15)

Reordering the entries of \( v \), we can assume without loss of generality that \( \text{supp}(v) = [k] \) for some \( k \leq m \). Then the existence of such a \( \sigma \in \mathbb{Z}^d \) as in (15) is equivalent to the following statement about \( A \in \mathbb{Z}^{m \times d} \):

if \( t = (t_1, \ldots, t_k) \in \mathbb{R}_k \setminus \{0\} \) is such that \( t_1 a_{i1} + \cdots + t_k a_{ik} = 0 \) for all \( i \in [d] \)

then at least two entries of \( t \) are of opposite sign. (16)

The equivalence of (15) and (16) is [28, Lemma 1.1], and is an analogue of Gordan’s theorem [24, Section 7.8 Equation (31)] over the rational numbers. Thus it remains to prove (16).

Since \( v \) has 0 in its orbit closure, there exists a sequence \( \lambda^{(n)} = (\lambda_1^{(n)}, \ldots, \lambda_d^{(n)}) \in \mathbb{G}_m \) with \( \lambda^{(n)} \cdot v \to 0 \) as \( n \to \infty \). In coordinates,

\[
(\lambda_1^{(n)})^{a_{i1}} \cdots (\lambda_d^{(n)})^{a_{id}} \to 0 \quad \text{as } n \to \infty \quad \text{for all } j \in [k].
\] (17)

The hypothesis of (16) is that we have \( t \in \mathbb{R}_k \setminus \{0\} \) with \( t_1 a_{i1} + \cdots + t_k a_{ik} = 0 \) for all \( i \in [d] \). Without loss of generality, we can assume \( t_1 \) is nonzero and therefore

\[
-a_{i1} = \frac{t_2}{t_1} a_{i2} + \cdots + \frac{t_k}{t_1} a_{ik} \quad \text{for all } i \in [d],
\] (18)

which implies

\[
\prod_{i=1}^d (\lambda_i^{(n)})^{-a_{i1}} = \left( \prod_{i=1}^d (\lambda_i^{(n)})^{a_{i2}} \right)^{t_2/t_1} \cdots \left( \prod_{i=1}^d (\lambda_i^{(n)})^{a_{ik}} \right)^{t_k/t_1}.
\] (19)
If $t_j/t_1 \geq 0$ for all $j \in \{2, \ldots, k\}$, then the right-hand side of (19) either equals one (if all $t_j/t_1$ are zero) or tends to zero (if there exists some $j$ with $t_j/t_1 > 0$). But the left-hand side of (19) tends to infinity as $n \to \infty$, since it is the inverse of (17) for $j = 1$. Hence $t_j/t_1$ must be strictly negative for some $j$, i.e., $t_1$ and $t_j$ have opposite signs.

Theorem A.2 has the following generalization. It can be proved via a polyhedral geometry argument similar to the proof of Theorem A.2.

**Theorem A.3** [20, page 173]. Consider the action of $\text{GT}_d$ on $\mathbb{C}^m$ given by matrix $A \in \mathbb{Z}^{d \times m}$, and fix $v \in \mathbb{C}^m$. If $w \in \text{GT}_d \cdot v \setminus \text{GT}_d \cdot v$, then there exists a one-parameter subgroup that scales $v$ to an element of $\text{GT}_d \cdot w$ in the limit.

**Remark A.4.** Theorem A.3 has an analogue for a general reductive group $G$, but it only applies with the additional assumption that $G \cdot w$ is the unique closed orbit in $\overline{G \cdot v}$, see [30, Section 6.8].

Equipped with Theorems A.2 and A.3, we now prove Theorem 3.4. The proof mostly rests on Theorem A.2; we only use the stronger statement of Theorem A.3 for one direction of one of the four cases. **Proof of Theorem 3.4.** We first prove parts (a) and (b). If $v = 0$, then the polytope $P_v(A)$ is empty, hence $0 \notin P_v(A)$. If $v \neq 0$ is unstable, then there exists some $\sigma \in \mathbb{Z}^d$ such that $\langle \sigma, a_j \rangle > 0$ for all $j \in \text{ supp}(v)$, by combining Theorem A.2 with (14). Hence $\sigma$ defines a hyperplane

$$H_\sigma = \{x \in \mathbb{R}^d \mid \langle \sigma, x \rangle = 0\}$$

that separates zero from $P_v(A)$. By Farkas’ lemma, see [24, Section 7.3], such a hyperplane exists if and only if $0 \notin P_v(A)$.

For (c), we first prove that if 0 is on the boundary of $P_v(A)$, then $v$ is not polystable. We construct a point in the orbit closure of $v$, with support strictly smaller than that of $v$, and hence deduce that the orbit of $v$ is not closed. Since 0 lies on the boundary of $P_v(A)$, it is contained in a minimal face $F \subsetneq P_v(A)$. Since $A$ has integer entries, there is a hyperplane $H_\sigma$, with $\sigma \in \mathbb{Z}^d$, such that $F = H_\sigma \cap P_v(A)$. We choose the sign of $\sigma$ so that it has nonnegative inner product with all of $P_v(A)$. This ensures that the limit $w := \lim_{\lambda \to 0} \sigma(\lambda) \cdot v$ exists. The limit $w$ has $\text{ supp}(w) \subsetneq \text{ supp}(v)$, since $P_w(A) \subsetneq F$. Hence $w \in \text{GT}_d \cdot v \setminus \text{GT}_d \cdot v$, and $\text{GT}_d \cdot v$ is not closed.

For the converse direction of (c), we show that if $v$ is semistable but not polystable, then $0 \notin \text{ relint}(P_v(A))$. Let $w' \in \text{GT}_d \cdot v \setminus \text{GT}_d \cdot v$. There exists $\sigma \in \mathbb{Z}^d$ such that $w := \lim_{\lambda \to 0} \sigma(\lambda) \cdot v \in \text{GT}_d \cdot w'$, by Theorem A.3. We have $\text{ supp}(w) \subseteq \text{ supp}(v)$ and, moreover, $\text{ supp}(w) \subsetneq \text{ supp}(v)$ (otherwise $w = v$ by (14), a contradiction). Hence $\langle \sigma, a_j \rangle > 0$ for all $j \in \text{ supp}(v) \setminus \text{ supp}(w)$, while $\langle \sigma, a_j \rangle = 0$ for all $j \in \text{ supp}(w)$, by (14). We obtain $P_v(A) \nsubseteq H_\sigma$ and $P_w(A) = H_\sigma \cap P_v(A)$, i.e., $P_w(A)$ is a proper face of $P_v(A)$. We have $\text{GT}_d \cdot w = \text{GT}_d \cdot w' \subseteq \text{GT}_d \cdot v$ and so $w$ is semistable as $v$ is semistable. By (b), $0 \in P_w(A)$ and hence 0 is on the boundary of $P_v(A)$.

It remains to prove (d). We can assume $v$ is polystable, i.e., $0 \in \text{ relint}(P_v(A))$. We want to show that the dimension of the stabilizer $\{\lambda \in \text{GT}_d \mid \lambda \cdot v = v\}$ is zero if and only if the interior of $P_v(A)$ equals its relative interior (i.e., if and only if $P_v(A)$ is full-dimensional). Since $0 \in P_v(A)$, the equality of the interior and relative interior holds if and only if $U := \text{ span}\{a_j \mid j \in \text{ supp}(v)\}$ equals $\mathbb{R}^d$. If the stabilizer is
positive dimensional, it must contain a one-parameter subgroup, i.e., some $\sigma \in \mathbb{Z}^d \setminus \{0\}$ with $\sigma(\lambda) \cdot v = v$ for all $\lambda \in \mathbb{C}^\times$. Then $\langle \sigma, a_j \rangle = 0$ for all $j \in \text{supp}(v)$, so the orthogonal complement $U^\perp \subseteq \mathbb{R}^d$ contains a line, and $U \neq \mathbb{R}^d$. Conversely, if $U \neq \mathbb{R}^d$ then there exists $\sigma \in U^\perp$, which can be chosen to have integer entries, since $A$ has integer entries. The one parameter subgroup $\sigma(\lambda)$ then lies in the stabilizer, which is therefore positive-dimensional.

\[\square\]

Appendix B: Kempf–Ness for a torus action

Many of the results in this paper use the Kempf–Ness theorem for a torus action, as stated in Theorem 3.5. An elementary proof of the Kempf–Ness theorem for torus actions can be found in Sections 1 and 2 of the original paper of Kempf and Ness [19]. In this appendix, we translate between the setting of [19] and our setting, to explain how the results of [19] give Theorem 3.5. In addition, we present an alternative proof of Theorem 3.5, obtaining it as a consequence of Theorem 3.4.

As before, we consider the action of $G_T^d$ on $\mathbb{C}^m$ given by a matrix $A \in \mathbb{Z}^{d \times m}$. We can assume without loss of generality that the linearization is $b = 0$, by Remark A.1. We first describe how Theorem 3.5 follows from [19].

First proof of Theorem 3.5. Let $v \in \mathbb{C}^m$ and consider the action of $G_T^d$ via the matrix $A \in \mathbb{Z}^{d \times m}$. Recall that the moment map at $v$ is the derivative $D_{\gamma} v$, where

$$\gamma_v : G_T^d \to \mathbb{R}, \quad \lambda \mapsto \|\lambda \cdot v\|^2.$$ Identifying the space of $\mathbb{R}$-linear functionals $\text{Hom}(\mathbb{C}^d, \mathbb{R})$ with $\mathbb{C}^d$ gives the moment map

$$\mu : \mathbb{C}^m \to \mathbb{C}^d, \quad v \mapsto 2A v^{(2)},$$

where $v^{(2)}$ is the vector with $j$-th coordinate $|v_j|^2$.

We translate between our notation and that in [19]. Most importantly, our notion of “polystable” is called “stable” by Kempf and Ness, see [19, page 234]. Our function $\gamma_v$ is denoted $p_v$ in [19]. Moreover, “a critical point $g \in G$ of $p_v$” in [19] means the vanishing of the moment map at $g \cdot v$:

$$0 = D_g p_v = D_{\gamma} p_v = D_{\gamma} \gamma_{g \cdot v} = \mu(g \cdot v).$$

Thus the polystable part of Theorem 3.5 is a direct consequence of [19, Theorems 0.1(a) and 0.2] for $G = G_T^d$. The semistable part follows from the polystable part, using the fact that any orbit closure for the group $G_T^d$ contains a unique closed orbit, see, e.g., [20, Bemerkung 1 on page 96]. That unique closed orbit is not the zero orbit if and only if the vector is semistable.

\[\square\]

We end with an alternative proof of Theorem 3.5, which uses an important connection between the polytope $P_v(A)$ and the moment map $\mu$, see [2; 16]. The connection relates $P_v(A)$ to the image of the orbit $G_T^d \cdot v$ under the moment map. It is a first example of a moment polytope, an important object of study in invariant theory.
Second proof of Theorem 3.5. To link the moment map $\mu$ to the polytope $P_v(A)$, we need to rescale $\mu$ as follows:

$$\tilde{\mu} : \mathbb{C}^m \setminus \{0\} \to \mathbb{R}^d, \quad v \mapsto \frac{\mu(v)}{\|v\|^2} = \frac{Av^{(2)}}{\|v\|^2}. \quad \text{(20)}$$

The vector $v^{(2)}/\|v\|^2$ consists of nonnegative numbers that sum up to one. Hence $\tilde{\mu}(v)$ is a convex combination of the columns of $A$, therefore $\tilde{\mu}(v) \in P_v(A)$. The stronger statement

$$\text{relint } P_v(A) = \tilde{\mu}(\text{GT}_d \cdot v), \quad \text{(21)}$$

was proven independently in [2, Theorem 2; 16, Theorem 4], see Remark B.1. Given (21), the polystable case of Theorem 3.5 is a direct consequence of Theorem 3.4(c), as follows. Polystability is equivalent to $0 \in \text{relint } P_v(A)$ which, by (21), implies $0 \in \tilde{\mu}(\text{GT}_d \cdot v)$, i.e., that $0 = Aw^{(2)}/\|w\|^2$ for some $w \in \text{GT}_d \cdot v$. As in the first proof, the semistable case of Theorem 3.5 can be deduced from the polystable case, since any orbit closure $\text{GT}_d \cdot v$ contains a unique closed orbit. \hfill \Box

Remark B.1. The statements in [2, Theorem 2; 16, Theorem 4] discuss a moment map whose domain is a projective space $\mathbb{P}^{m-1}_C$, rather than the space $\mathbb{C}^m \setminus \{0\}$ in (20). However, the projective results still allow us to obtain (21), as follows. For nonzero $v \in \mathbb{C}^m$, let $[v]$ be the point in $\mathbb{P}^{m-1}_C$ that represents the line $Cv$. We consider the action of $\text{GT}_d$ on $\mathbb{P}^{m-1}_C$ given by the matrix $A \in \mathbb{Z}^{d \times m}$. We denote the $\text{GT}_d$-orbit of $[v]$ in $\mathbb{P}^{m-1}_C$ by $\text{GT}_d \cdot [v]$ (and denote the orbit closure by $\overline{\text{GT}_d \cdot [v]}$). The map $\tilde{\mu}$ factors through the projective space $\mathbb{P}^{m-1}_C$ via a unique map $\tilde{\mu} : \mathbb{P}^{m-1}_C \to \mathbb{R}^d$. In fact, $\tilde{\mu}$ is the moment map for the action of $\text{GT}_d$ on $\mathbb{P}^{m-1}_C$ given by $A \in \mathbb{Z}^{d \times m}$. This action fits the setting of [2; 16] because $\mathbb{P}^{m-1}_C$ is a compact Kähler manifold. The results [2, Theorem 2; 16, Theorem 4] give

$$P_v(A) = \tilde{\mu}(\overline{\text{GT}_d \cdot [v]}).$$

For (21), we require a statement for the orbit of $v$ rather than the orbit closure of $[v]$. The closure $\overline{\text{GT}_d \cdot [v]}$ is the disjoint union of finitely many $\text{GT}_d$ orbits. The orbits relate to $P_v(A)$ as follows. For each open face $F$ of $P_v(A)$ the set $\tilde{\mu}^{-1}(F) \cap \overline{\text{GT}_d \cdot [v]}$ is a single $\text{GT}_d$ orbit in $\mathbb{P}^{m-1}_C$, see [2, Theorem 2]. In particular, when $F = \text{relint } P_v(A)$ we obtain the orbit $\text{GT}_d \cdot [v]$. This yields (21), since $\tilde{\mu}(\text{GT}_d \cdot [v]) = \tilde{\mu}(\text{GT}_d \cdot v)$.

Acknowledgements

We are grateful to Peter Bürgisser, Mathias Drton, and Bernd Sturmfels for fruitful discussions. We are also grateful to an anonymous referee for helpful comments that improved the paper. Améndola was partially supported by the Deutsche Forschungsgemeinschaft (DFG) in the context of the Emmy Noether junior research group KR 4512/1-1. Kohn was partially supported by the Knut and Alice Wallenberg Foundation within their WASP (Wallenberg AI, Autonomous Systems and Software Program) AI/Math initiative. Research of Reichenbach is funded by the European Research Council (ERC) under the European’s Horizon 2020 research and innovation programme (grant agreement no. 787840).
References


Received 2021-01-29. Revised 2021-07-28. Accepted 2021-09-27.

CARLOS AMÉNDOLA: carlos.amendola@tum.de
Department of Mathematics, Technical University of Munich, Germany

KATHLÉN KOHN: kathlen@kth.se
Department of Mathematics, KTH Royal Institute of Technology, Stockholm, Sweden

PHILIPP REICHENBACH: reichenbach@tu-berlin.de
Institut für Mathematik, Technische Universität Berlin, Germany

ANNA SEIGAL: seigal@maths.ox.ac.uk
Mathematical Institute, University of Oxford, United Kingdom

Current address: Department of Mathematics, Harvard University, Cambridge, MA, United States
DISCRETE MAX-LINEAR BAYESIAN NETWORKS

BENJAMIN HOLLERING AND SETH SULLIVANT

Discrete max-linear Bayesian networks are directed graphical models specified by the same recursive structural equations as max-linear models but with discrete innovations. When all of the random variables in the model are binary, these models are isomorphic to the conjunctive Bayesian network (CBN) models of Beerenwinkel, Eriksson, and Sturmfels. Many of the techniques used to study CBN models can be extended to discrete max-linear models and similar results can be obtained. In particular, we extend the fact that CBN models are toric varieties after linear change of coordinates to all discrete max-linear models.

1. Introduction

Max-linear Bayesian networks are a special class of graphical models introduced in [6] to model extreme events. A max-linear Bayesian network, $X = (X_1, \ldots, X_n)$, is determined by a directed acyclic graph $D = (V, E)$, edge weights $c_{ij} \geq 0$, and independent positive random variables $Z_1, \ldots, Z_n$ called innovations. The $Z_i$ have support $(0, \infty)$ and are required to have atom-free distributions. Then the random vector $X$ is required to satisfy the structural equations

$$X_i = \bigvee_{j \in \text{pa}(i)} c_{ij} X_j \lor Z_i,$$

where $\lor$ denotes maximum and $\text{pa}(i)$ denotes the parents of vertex $i$ in $D$. These models are able to more accurately model extreme events spreading throughout a network than standard discrete or Gaussian Bayesian networks [6; 10] and also have interesting ties to tropical geometry [1].

Max-linear models also exhibit a remarkable property concerning conditional independence: they are usually not faithful to their underlying DAG, meaning that they often satisfy more conditional independence statements than those implied by the d-separation criterion [10]. Améndola, Klüppelberg, Lauritzen, and Tran recently gave a new criterion named $\star$-separation which gives a complete set of conditional independence statements for max-linear models [1]. There has also been work done to establish when the parameters of the model are identifiable [6; 7; 10]. In all of this work, the atom-free property of the innovations is key.

In this note, we consider discrete max-linear Bayesian networks so we assume the $Z_i$ are all $k$-state discrete random variables. This means the $Z_i$ are now atomic so much of the above work on max-linear models does not apply; however, these models are closely related to the conjunctive Bayesian network (CBN) models discussed in [3; 4]. CBN models were originally introduced in [3] to model HIV drug
resistance and cancer development via mutation in a genome. In CBN models, events accumulate as you move up a poset (or DAG) similar to the way extreme events spread throughout a max-linear Bayesian network. Part of our goal in this note is to show that CBN models are naturally included in the family of max-linear models. Though we do not develop the details here, we also believe it would be interesting to find a relationship between the typical max-linear model and our discrete analogue similar to the relationship found between the CBN model and its continuous analogue as the authors did in [2].

We also show that all D-MLBN models are toric varieties after a natural change of coordinates which is similar to the coordinate change used to show that CBN models are toric varieties. By generalizing the approach used for CBN models, we are able to find the vanishing ideal of the D-MLBN model in the transformed coordinates which gives part of the implicit description of the model.

In Section 2 we give a brief overview of the combinatorial objects we use throughout the note and some background on the CBN model. In Section 3 we define the discrete max-linear Bayesian network model and show that when all of the random variables involved are binary, it is equal to the CBN model after a natural change of coordinates. In Section 4 we describe the algebraic structure of discrete max-linear models. In Section 5 we discuss some open problems concerning discrete max-linear models.

2. Preliminaries

In this section we provide some basic background on graphs and posets which we will use throughout this note. We also discuss conjunctive Bayesian network (CBN) models which were first introduced in [3]. These models are directly related to the discrete max-linear models we discuss in this note.

Graphs, posets, and lattices. Our notation for graphs follows that of [1] while our notation for posets and lattices follows that of [11]. The graphs in this note will primarily be directed graphs which are given by a vertex set \( V = \{ 1, \ldots, n \} \) and an edge set \( E = \{ j \rightarrow i : i, j \in V, i \neq j \} \). A vertex \( j \) is a parent of \( i \) if \( j \rightarrow i \in E \) and we denote the set of parents of \( i \) by \( \text{pa}(i) \). A path from \( j \) to \( i \) is a sequence of distinct nodes \( j = \ell_0, \ell_1, \ldots, \ell_m = i \) such that either \( \ell_r \rightarrow \ell_{r+1} \) or \( \ell_{r+1} \rightarrow \ell_r \) for each \( r \). A directed path from \( j \) to \( i \) is a path where \( \ell_r \rightarrow \ell_{r+1} \) for all \( r \) and a directed cycle is a directed path where \( j = i \). A vertex \( j \) is an ancestor of \( i \) if there exists a directed path from \( j \) to \( i \) and we denote the set of ancestors of \( i \) by \( \text{an}(i) \). Most directed graphs in this note will be directed acyclic graphs (DAG) which are directed graphs with no directed cycles.

We will also frequently use partially ordered sets, typically called posets, throughout this note. A poset, \( P \), is a set equipped with a binary relation \( \leq \) that is reflexive, antisymmetric, and transitive. A chain is a poset where any two elements are comparable and we denote a chain whose elements are \( [k] = \{ 1, \ldots, k \} \) with the natural relation by \( k \). An element \( j \) of \( P \) is said to cover \( i \), if \( j > i \) and there is no other element \( k \) such that \( j > k > i \). A poset can be represented by an undirected graph called the Hasse diagram of the poset where the vertices correspond to elements of \( P \), edges represent cover relations, and the elements are ordered in the figure so smaller elements are at the bottom. See Figure 1 for an example.

An order ideal of \( P \) is a subset \( I \) of \( P \) such that if \( i \in I \) and \( j \leq i \) then \( j \in I \). A dual order ideal of \( P \) is a subset \( I \) of \( P \) such that if \( i \in P \) and \( j \geq i \) then \( j \in P \). An order-preserving map is a map \( \phi : P \rightarrow Q \) between posets such that if \( j, i \in P \) satisfy \( j \leq i \) then \( \phi(j) \leq \phi(i) \).
Many of the posets we work with will be obtained by taking the transitive closure of a DAG. Let $D$ be a DAG and define a relation on the vertices by $j \leq i$ if and only if there is a directed path from $j$ to $i$. We refer to this poset as $D^\vee$, which denotes the transitive closure poset.

A lattice, $L$, is a poset equipped with two binary operations, denoted $\lor$ and $\land$, such that for any elements $s, t \in L$ the least upper bound of $s$ and $t$ exists which is $s \lor t$ and the greatest lower bound of $s$ and $t$ exists which is $s \land t$. The operation $\lor$ is typically referred to as the join while $\land$ is called the meet. A classic example that will arise in this note is the lattice of order ideals of a finite poset. Let $P$ be a finite poset and denote by $J(P)$ the order ideals of $P$ ordered by inclusion then $J(P)$ is a lattice with meet given by intersection and join given by union. Furthermore, $J(P)$ is a distributive lattice meaning the operations $\lor$ and $\land$ distribute over each other. Throughout this note we will also use $\lor$ to denote the maximum of a subset of a totally ordered set and $\land$ to denote the minimum since these operations are the join and meet respectively for finite subsets of totally ordered sets.

**Conjunctive Bayesian networks.** In this section we provide some basic background on the CBN model which was first introduced by Beerenwinkel, Eriksson, and Sturmfels in [3] as a mathematical model for mutations occurring in a genome. For further information we refer the reader to [3; 4]. In Section 3, we show that CBN models can be viewed as a subclass of D-MLBN models which are our main focus in this note.

The CBN model begins with a poset $P$ whose elements are called events which are usually taken to be the elements of $[n]$. The state space of the model is the lattice of order ideals $J(P)$ and elements $g \in J(P)$ are called genotypes. It is often convenient to think of a state $g$ both as a subset of the ground set $[n]$ and as a 0-1 string and we will do so throughout the note.

The parametrization of the CBN model can be explicitly written down in terms of the poset $P$ but it is often convenient to instead view it as a directed graphical model. We refer the reader to [12, Chapter 13] for additional information on parametrizations of graphical models. We first form a DAG with edges $i \rightarrow j$ if $i < j$ is a cover relation in $P$ and associate a binary random variable $X_i$ to each $i \in P$. Then the CBN model is a directed graphical model on the $X_i$ with conditional probabilities given by

$$
(P(X_i = b|X_{pa(i)} = a))_{a \in \{0,1\}^{pa(i)}, b \in \{0,1\}} = \begin{pmatrix}
1 & 0 \\
\vdots & \vdots \\
0 & \theta_0^{(i)} \\
1 & \theta_1^{(i)}
\end{pmatrix},
$$

where the rows and columns are ordered lexicographically and $\theta_1^{(i)}$ is simply the conditional probability that the event $i$ occurs given all of its parents have occurred while $\theta_0^{(i)}$ is the conditional probability that the event $i$ does not occur given all of its parents have occurred. This means that $\theta_0^{(i)} + \theta_1^{(i)} = 1$. We note that this slightly different than the matrix of conditional probabilities shown in [4] but it is equivalent. The probability of observing an event $g$ is the product of these conditional probabilities so

$$p_g = P(X = g) = \prod_{i \in P} P(X_i = g_i|X_{pa(i)} = g_{pa(i)}).$$

**Example 2.1.** Let $P$ be the poset pictured on the left of Figure 1. The state space of the CBN model on $P$ is the lattice of order ideals $J(P)$ pictured on the right of Figure 1. In this representation, the 0-1
strings represent the corresponding subsets of the ground set of P. For instance, the element (1, 1, 0, 1) of J(P) represents the order ideal {1, 2, 3, 5} of P. The parametrization of the CBN model on P is

\[
p_{00000} = \theta_0^{(1)} \theta_0^{(2)}, \quad p_{10000} = \theta_1^{(1)} \theta_0^{(2)}, \\
p_{01000} = \theta_0^{(1)} \theta_1^{(2)}, \quad p_{11000} = \theta_1^{(1)} \theta_1^{(2)} \theta_0^{(3)}, \\
p_{11100} = \theta_1^{(1)} \theta_1^{(2)} \theta_1^{(3)} \theta_0^{(4)} \theta_0^{(5)}, \quad p_{11110} = \theta_1^{(1)} \theta_1^{(2)} \theta_1^{(3)} \theta_1^{(4)} \theta_0^{(5)}, \\
p_{11111} = \theta_1^{(1)} \theta_1^{(2)} \theta_1^{(3)} \theta_1^{(4)} \theta_1^{(5)}.
\]

Observe that for this note we suppress the commas and parentheses when writing \( p_g \).

3. Discrete max-linear Bayesian networks

In this section we introduce a discrete version of the max-linear models studied in [1; 6; 10]. When all of the random variables in the model are binary, there is a direct correspondence between discrete max-linear models and the conjunctive Bayesian networks discussed in [3; 4]. For this reason, we try to keep our notation consistent with that of [4] when possible.

Let \( D = (V, E) \) be a directed acyclic graph (DAG) and associated a discrete random variable \( Z_i \) with \( k \) states to each vertex \( i \in V \). Then the discrete max-linear Bayesian network (D-MLBN) is the family of joint distributions of the random variables \( (X_i)_{i \in V} \) specified by

\[
X_i = \bigvee_{j \in \text{pa}(i)} X_j \lor Z_i, \quad i \in V,
\]

where \( \text{pa}(i) \) denotes the parents of vertex \( i \) in \( D \). These are the same structural equations used to specify the max-linear Bayesian networks discussed in [1; 6; 10] except there are no coefficients and the random
variables $Z_1, \ldots, Z_n$ are now discrete instead of continuous and atom-free. Despite these alterations, this system of equations still has the same solution which is

$$X_i = \bigvee_{j \in \text{an}(i) \cup \{i\}} Z_j, \quad i \in V,$$

(3)

where $\text{an}(i)$ denotes the ancestors of vertex $i$ in $D$.

The state space of the $k$-state discrete max-linear model is the set of order-preserving maps from the transitive closure of $D$ to a chain of length $k$. More explicitly, let $k$ be a chain of size $k$ and let $D^\text{tr}$ be the poset obtained by taking the transitive closure of $D$. Then $g = (g_1, \ldots, g_n)$ is a possible state of the $k$-state D-MLBN if there exists an order-preserving map $\pi : D^\text{tr} \rightarrow k$ such that $\pi(i) = g_i$. This can be seen by directly examining the structural equations and their solution. Note that if $i \geq j$ in the partial order $D^\text{tr}$, then $j$ is an ancestor of $i$ so there is a directed path $[j = \ell_0, \ell_1, \ldots, \ell_m = i]$ and for any $r$ it is immediate that, $\ell_r \in \text{pa}(\ell_{r+1})$ which immediately implies that $X_{\ell_{r+1}} \geq X_{\ell_r}$. This gives a chain of inequalities from which we get $X_j \leq X_i$. Denote this set of states by $\mathcal{G}(D, k)$ and note that this forms a distributive lattice with meet given by taking the coordinatewise minimum and join given by taking the coordinatewise maximum. When it is clear from context, we simply write $\mathcal{G}$ instead of $\mathcal{G}(D, k)$. This is analogous to the state space of the CBN model being the lattice of order ideals of a poset [4].

**Example 3.1.** Let $D$ be the DAG pictured in Figure 2. Then the structural equations of the D-MLBN model on $D$ are

$$X_1 = Z_1, \quad X_2 = Z_2, \quad X_3 = X_1 \lor Z_3, \quad X_4 = X_1 \lor X_2 \lor Z_4, \quad X_5 = X_3 \lor X_4 \lor Z_5$$

which have the solution

$$X_1 = Z_1, \quad X_2 = Z_2, \quad X_3 = Z_1 \lor Z_3, \quad X_4 = Z_1 \lor Z_2 \lor Z_4, \quad X_5 = Z_1 \lor Z_2 \lor Z_3 \lor Z_4 \lor Z_5.$$ 

If each of the $Z_i$ has two states, so $k = 2$, then the state space of the model is the lattice $\mathcal{G}$ that is also pictured in Figure 2.

Similarly to the CBN model, the D-MLBN model can also be thought of as a directed graphical model on a DAG $D$ with additional restrictions on the parameters. In the usual directed graphical model, the parameters of the model are the conditional probabilities $P(X_i = x_i \mid X_{pa(i)} = x_{pa(i)})$ [5]. In the D-MLBN model, we can compute these conditional probabilities in terms of the distributions of the $Z_i$. Let each $Z_i$ have distribution $\theta^{(i)} = (\theta^{(i)}_0, \ldots, \theta^{(i)}_{k-1}) \in \Delta_{k-1}$ and $g \in \{0, \ldots, k-1\}^n$. For any $i$ define $M_i = \bigvee_{j \in \text{pa}(i)} g_j$ then we have

$$P(X_i = g_i \mid X_{pa(i)} = g_{pa(i)}) = \begin{cases} 0, & g_i < M_i, \\ \sum_{\ell \leq M_i} \theta^{(i)}_\ell, & g_i = M_i, \\ \theta^{(i)}_{g_i}, & g_i > M_i. \end{cases}$$

(4)

Using these conditional probabilities, we can then compute the probability $p_g = P(X = g)$ which is

$$p_g = \prod_{i \in V} P(X_i = g_i \mid X_{pa(i)} = g_{pa(i)}).$$
Note that if \( g \) does not come from an order preserving map, that is \( g \notin \mathcal{G}(\mathcal{D}, k) \), then \( \rho_g = 0 \). This means we can either think of the D-MLBN model as just being a model for the states \( \mathcal{G}(\mathcal{D}, k) \) or a model for all \( g \in \{0, \ldots, k - 1\}^n \) where the \( g \notin \mathcal{G}(\mathcal{D}, k) \) have probability 0. Both of these perspectives can be useful when studying the algebraic structure of the model. The following example illustrates this parametric description of the model.

**Example 3.2.** We again let \( \mathcal{D} \) and \( \mathcal{G} \) be the DAG and lattice pictured in Figure 2. Then from the above discussion the parametrization of the D-MLBN model is

\[
\begin{align*}
p_{00000} &= \theta_0^{(1)} \theta_0^{(2)} \theta_0^{(3)} \theta_0^{(4)} \theta_0^{(5)}, & p_{00001} &= \theta_0^{(1)} \theta_0^{(2)} \theta_0^{(3)} \theta_0^{(4)} \theta_1^{(5)}, \\
p_{00011} &= \theta_0^{(1)} \theta_0^{(2)} \theta_1^{(3)} \theta_0^{(4)}, & p_{00101} &= \theta_0^{(1)} \theta_0^{(2)} \theta_0^{(3)} \theta_1^{(4)}, \\
p_{00111} &= \theta_0^{(1)} \theta_0^{(2)} \theta_0^{(3)} \theta_1^{(4)}, & p_{01011} &= \theta_0^{(1)} \theta_1^{(2)} \theta_0^{(3)}, \\
p_{01111} &= \theta_1^{(1)} \theta_1^{(2)} \theta_0^{(3)}, & p_{10111} &= \theta_1^{(1)} \theta_0^{(2)}, \\
p_{11111} &= \theta_1^{(1)} \theta_1^{(2)}. &
\end{align*}
\]

Note that while this appears to be a monomial map, the relationship \( \theta_0^{(i)} + \theta_1^{(i)} = 1 \) implies that it is not and so the ideal in these coordinates is not toric.

We now describe the relationship between the 2-state D-MLBN model and the CBN model of [5].

**Theorem 3.3.** Let \( \mathcal{D} = (V, E) \) be a DAG and \( \rho \) be the map parametrizing the CBN model on \( \mathcal{D}^\mathfrak{f} \). Let \( \psi \) be the map parametrizing the 2-state D-MLBN model on \( \mathcal{D} \). Then image(\( \rho \)) is equal to image(\( \psi \)) after a natural relabeling of coordinates. In particular, there exists a bijection, \( \phi : J(\mathcal{D}^\mathfrak{f}) \to \mathcal{G}(\mathcal{D}, 2) \) such that \( \rho_g(\theta_0^{(1)}, \theta_1^{(1)}, \ldots, \theta_0^{(n)}, \theta_1^{(n)}) = \psi_{\phi(g)}(\theta_1^{(1)}, \theta_0^{(1)}, \ldots, \theta_1^{(n)}, \theta_0^{(n)}) \).

\[\text{Figure 2.} \quad \text{A DAG } \mathcal{D} \text{ and the state space, } \mathcal{G}, \text{ of the 2-state D-MLBN model pictured as the lattice of order preserving maps from } \mathcal{D} \text{ to the chain 2. Also pictured is the lattice of order ideals } J(\mathcal{D}^\mathfrak{f}) \text{ of the poset } \mathcal{D}^\mathfrak{f} \text{ written as 0-1 vectors.} \]
Proof. We first describe the bijection between the state spaces of the two models. Recall that the state space of the CBN model is the distributive lattice $J(D^n)$ of order ideals. There is a natural bijection between elements $g \in J(D^n)$ and order-preserving maps from $\pi : D^r \to \{0, 1\}$. The general form of the following bijection can be found in [11, Proposition 3.5.1] but we describe the special case for order-preserving maps to $\{0, 1\}$ here. For any order ideal $g$ of $D^r$ let $\pi_g$ be the map defined by

$$
\pi_g(i) = \begin{cases} 
0, & i \in g, \\
1, & i \notin g.
\end{cases}
$$

It remains to show that $\rho_g(\theta_{0}^{(1)}, \theta_{1}^{(1)}, \ldots, \theta_{0}^{(n)}, \theta_{1}^{(n)}) = \psi(\theta)(\theta_{1}^{(1)}, \theta_{0}^{(1)}, \ldots, \theta_{1}^{(n)}, \theta_{0}^{(n)})$. We do this by showing that the two models have the same conditional probabilities after interchanging $\theta_{0}^{(i)}$ and $\theta_{1}^{(i)}$.

Let $\phi(g) \in G(D, 2)$ be a state of the D-MLBN model and for each $i$ let $M_i = \bigvee_{j \in \text{pa}(i)} \phi(g)_j$. Observe that

$$
M_i = \bigvee_{j \in \text{pa}(i)} \phi(g)_j = \begin{cases} 
0, & \phi(g)_j = 0 \text{ for all } j \in \text{pa}(i), \\
1, & \text{otherwise}
\end{cases}
$$

with the second equality following from the definition of $\phi$. We now examine the different possibilities for $\phi(g)_i$ and $M_i$ and compute the conditional probabilities in each case.

Suppose $\phi(g)_i = 0$ and $M_i = 0$, then under the D-MLBN model, we have

$$
P(X_i = \phi(g)_i | X_{\text{pa}(i)} = \phi(g)_{\text{pa}(i)}) = \sum_{\ell \leq 0} \theta_{\ell}^{(i)} = \theta_{0}^{(i)}.
$$

We know from the above formula for $M_i$ that if $M_i = 0$, then for all $j \in \text{pa}(i)$, $g_j = 1$. Since $\phi(g)_i = 0$ we have that $g_i = 1$ and so the corresponding entry of the matrix of conditional probabilities for the CBN model in (1) is $\theta_{1}^{(i)}$.

If $\phi(g)_i = 1$ and $M_i = 0$, then under the D-MLBN model, we have

$$
P(X_i = \phi(g)_i | X_{\text{pa}(i)} = \phi(g)_{\text{pa}(i)}) = \theta_{1}^{(i)}.
$$

On the other hand, we now have $g_i = 0$ and $M_i = 0$ so the conditional probability for the CBN model is $\theta_{0}^{(i)}$.

It is straightforward to check the remaining cases so we omit it here. In these cases both models have the same conditional probabilities which are either 0 or 1. Since each model is a directed graphical model, the probability of observing $g$ (or $\phi(g)$) is simply

$$
\rho_g(\theta) = \prod_{i \in V} P_{CBN}(X_i = g_i | X_{\text{pa}(i)} = g_{\text{pa}(i)}),
$$

$$
\psi_g(\theta) = \prod_{i \in V} P_{D-MLBN}(X_i = g_i | X_{\text{pa}(i)} = g_{\text{pa}(i)}).
$$

Since we know that the conditional probabilities are equal after interchanging the parameters $\theta_{0}^{(i)}$ and $\theta_{1}^{(i)}$ for each $i$, we have that the above products are equal after interchanging the corresponding parameters as claimed above.

\[\Box\]
We end this section with an example that illustrates the previous theorem.

**Example 3.4.** Let \( \mathcal{D} \) be the DAG pictured on the left of Figure 2. We have already seen that \( \mathcal{G} = \mathcal{G}(\mathcal{D}, 2) \) pictured in the middle of Figure 2 is the state space of the D-MLBN model on \( \mathcal{D} \) while the lattice of order ideals \( J(\mathcal{D}^{tr}) \) that is pictured on the right of Figure 2 is the state space of the CBN model on \( \mathcal{D}^{tr} \).

The map \( \phi \) from Theorem 3.3 maps the element \( g = (1, 0, 1, 0, 0) \in J(\mathcal{D}^{tr}) \) to the element \( \phi(g) = (0, 1, 0, 1, 1) \in \mathcal{G} \). The probability of \( g \) under the CBN model is

\[
\rho_{10100}(\theta_0^{(1)}, \theta_1^{(1)}, \ldots, \theta_0^{(5)}, \theta_1^{(5)}) = \theta_1^{(1)} \theta_0^{(2)} \theta_1^{(3)},
\]

while the probability of \( \phi(g) \) under the D-MLBN model is

\[
\psi_{01011}(\theta_0^{(1)}, \theta_1^{(1)}, \ldots, \theta_0^{(5)}, \theta_1^{(5)}) = \theta_0^{(1)} \theta_1^{(2)} \theta_0^{(3)}.
\]

We can see that these two probabilities will be equal after interchanging the parameters \( \theta_0^{(i)} \) and \( \theta_1^{(i)} \).

### 4. Algebraic structure of the D-MLBN model

In this section we describe the algebraic structure of the D-MLBN model. We do this by extending the techniques developed for CBN models in [4] to all D-MLBN models. The main tool here is Möbius inversion which corresponds to a linear change of coordinates on the D-MLBN model. In these new coordinates, the ideal of polynomials that vanish on a D-MLBN belongs to a special class of toric ideals whose Gröbner bases were described by Hibi in [9].

Let \( \mathcal{D} \) be a DAG and \( \psi_D^{(k)} \) be the map parametrizing the \( k \)-state D-MLBN model on \( \mathcal{D} \) and \( \mathcal{G} \) be the state space of the model. Also let \( \mathbb{R}[p_g] = \mathbb{R}[p_g : g \in \mathcal{G}] \), then our goal is to find a Gröbner basis for the ideal

\[
I_D^{(k)} = \{ f \in \mathbb{R}[p_g] : f(a) = 0 \text{ for all } a \in \text{image}(\psi_D^{(k)}) \}.
\]

Finding a Gröbner basis for the ideal \( I_D \) is a first step in obtaining an implicit description of the D-MLBN model on \( \mathcal{D} \).

**Example 4.1.** Again let \( \mathcal{D} \) be the DAG pictured on the left in Figure 2. Then the ideal \( I_D^{(2)} \) is generated by the polynomials

\[
\begin{align*}
p_{00000} + p_{00001} + p_{00011} + p_{00101} + p_{00111} + p_{01011} + p_{01111} + p_{11111} - 1, \\
p_{00111} p_{10111} - p_{00101} p_{11111} - p_{00111} p_{11111}, \\
p_{00111} p_{10111} - p_{00000} p_{11111} - p_{00001} p_{11111} - p_{00011} p_{11111}, \\
p_{00111} p_{01011} - p_{00011} p_{01111}, \\
p_{00101} p_{01011} - p_{00000} p_{01111} - p_{00001} p_{01111}, \\
p_{00011} p_{00101} - p_{00000} p_{00111} - p_{00001} p_{00111}.
\end{align*}
\]

As we noted before, the state space of the D-MLBN model is a distributive lattice. Hibi showed in [9] that there is a toric ideal naturally associated to such a lattice. In [4] the authors show that the ideal of the CBN model is the toric ideal defined by Hibi, after a suitable change of coordinates. So it is natural...
to attempt to extend this result from the binary case to D-MLBN models with an arbitrary number of states. We describe the construction of Hibi here but for additional information we refer the reader to [8] or [9]. Let $L$ be a distributive lattice and recall that there is unique poset $P$ (up to isomorphism) such that $L = J(P)$ (see [11, Theorem 3.4.1]). Let $P$ have ground set $[n]$. Then the map

$$\varphi_L : \mathbb{R}[q_g : g \in L] \rightarrow \mathbb{R}[t, x_1, \ldots, x_n], \quad q_g \mapsto t \prod_{i \in g} x_i$$

has kernel $I_L = \ker(\varphi_L)$ generated by $I_L = \langle q_g q_h - q_g \land h q_g \lor h : g, h \in \mathcal{G} \text{ are incomparable} \rangle$ [9]. Recall that elements $g$ and $h$ in a poset are incomparable if neither $g \leq h$ nor $h \leq g$. The following example illustrates this construction.

**Example 4.2.** Let $P$ be the poset pictured on the left in Figure 1 whose lattice of order ideals $L = J(P)$ is pictured on the right. Then $\varphi_L$ is given by

$$q_{00000} = t, \quad q_{10000} = tx_1, \quad q_{01000} = tx_2, \quad q_{11000} = tx_1x_2, \quad q_{11100} = tx_1x_2x_3, \quad q_{11110} = tx_1x_2x_3x_4, \quad q_{11111} = tx_1x_2x_3x_4x_5.$$  

There are two pairs of incomparable elements in the lattice $L$ which are $(1, 0, 0, 0, 0)$ and $(0, 1, 0, 0, 0)$ as well as $(1, 1, 1, 0)$ and $(1, 1, 1, 0, 1)$. This means the generators of the ideal $I_L$ are the binomials

$$q_{10000}q_{01000} - q_{00000}q_{11000}, \quad q_{11110}q_{11101} - q_{11100}q_{11111}$$

which correspond to these two pairs of incomparable elements.

We are now ready to state our main result.

**Theorem 4.3.** Let $\mathcal{D} = (V, E)$ be a DAG and $I_D^{(k)}$ be vanishing ideal of the $k$-state D-MLBN model on $\mathcal{D}$. Then after homogenizing the map $\psi_D^{(k)}$ and applying the linear change of coordinates

$$q_g = \sum_{h \leq g} p_h$$

on $\mathbb{R}[p_g]$ the ideal $I_D^{(k)}$ is equal to the toric ideal associated to the distributive lattice $J(\mathcal{D}^F \times k - 1)$ with generating set

$$I_D^{(k)} = \langle q_g q_h - q_g \land h q_g \lor h : g, h \in \mathcal{G} \text{ are incomparable} \rangle.$$  

**Proof.** First we note that for any $g \in \mathcal{G} = \mathcal{G}(\mathcal{D}, k)$ we have that

$$\sum_{h \leq g} p_h = \prod_{i \in V} \sum_{\ell \leq g_i} \theta_\ell^{(i)}.$$  

This leads us to consider a transform of the parameter space given by

$$\alpha_j^{(i)} = \sum_{\ell \leq j} \theta_\ell^{(i)}.$$
Note that the matrix of this transformation is block diagonal and each block can be made into a lower triangular matrix with ones on the diagonal so this is truly a linear change of coordinates. Combining equations (5) and (6) we see that in the transformed coordinates, the map \( \psi_D^{(k)} \) is given by

\[
q_g = t \prod_{i \in V} \alpha_{gi}^{(i)}.
\]

We also have introduced a new variable \( t \) which homogenizes the parametrization so that \( I_D^{(k)} \) will be a homogeneous ideal. This simply removes the trivial relation that all of the coordinates sum to one.

We now consider the parametrization of the Hibi ideal associated to \( I_L \) where

\[
L = J(D^\text{tr} \times \{0, 1, \ldots, k - 2\}).
\]

By [11, Proposition 3.5.1], there is a bijection between order-preserving maps \( g \in G \) and order ideals of the poset \( D^\text{tr} \times \{0, 1, \ldots, k - 2\} \). Under this bijection, a map \( g \) is sent to an order ideal

\[
\tilde{g} = \{(i, r) \in D^\text{tr} \times \{0, 1, \ldots, k - 2\} : 0 \leq r \leq k - g_i - 1\}.
\]

Then \( I_L \) is the kernel of the map

\[
\varphi_L : \mathbb{R}[Q_g : g \in L] \to \mathbb{R}[t, x_r^{(i)} : i \in [n], r \in 0, \ldots, k - 2], \quad q_g \mapsto t \prod_{(i, r) \in \tilde{g}} x_r^{(i)}.
\]

At first, this parametrization might look quite different when compared to the parametrization \( \psi_D^{(k)} \) but we can transform the parameter space of \( \psi_D^{(k)} \) again so that they agree. Consider the transform given by

\[
\alpha_j^{(i)} = \prod_{r=0}^{k-j-2} x_r^{(i)}
\]

and note that this is invertible with inverse given by

\[
x_{i-j}^{(i)} = \frac{\alpha_{j-1}^{(i)}}{\alpha_j^{(i)}}.
\]

After applying this transform on the parameter space of \( \psi_D^{(k)} \) the map is given by

\[
q_g = t \prod_{i \in V} \prod_{r=0}^{k-g_i-2} x_r^{(i)} = t \prod_{(i, r) \in \tilde{g}} x_r^{(i)},
\]

where the last equality follows directly from the definition of \( \tilde{g} \). Since the ideals \( I_D^{(k)} \) and \( I_L \) are now the kernel of the exact same map, they are equal and the generating set stated above is exactly the generating set described by Hibi in [9].

\( \square \)

Remark. Note that the coordinate transform that takes the \( p_g \) coordinates to the \( q_g \) coordinates corresponds to Möbius inversion on the lattice \( G \). This is the same transform that is used in [4] for CBN models but in this case the resulting toric ideal is that associated to \( J(D^\text{tr}) \) since it is a 2-state model.

We conclude this section with two examples which illustrate the previous theorem.
Example 4.4. Let $\mathcal{D}$ be the graph pictured in Figure 3 on the left. Consider the state $g = (0, 1, 2) \in G$ of the 3-state D-MLBN model on $\mathcal{D}$. The original probability of this state under the model is $p_{012} = \theta_0^{(1)}\theta_1^{(2)}\theta_2^{(3)}$. Then after our first coordinate transform

$$q_{012} = \sum_{h \leq g} p_h = p_{000} + p_{001} + p_{002} + p_{010} + p_{011} + p_{012} = \theta_0^{(1)}(\theta_0^{(2)} + \theta_1^{(2)})(\theta_0^{(3)} + \theta_1^{(3)} + \theta_2^{(3)}) = \alpha_0^{(1)}\alpha_1^{(2)}.$$ 

Note that we omit $\alpha_2^{(3)}$ since the parameter corresponding to the state $k - 1$ is always 1. Under our second transform of the parameter space the relevant parameters become $\alpha_0^{(1)} = x_0^{(1)}x_1^{(1)}$ and $\alpha_1^{(2)} = x_0^{(2)}$. After homogenizing with a new parameter $t$ and applying this second transform we have that

$$\psi_\mathcal{D}^{(3)}(q_{012}) = tx_0^{(1)}x_1^{(1)}x_0^{(2)}.$$ 

The state $g$ also corresponds to the order ideal $\tilde{g} = \{(1, 0), (1, 1), (2, 0)\} \in L = J(\mathcal{D}^\text{tr} \times \{0, 1\})$. This means the map $\varphi_L$ takes $q_g$ to

$$\varphi_L(q_{012}) = tx_0^{(1)}x_1^{(1)}x_0^{(2)}$$

and we can see that $\varphi_L(q_g) = \psi_\mathcal{D}^{(3)}(q_g)$ as was shown in the proof of Theorem 4.3.

Example 4.5. Let $\mathcal{D}$ be the graph pictured in Figure 2. Then after homogenizing the parametrization $\psi_\mathcal{D}^{(2)}$ and applying the coordinate transform described in Theorem 4.3 the ideal $I_\mathcal{D}^{(2)}$ is generated by the polynomials

$$q_{01111}q_{10111} - q_{00111}q_{11111}, \quad q_{01011}q_{10111} - q_{00011}q_{11111}, \quad q_{00111}q_{01011} - q_{00011}q_{01111},$$

$$q_{00101}q_{01011} - q_{00001}q_{01111}, \quad q_{00011}q_{00101} - q_{00001}q_{00111}.$$ 

Note that the first monomial in each polynomial corresponds to a pair of incomparable elements $g, h \in G$ while the second corresponds to their meet and join which are given by taking coordinatewise minimums and maximums respectively.

5. Open problems

In this section we discuss some interesting open problems concerning both the theory and applications of D-MLBN models.
In the original max-linear model, the random vector $X$ is given by the recursive structural equations

$$X_i = \bigvee_{j \in \text{pa}(i)} c_{ij} X_j \lor Z_i,$$

where the $c_{ij}$ are coefficients. These recursive structural equations have a similar solution to that shown in (3) which is

$$X_i = \bigvee_{j \in \text{an}(i) \cup \{i\}} c_{*ij} Z_j, \quad i \in V,$$

where $C^*$ is the Kleene star of the matrix $C$. That is, $C^* = \bigvee_{k=0}^{n-1} C^k$ where $C^k$ represents the matrix $C$ raised to the $k$-th power using tropical matrix multiplication. The entries of this matrix can also be thought of as the weight of the highest weight path from the vertex $j$ to the vertex $i$.

In our formulation of the D-MLBN model we set all of these coefficients to 1 but one could consider a version of the D-MLBN model with arbitrary coefficients. In this case, one can write the random vector $X$ as

$$X = \bigvee_{i=1}^{n} c_{i}^* Z_i,$$

where $c_{i}^*$ is the $i$-th column of the matrix $C^*$. Note that this means that $X$ is in the tropical convex hull of the columns of $C^*$ so the D-MLBN model with coefficients can be thought of as a tropical polytope whose vertices are the columns of $C^*$.

**Question 5.1.** Can the tropical polyhedral structure of the D-MLBN model with coefficients be used to better understand the model? In particular, can it be used to deduce the vanishing ideal of the model or answer statistical questions about the model such as finding maximum likelihood estimates?

It would also be interesting to further explore the applications of the D-MLBN model. Estimating the continuous max-linear model can be quite difficult when the data has noise [7]. Estimating the continuous time analogue of the CBN model can also be quite difficult but the authors in [2] used the known maximum likelihood estimate for the discrete CBN model as a starting point in an iterative method to find the maximum likelihood estimate for some continuous time CBN models. This motivates the following question.

**Question 5.2.** Is there a relationship between the discrete max-linear model and the continuous max-linear model? Can the toric structure of the discrete model be used to help estimate the continuous model?

An answer to Question 5.2 would help solve a major open problem concerning max-linear models and make it much easier to apply them to actual data.

**Acknowledgments**

Benjamin Hollering and Seth Sullivant were partially supported by the US National Science Foundation (DMS 1615660).
References


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 \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.
<table>
<thead>
<tr>
<th>115</th>
<th>An algebraic Monte-Carlo algorithm for the partition adjacency matrix realization problem</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ÉVA CZABARKA, LÁSZLÓ SZÉKELY, ZOLTÁN TOROCZKAI and SHANISE WALKER</td>
</tr>
<tr>
<td>125</td>
<td>Bimonotone subdivisions of point configurations in the plane</td>
</tr>
<tr>
<td></td>
<td>ELINA ROBEVA and MELINDA SUN</td>
</tr>
<tr>
<td>139</td>
<td>A generalization for the expected value of the earth mover’s distance</td>
</tr>
<tr>
<td></td>
<td>WILLIAM Q. ERICKSON</td>
</tr>
<tr>
<td>167</td>
<td>Likelihood equations and scattering amplitudes</td>
</tr>
<tr>
<td></td>
<td>BERND STURMFELS and SIMON TELEN</td>
</tr>
<tr>
<td>187</td>
<td>Toric invariant theory for maximum likelihood estimation in log-linear models</td>
</tr>
<tr>
<td></td>
<td>CARLOS AMÉNDOLA, KATHLÉN KOHN, PHILIPP REICHENBACH and ANNA SEIGAL</td>
</tr>
<tr>
<td>213</td>
<td>Discrete max-linear Bayesian networks</td>
</tr>
<tr>
<td></td>
<td>BENJAMIN HOLLERING and SETH SULLIVANT</td>
</tr>
</tbody>
</table>