Wasserstein Distance

Bernd Sturmfels

January 31, 2023

A basic problem in metric algebraic geometry is finding a point in a variety X in $\mathbb{R}^n$ that is closest to a given data point $u \in \mathbb{R}^n$. Thus, we wish to solve the optimization problem

$$\text{minimize } ||x-u|| \text{ subject to } x \in X. \tag{1}$$

In what follows, this minimum in (1) is always attained because X is non-empty and closed. Hence there exists an optimal solution. If that solution is unique then we denote it by x^*.

We already discussed this problem for the Euclidean norm on $\mathbb{R}^n$. In this lecture we study (1) in the case when $||\cdot||$ is a *polyhedral norm*. This means that the unit ball

$$B \;=\; \{x \in \mathbb{R}^n : ||x|| \leq 1\}$$

is a centrally symmetric convex polytope. Every centrally symmetric convex polytope B in $\mathbb{R}^n$ defines a polyhedral norm on $\mathbb{R}^n$. Using the unit ball, we can paraphrase (1) as follows:

$$\text{minimize } \lambda \text{ subject to } \lambda \geq 0 \text{ and } (u+\lambda B) \cap X \;\neq\; \emptyset. \tag{2}$$

Familiar examples of polyhedral norms are $||\cdot||_\infty$ and $||\cdot||_1$, where the unit ball B is the cube and the crosspolytope respectively. Polyhedral norms are very important in optimal transport theory, where one uses Wasserstein norm on the space of probability distributions. This will be our main application in this lecture, and it will be discussed in detail later on.

We begin our discussion with a general polyhedral norm, that is, we allow B to be an arbitrary full-dimensional centrally symmetric polytope in $\mathbb{R}^n$. The boundary of B consists of faces whose dimensions range from 0 to $n-1$. We use the dot $\cdot$ for the standard inner product on $\mathbb{R}^n$. Recall that a subset F of B is a *face* if there exists $\ell \in \mathbb{R}^n \backslash \{0\}$ such that

$$F \;=\; \{\, x \in B \,:\, \ell \cdot x \,\geq\, \ell \cdot y \text{ for all } y \in B \,\}. \tag{3}$$

The set of all faces, ordered by inclusion, is a partial ordered set, called the *face poset* of B. An important combinatorial invariant of our polytope B is its *f-vector* $f(B) = (f_0, f_1, \ldots, f_{n-1})$. The ith coordinate f_i of the f-vector is the number of i-dimensional faces of B.

The dual of B is also a centrally symmetric polytope, namely

$$B^* \;=\; \{\, \ell \in \mathbb{R}^n \,:\, \ell \cdot x \,\leq\, 1 \text{ for all } x \in B \,\}.$$

The norm $||\cdot||_*$ defined by B^* is dual to the norm $||\cdot||$ given by B. The f-vector of B^* is the reverse of the f-vector of B. More precisely, we have $f_i(B^*) = f_{n-1-i}(B)$ for $i = 0, 1, \ldots, n-1$.

Example 1. Fix the unit cube $B = [-1, 1]^n$. Its dual is the cross-polytope

$$B^* \;=\; \mathrm{conv}\{\pm e_1, \pm e_2, \ldots, \pm e_n\} \;\subset\; \mathbb{R}^n.$$

Here e_j is the jth standard basis vector. The number of i-dimensional faces of the cube is

$$f_i(B) \;=\; \binom{n}{i} \cdot 2^{n-i}.$$

The 3-dimensional crosspolytope is the octahedron. The 3-cube and the octahedron satisfy

$$f(B) = (8, 12, 6) \quad \text{and} \quad f(B^*) = (6, 12, 8).$$

These numbers govern the combinatorial structure of the norms $||\cdot||_\infty$ and $||\cdot||_1$ on $\mathbb{R}^3$. $\diamond$

We now turn to the critical equations for the optimization problem given in (1) or (2). To derive these a combinatorial stratification of the problem is used, given by the face poset of the polytope B. Suppose that the variety X is in sufficiently general position in $\mathbb{R}^n$. This hypothesis implies that $(u + \lambda^* B) \cap X \;=\; \{x^*\}$ is a singleton for the optimal value λ^* in (2). The point $\frac{1}{\lambda^*}(x^* - u)$ lies in boundary of the unit ball B. Hence it lies in the relative interior of a unique face F of the polytope B. Let L_F denote the linear span of F in $\mathbb{R}^n$. We have $\dim(L_F) = \dim(F) + 1$. Let ℓ be any linear functional on $\mathbb{R}^n$ that attains its maximum over the polytope B at the face F. This means that (3) holds.

Lemma 2. *The optimal point x^* in (1) is the unique solution to the optimization problem*

$$\textit{Minimize} \;\; \ell(x) \;\; \textit{subject to } x \in (u + L_F) \cap X. \tag{4}$$

Proof. The general position hypothesis ensures that $u + L_F$ intersects X transversally, and x^* is a smooth point of that intersection. Moreover, x^* is a minimum of the restriction of ℓ to the variety $(u + L_F) \cap X$. By our hypothesis, this linear function is generic relative to the variety, so the number of critical points is finite and the function values are distinct. $\square$

Example 3 (Touching at a facet)**.** Suppose that the face F is a facet of the unit ball B. Then $L_F = \mathbb{R}^n$, and ℓ is an outer normal vector to that facet, which is unique up to scaling. Here, (4) asks for the minimum of ℓ over X. This corresponds to the left diagram in Figure 1.

Example 4 (Touching at a vertex)**.** Suppose F is a vertex of the unit ball B. This case arises when X is a hypersurface. It corresponds to the middle diagram in Figure 1. Here, the affine space $u + L_F$ is the line that connects u and x^*. That line intersects X in a finite set of cardinality degree(X), and x^* is the real point among them at which ℓ is minimal.

The problem (4) amounts to linear programming over a real variety. We now determine the algebraic degree of this optimization task when F is a face of codimension i. To this end, we replace the affine variety $X \subset \mathbb{R}^n$ by its closure in complex projective space $\mathbb{P}^n$, and we retain the same symbol X for that projective variety. We consider the affine space $L = u + L_F$ in $\mathbb{R}^n$ and we also identify it with its closure in $\mathbb{P}^n$. If the face F has codimension i then the linear space L has codimension $i - 1$. The following result assumes that that this space is in general position relative to the variety X and relative to the isotropic quadric.

Theorem 5. *Let L be a general affine-linear space of codimension $i-1$ in $\mathbb{R}^n$ and ℓ a general linear form. The number of critical points of ℓ on $L \cap X$ is the polar degree $\delta_i(X)$.*

Proof. This result is [5, Theorem 5.1]. The number of critical points of a linear form is the degree of the dual variety $(L \cap X)^\vee$. That degree coincides with the polar degree $\delta_i(X)$. □

Example 6. Examples 3 and 4 explain Theorem 5 in the two extreme cases $i = 1$ and $i = n$. Touching at a vertex ($i = n$) can only happen when X is a hypersurface, and here $\delta_n(X) = \text{degree}(X)$. Touching at a facet ($i = 1$) can happen for varieties of any dimension, as long as the dual variety $X^\vee$ is a hypersurface. In that case we have $\delta_1(X) = \text{degree}(X^\vee)$.

Theorem 5 offers a direct interpretation of each polar degree $\delta_i(X)$ in terms of optimization on X. Some readers might prefer this interpretation as a definition of polar degrees.

Example 7. Consider (1) and (2) where X is a general surface of degree d in $\mathbb{R}^3$. The optimal face F of the unit ball B depends on the location of the data point u. The algebraic degree of the solution x^* equals $\delta_3(X) = d$ if $\dim(F) = 0$, it is $\delta_2(X) = d(d-1)$ if $\dim(F) = 1$, and it is $\delta_1(X) = d(d-1)^2$ if $\dim(F) = 2$. Here x^* is the unique point in $(u + \lambda^* B) \cap X$, where λ^* is the optimal value in (2). Figure 1 visualizes this scenario for $d = 2$ and $||\cdot||_\infty$. The variety X is the green sphere, which is a surface of degree $d = 2$. The unit ball for the norm $||\cdot||_\infty$ is the cube $B = [-1, 1]^3$. The picture shows the smallest λ^* such that $u + \lambda^* B$ touches the sphere X. The cross marks the point of contact. This is the point x^* in X which is closest in ∞-norm to the green point u in the center of the cube. Point of contact is either on a facet, or on an edge, or it is a vertex. The algebraic degree of x^* is two in all three cases, i.e. we can write the solution x^* in terms of the data u by solving the quadratic formula. If the green variety were a cubic surface then these degrees would be 3, 6 and 12. ⋄

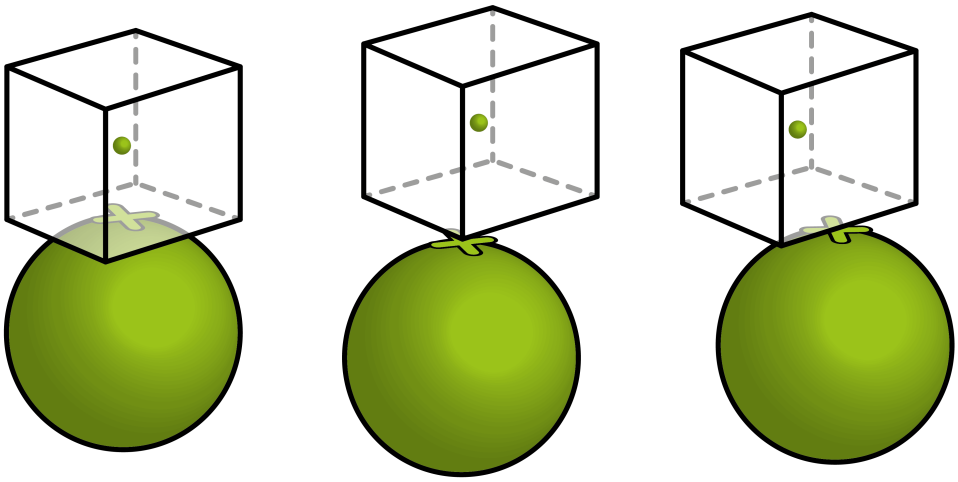

Figure 1: The cube is an $||\cdot||_\infty$ ball around the green point u. The variety X is the sphere. The contact point x^* is marked with a cross. The optimal face F is a facet, vertex, or edge.

We have learned that the conormal variety N_X and its cohomology class $[N_X]$ are key players when it comes to reliably solving the distance minimization problem for a variety X.

The polar degrees $\delta_i(X)$ reveal precisely how many paths need to be tracked by numerical solvers like [1, 3] in order to find and certify [2] the optimal solution x^* in (1) or (4).

We now come to the title of this lecture. The variety X will be an independence model in a probability simplex, described algebraically by matrices or tensors of low rank, and we measure distances using Wasserstein metrics on that simplex. This is a class of polyhedral norms which are important in optimal transport theory. We now present the relevant definitions.

A probability distribution on the finite set $[n] = \{1, 2, \ldots, n\}$ is a point ν in the simplex $\Delta_{n-1} = \{(\nu_1, \ldots, \nu_n) \in \mathbb{R}^n_{\geq 0} : \sum_{i=1}^n \nu_i = 1\}$. We metrize this simplex by the *Wasserstein distance*. To define this, we first turn the state space $[n]$ into a finite metric space by fixing a symmetric $n \times n$ matrix $d = (d_{ij})$ with nonnegative entries. These entries satisfy $d_{ii} = 0$ and $d_{ik} \leq d_{ij} + d_{jk}$ for all i, j, k. Given two probability distributions $\mu, \nu \in \Delta_{n-1}$, we consider the following linear programming problem, where $z = (z_1, \ldots, z_n)$ denotes the decision variables:

$$\text{Maximize} \ \sum_{i=1}^{n} (\mu_i - \nu_i)\, z_i \ \ \text{subject to} \ \ |z_i - z_j| \ \leq \ d_{ij} \ \ \text{for all} \ \ 1 \leq i < j \leq n. \tag{5}$$

The optimal value of (5), denoted $W_d(\mu, \nu)$, is the *Wasserstein distance* between μ and ν.

An optimal solution z^* to problem (5) is an *optimal discriminator* for the two probability distributions μ and ν. It satisfies $W_d(\mu, \nu) = \langle \mu - \nu, z^* \rangle$, and its coordinates z_i^* are weights on the state space $[n]$ that tell μ and ν apart. Here $\langle \cdot, \cdot \rangle$ is the standard inner product on $\mathbb{R}^n$. The linear program (5) is the *Kantorovich dual* of the *optimal transport problem*.

The feasible region of the linear program (5) is unbounded because it is invariant under translation by $\mathbf{1} = (1, 1, \ldots, 1)$. It is compact after taking the quotient modulo the line $\mathbb{R}\mathbf{1}$:

$$P_d \ = \ \big\{ \, z \in \mathbb{R}^n / \mathbb{R}\mathbf{1} \ : \ |z_i - z_j| \ \leq \ d_{ij} \ \ \text{for all} \ \ 1 \leq i < j \leq n \, \big\}. \tag{6}$$

This $(n-1)$-dimensional polytope is the *Lipschitz polytope* of the metric space $([n], d)$. In tropical geometry, one calls P_d a *polytrope* because it is convex both classically and tropically.

The polytope P_d^* that is dual to P_d lies in the hyperplane perpendicular to the line $\mathbb{R}\mathbf{1}$. We call P_d^* the *root polytope* because its vertices are, up to scaling, the elements $e_i - e_j$ in the root system of Lie type A_{n-1}. More precisely, we have

$$P_d^* \ = \ \{ \, x \in \mathbb{R}^n \, : \, \max_{z \in P_d} \langle x, z \rangle \leq 1 \, \} \ = \ \text{conv} \, \big\{ \, \frac{1}{d_{ij}} (e_i - e_j) \ : \ \ 1 \leq i, j \leq n \, \big\}.$$

This is a centrally symmetric polytope since the finite metric space $([n], d)$ satisfies $d_{ij} = d_{ji}$.

Proposition 8. *The Wasserstein metric W_d on the probability simplex Δ_{n-1} is given by the polyhedral norm whose unit ball is the root polytope P_d^*.*

Proof. Fix the polyhedral norm with unit ball P_d^*. The distance between μ and ν in this norm the smallest λ such that $\mu \in \nu + \lambda P_d^*$, or, equivalently, $\frac{1}{\lambda}(\mu - \nu) \in P_d^*$. By definition of dual polytope, this minimal λ is the maximum inner product $\langle \mu - \nu, z \rangle$ over all points z in the dual $(P_d^*)^*$ of the unit ball. But this specifies the Lipschitz polytope, i.e. $(P_d^*)^* = P_d$. Hence the distance between μ and ν is equal to $W_d(\mu, \nu)$, which is the optimal value in (5). $\square$

Example 9. Let $n = 4$ and fix the finite metric space graph distance on the 4-cycle

$$d \;=\; \begin{pmatrix} 0 & 1 & 1 & 2 \\ 1 & 0 & 2 & 1 \\ 1 & 2 & 0 & 1 \\ 2 & 1 & 1 & 0 \end{pmatrix}. \tag{7}$$

The induced metric on the tetrahedron Δ_3 is given by the Lipschitz polytope

$$\begin{aligned} P_d \;&=\; \big\{\,(x_1,x_2,x_3,x_4) \in \mathbb{R}^4/\mathbb{R}\mathbf{1} \,:\, |x_1-x_2| \leq 1, |x_1-x_3| \leq 1, |x_2-x_4| \leq 1, |x_3-x_4| \leq 1\,\big\} \\ &=\; \operatorname{conv}\big\{(1,0,0,-1),(1,0,0,-1),(\tfrac{1}{2},-\tfrac{1}{2},-\tfrac{1}{2},\tfrac{1}{2}),(-\tfrac{1}{2},\tfrac{1}{2},\tfrac{1}{2},-\tfrac{1}{2}),(0,1,-1,0),(0,-1,1,0)\big\}. \end{aligned}$$

Note that this 3-dimensional polytope is an octahedron. Therefore, its dual is a cube:

$$\begin{aligned} P_d^* \;&=\; \big\{\,(y_1,y_2,y_3,y_4) \in (\mathbb{R}\mathbf{1})^\perp \,:\, |y_1-y_4| \leq 1,\, |y_2-y_3| \leq 1,\, |y_2+y_3| \leq 1\big\} \\ &=\quad \operatorname{conv}\big\{(1,-1,0,0),(1,0,-1,0),(0,1,0,-1),(0,0,1,-1) \\ &\qquad\qquad (-1,1,0,0),(-1,0,1,0),(0,-1,0,1),(0,0,-1,1)\big\}. \end{aligned}$$

This is the unit ball for the Wasserstein metric on Δ_3 that is induced by d. Measuring the distance from a point to a surface with respect to this metric is illustrated in Figure 1.

We wish to compute the Wasserstein distance from a given distribution μ to a fixed *discrete statistical model* $\mathcal{M} \subset \Delta_{n-1}$. This is the problem studied in [4, 5]. Our discussion serves as an introduction. As is customary in algebraic statistics, we assume that $\mathcal{M}$ is defined by polynomials in $\nu_1, \ldots, \nu_n$. Our task is to solve the following optimization problem:

$$W_d(\mu,\mathcal{M}) \quad := \quad \min_{\nu \in \mathcal{M}} W_d(\mu,\nu) \quad = \quad \min_{\nu\in\mathcal{M}} \max_{x \in P_d} \langle \mu - \nu, x \rangle. \tag{8}$$

Computing this quantity means solving a non-convex optimization problem. Our aim is to study this problem and propose solution strategies, using methods from geometry, algebra and combinatorics. The analogous problem for the Euclidean metric was treated earlier.

We now present a detailed case study for the tetrahedron Δ_3 whose points are joint probability distributions of two binary random variables. The 2-*bit independence model* $\mathcal{M} \subset \Delta_3$ consists of all nonnegative 2×2 matrices of rank one whose entries sum to one:

$$\begin{pmatrix} \nu_1 & \nu_2 \\ \nu_3 & \nu_4 \end{pmatrix} \;=\; \begin{pmatrix} pq & p(1-q) \\ (1-p)q & (1-p)(1-q) \end{pmatrix}, \qquad (p,q) \in [0,1]^2. \tag{9}$$

Thus, $\mathcal{M}$ is the quadratic surface in the tetrahedron Δ_3 defined by the equation $\nu_1\nu_4 = \nu_2\nu_3$. The next theorem gives the optimal value function and the solution function for this independence model. We use the Wasserstein metric W_d that was defined in Example 9.

Theorem 10. *The Wasserstein distance from a distribution $\mu \in \Delta_3$ to the surface $\mathcal{M}$ is*

$$W_d(\mu,\mathcal{M}) \;=\; \begin{cases} 2\sqrt{\mu_1}(1-\sqrt{\mu_1}) - \mu_2 - \mu_3 & \text{if } \mu_1 \geq \mu_4\,,\ \sqrt{\mu_1} \geq \mu_1+\mu_2\,,\ \sqrt{\mu_1} \geq \mu_1+\mu_3, \\ 2\sqrt{\mu_2}(1-\sqrt{\mu_2}) - \mu_1 - \mu_4 & \text{if } \mu_2 \geq \mu_3\,,\ \sqrt{\mu_2} \geq \mu_1+\mu_2\,,\ \sqrt{\mu_2} \geq \mu_2+\mu_4, \\ 2\sqrt{\mu_3}(1-\sqrt{\mu_3}) - \mu_1 - \mu_4 & \text{if } \mu_3 \geq \mu_2\,,\ \sqrt{\mu_3} \geq \mu_1+\mu_3\,,\ \sqrt{\mu_3} \geq \mu_3+\mu_4, \\ 2\sqrt{\mu_4}(1-\sqrt{\mu_4}) - \mu_2 - \mu_3 & \text{if } \mu_4 \geq \mu_1\,,\ \sqrt{\mu_4} \geq \mu_2+\mu_4\,,\ \sqrt{\mu_4} \geq \mu_3+\mu_4, \\ |\mu_1\mu_4 - \mu_2\mu_3|/(\mu_1+\mu_2) & \text{if } \mu_1 \geq \mu_4,\ \mu_2 \geq \mu_3,\ \mu_1{+}\mu_2 \geq \sqrt{\mu_1},\ \mu_1{+}\mu_2 \geq \sqrt{\mu_2}, \\ |\mu_1\mu_4 - \mu_2\mu_3|/(\mu_1+\mu_3) & \text{if } \mu_1 \geq \mu_4,\ \mu_3 \geq \mu_2,\ \mu_1{+}\mu_3 \geq \sqrt{\mu_1},\ \mu_1{+}\mu_3 \geq \sqrt{\mu_3}, \\ |\mu_1\mu_4 - \mu_2\mu_3|/(\mu_2+\mu_4) & \text{if } \mu_4 \geq \mu_1,\ \mu_2 \geq \mu_3,\ \mu_2{+}\mu_4 \geq \sqrt{\mu_4},\ \mu_2{+}\mu_4 \geq \sqrt{\mu_2}, \\ |\mu_1\mu_4 - \mu_2\mu_3|/(\mu_3+\mu_4) & \text{if } \mu_4 \geq \mu_1,\ \mu_3 \geq \mu_2,\ \mu_3{+}\mu_4 \geq \sqrt{\mu_4},\ \mu_3{+}\mu_4 \geq \sqrt{\mu_3}. \end{cases}$$

The solution function $\Delta_3 \to \mathcal{M}$, $\mu \mapsto \nu^*(\mu)$ is given (with the same case distinction) by

$$\nu^*(\mu) \;=\; \begin{cases} \big(\mu_1\,,\ \sqrt{\mu_1}-\mu_1\,,\ \sqrt{\mu_1}-\mu_1\,,\ -2\sqrt{\mu_1}+\mu_1+1\big), \\ \big(\sqrt{\mu_2}-\mu_2\,,\ \mu_2\,,\ -2\sqrt{\mu_2}+\mu_2+1\,,\ \sqrt{\mu_2}-\mu_2\big), \\ \big(\sqrt{\mu_3}-\mu_3\,,\ -2\sqrt{\mu_3}+\mu_3+1\,,\ \mu_3\,,\ \sqrt{\mu_3}-\mu_3\big), \\ \big(-2\sqrt{\mu_4}+\mu_4+1\,,\ \sqrt{\mu_4}-\mu_4\,,\ \sqrt{\mu_4}-\mu_4\,,\ \mu_4\big), \\ \big(\mu_1\,,\ \mu_2\,,\ \mu_1(\mu_3{+}\mu_4)/(\mu_1{+}\mu_2)\,,\ \mu_2(\mu_3{+}\mu_4)/(\mu_1{+}\mu_2)\big), \\ \big(\mu_1\,,\ \mu_1(\mu_2{+}\mu_4)/(\mu_1{+}\mu_3)\,,\ \mu_3\,,\ \mu_3(\mu_2{+}\mu_4)/(\mu_1{+}\mu_3)\big), \\ \big(\mu_2(\mu_1{+}\mu_3)/(\mu_2{+}\mu_4)\,,\ \mu_2\,,\ \mu_4(\mu_1{+}\mu_3)/(\mu_2{+}\mu_4)\,,\ \mu_4\big), \\ \big(\mu_3(\mu_1{+}\mu_2)/(\mu_3{+}\mu_4)\,,\ \mu_4(\mu_1{+}\mu_2)/(\mu_3{+}\mu_4)\,,\ \mu_3\,,\ \mu_4\big). \end{cases}$$

The boundaries separating the various cases are given by the surfaces $\{\mu \in \Delta_3 : \mu_1 - \mu_4 = 0, \mu_1{+}\mu_2 \geq \sqrt{\mu_1}, \mu_1{+}\mu_3 \geq \sqrt{\mu_1}\}$ and $\{\mu \in \Delta_3 : \mu_2 - \mu_3 = 0, \mu_1{+}\mu_2 \geq \sqrt{\mu_2}, \mu_2{+}\mu_4 \geq \sqrt{\mu_2}\}$.

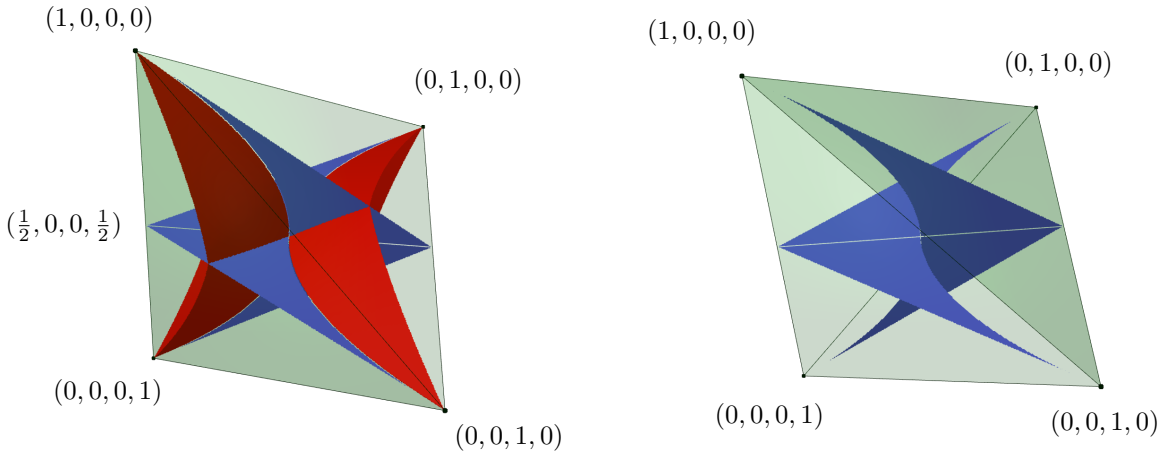

Figure 2: The optimal value function of Theorem 10 subdivides the tetrahedron of probability distributions μ (left). The surfaces that separate the various cases are shown in blue (right).

Theorem 10 involves a distinction into eight cases. This division of Δ_3 is shown in Figure 2. Each of the last four cases breaks into two subcases, since the numerator in the formulas is the absolute value of $\mu_1\mu_4 - \mu_2\mu_3$. The sign of this 2×2 determinant matters for the pieces of our piecewise algebraic function. Thus, the tetrahedron Δ_3 is divided into 12 regions on which $\mu \mapsto W_d(\mu, \mathcal{M})$ is algebraic. We now explain how to visualize Figure 2. The red surface consists of eight pieces that, together with the blue surface, separate the eight cases (this surface is not the model). Four convex regions are enclosed between the red surfaces and the edges they meet. These regions represent the first four cases in Theorem 10. For instance, the region containing the points $(1,0,0,0), (1/2,0,0,1/2)$ corresponds to the first case. The remaining four regions are each bounded by two red and two blue pieces, and correspond to the last four cases. Each of these four regions is further split in two by the model which we do not depict for the sake of visualization. The two sides are determined by the sign of the determinant $\mu_1\mu_4 - \mu_2\mu_3$. The two blue surfaces in the right figure separate the various cases. These specify the points $\mu \in \Delta_3$ with more than one optimal solution. For the proof of Theorem 10 and a simpler example see [5]. For further details see [4].

Returning to the general case, suppose that $\mathcal{M}$ is a smooth variety in $\Delta_{n-1} \subset \mathbb{R}^n$. For any point $\nu \in \Delta_{n-1}$, we seek its distance to $\mathcal{M}$ under our polyhedral norm. As before, the optimal point ν^* determines a unique face F of the unit ball $B = P_d^*$. Given that face F, we now characterize optimality as in Lemma 2. Let $\mathcal{F}$ be the set of all index pairs (i,j) such that the point $\frac{1}{d_{ij}}(e_i - e_j)$ is a vertex and it lies in F. Let ℓ_F be any linear functional on $\mathbb{R}^m$ that attains its maximum over B at F. We work in the linear space spanned by the face:

$$L_F \;\; = \;\; \Big\{ \sum_{(i,j)\in\mathcal{F}} \lambda_{ij}(e_i - e_j) \;\; : \;\; \lambda_{ij} \in \mathbb{R} \Big\}. \tag{10}$$

The point ν^* on $\mathcal{M}$ that is closest to μ is the solution of the following optimization problem:

$$\text{Minimize } \; \ell_F = \ell_F(\nu) \;\; \text{subject to } \; \nu \in (\mu + L_F) \cap \mathcal{M}. \tag{11}$$

This is a polynomial optimization problem in the linear subspace L_F of $\mathbb{R}^{n-1}$. With the notation in (10), the decision variables are λ_{ij} for $(i,j) \in \mathcal{F}$. The algebraic complexity of this problem is given by the polar degree (Theorem 5). The combinatorial complexity of (11) is governed by the facial structure of the Wasserstein ball $B = P_d^*$ associated to a finite metric space $([n], d)$. We now focus on the polar dual, which is the $(n-1)$-dimensional Lipschitz polytope $B^* = P_d$. This polytope lives in $\mathbb{R}^n/\mathbb{R}\mathbf{1} \simeq \mathbb{R}^{n-1}$, and is defined in (6).

In the study of independence models $\mathcal{M} \subset \Delta_{n-1}$, the following metrics $([n], d)$ arise:

- The discrete metric on any finite set $[n]$ where $d_{ij} = 1$ for distinct i, j.
- The L_0-metric on the Cartesian product $[m_1] \times \cdots \times [m_k]$ where $d_{ij} = \#\{l : i_l \neq j_l\}$. Here $i = (i_1, \ldots, i_k)$ and $j = (j_1, \ldots, j_k)$ are elements in that Cartesian product.
- The L_1-metric on the Cartesian product $[m_1] \times \cdots \times [m_k]$ where $d_{ij} = \sum_{l=1}^k |i_l - j_l|$.

For the last two metrics we have $n = m_1 \cdots m_k$. To compute Wasserstein distances, we need to describe the Lipschitz polytope P_d as explicitly as possible. All three metrics above are

graph metrics. This means that there exists an undirected simple graph G with vertex set $[n]$ such that d_{ij} is the length of the shortest path from i to j in G. The corresponding Wasserstein balls are called *symmetric edge polytopes*. They are studied in [5, [Section 4]].

The following four independence models are used for the case studies in [5, Section 6]. We use the tuple $((m_1)_{d_1}, \ldots, (m_k)_{d_k})$ to denote the independence model with $n = \prod_{i=1}^{k} \binom{m_i+d_i-1}{d_i}$ states where the ith entry $(m_i)_{d_i}$ refers to a multinomial distribution with m_i possible outcomes and d_i trials, which can be interpreted as an unordered set of d_i identically distributed random variables on $[m_i] = \{1, 2, ..., m_i\}$. The subscript d_i is omitted if $d_i = 1$. For example, $(2_2, 2)$ is the independence model for three binary random variables where the first two are identically distributed. We list the $n = 6$ states in the order $00, 10, 20, 01, 11, 21$. These are the vertices of the associated graph G, which is the product of a 3-chain and a 2-chain. This model $\mathcal{M}$ is the image of the map from the square $[0, 1]^2$ into the simplex Δ_5 given by

$$(p, q) \mapsto \big(p^2 q,\, 2p(1-p)q,\, (1-p)^2 q,\, p^2(1-q),\, 2p(1-p)(1-q),\, (1-p)^2(1-q) \big). \quad (12)$$

Example 11. We consider the 3-bit model $(2, 2, 2)$ with the L_0-metric on $[2]^3$; the model $(3, 3)$ for two ternary variables with the L_1-metric on $[3]^2$; the model (2_6) for six identically distributed binary variables with the discrete metric on $[7]$; the model $(2_2, 2)$ in (12) with the L_1-metric on $[3] \times [2]$. In Table 1, we report the f-vectors of their Wasserstein balls.

$\mathcal{M}$	n	$\dim(\mathcal{M})$	Metric d	f-vector of the $(n-1)$-polytope P_d^*
$(2, 2, 2)$	8	3	$L_0 = L_1$	$(24, 192, 652, 1062, 848, 306, 38)$
$(3, 3)$	9	4	L_1	$(24, 216, 960, 2298, 3048, 2172, 736, 82)$
(2_6)	7	1	discrete	$(42, 210, 490, 630, 434, 126)$
$(2_2, 2)$	6	2	L_1	$(14, 60, 102, 72, 18)$

Table 1: f-vectors of the Wasserstein balls for the four models in Example 11.

Independence models correspond in algebraic geometry to *Segre-Veronese varieties*. They are of considerable current interest study of tensor decompositions. We here replace the model, which is a semialgebraic set inside a simplex, by its complex Zariski closure in a projective space. This allows us to compute the algebraic degrees of our optimization problem.

The Segre-Veronese variety $\mathcal{M} = ((m_1)_{d_1}, \ldots, (m_k)_{d_k})$ is the embedding of $\mathbb{P}^{m_1-1} \times \cdots \times \mathbb{P}^{m_k-1}$ in the projective space of partially symmetric tensors $\mathbb{P}(\mathrm{Sym}_{d_1}\mathbb{R}^{m_1} \otimes \cdots \otimes \mathrm{Sym}_{d_k}\mathbb{R}^{m_k})$. That projective space equals $\mathbb{P}^{n-1}$ where $n = \prod_{i=1}^{k} \binom{m_i+d_i-1}{d_i}$. By definition, the Segre-Veronese variety $\mathcal{M}$ is the set of all tensors of rank one inside this projective space.

Example 12. Let $k = 2$. The Segre-Veronese variety $\mathcal{M}\big((2)_2, (2)_1\big)$ is an embedding of $\mathbb{P}^1 \times \mathbb{P}^1$ into $\mathbb{P}^5$, where it is a quartic surface. Its points are rank one tensors of format $2 \times 2 \times 2$ which are symmetric in the first two indices. This model appears in the last row of Table 3.

We identify the variety $\mathcal{M}$ with its real nonnegative points with the simplex Δ_{n-1}. The independence model $\mathcal{M}$ consists of nonnegative rank one tensors whose entries sum to 1. The dimension of $\mathcal{M}$ is denoted $\mathbf{m} := (m_1 - 1) + \cdots + (m_k - 1)$. The computation of the polar degrees of $\mathcal{M}$ appears in the doctoral dissertation of Luca Sodomaco [6, Chapter 6].

Theorem 13 (Sodomaco). *The polar degrees of the Segre-Veronese variety are*

$$\delta_{r-1}(\mathcal{M}) \;=\; \sum_{s=0}^{\mathbf{m}-n+1+r} (-1)^s \binom{\mathbf{m}-s+1}{n-r} (\mathbf{m}-s)! \left(\sum_{i_1+\cdots+i_k=s} \prod_{l=1}^{k} \frac{\binom{m_l}{i_l} d_l^{m_l-1-i_l}}{(m_l-1-i_l)!} \right). \qquad (13)$$

Here r is any integer in the range $n-1-\dim(\mathcal{M}) \leq r \leq \dim(\mathcal{M}^*)$.

We next examine this formula for various special cases starting with the binary case.

Corollary 14. *Let $\mathcal{M}$ be the k-bit independence model. The formula (13) specializes to*

$$\delta_{r-1}(\mathcal{M}) \;=\; \sum_{s=0}^{k-2^k+1+r} (-1)^s \binom{k+1-s}{2^k-r} (k-s)!\, 2^s \binom{k}{s}. \qquad (14)$$

In algebraic geometry language, our model $\mathcal{M}$ here is the Segre embedding of $(\mathbb{P}^1)^k$ into $\mathbb{P}^{2^k-1}$. This is the toric variety associated with the k-cube, so its degree is the normalized volume of the cube, which is $k!$. The polar degrees δ_{r-1} in (14) are shown for $k \leq 7$ in Table 2. The indices r with $\delta_{r-1} \neq 0$ range from $\operatorname{codim}(\mathcal{M}) = 2^k-1-k$ to $\dim(\mathcal{M}^*) = 2^k-1$. For the sake of the table's layout, we shift the indices on each row so that the row labeled 0 contains $\delta_{\operatorname{codim}(\mathcal{M})-1} = \operatorname{degree}(\mathcal{M}) = k!$. The dual variety $\mathcal{M}^*$ is a hypersurface of degree δ_{2^k-2} known as the *hyperdeterminant* of format 2^k. For instance, for $k=3$, this hypersurface in $\mathbb{P}^7$ is the $2\times 2\times 2$-hyperdeterminant which has degree four. The entries in the first column ($k=2$) corresponds to the three scenarios in Figure 1, where the algebraic degree equals 2.

$r-\operatorname{codim}(\mathcal{M})$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$
0	2	6	24	120	720	5040
1	2	12	72	480	3600	30240
2	2	12	96	840	7920	80640
3		4	64	800	9840	124320
4			24	440	7440	120960
5				128	3408	75936
6					880	30016
7						6816

Table 2: The polar degrees $\delta_{r-1}(\mathcal{M})$ of the k-bit independence model for $k \leq 7$.

We briefly discuss the independence models (m_1, m_2) for two random variables. These are the classical contingency tables of format $m_1 \times m_2$. Here, $n = m_1 m_2$ and $\mathbf{m} = m_1 + m_2 - 2$. The Segre variety $\mathcal{M} = \mathbb{P}^{m_1-1} \times \mathbb{P}^{m_2-1} \subset \mathbb{P}^{n-1}$ consists of $m_1 \times m_2$ matrices of rank one.

Corollary 15. *The Segre variety of $m_1 \times m_2$ matrices of rank one has the polar degrees*

$$\delta_{r-1}(\mathcal{M}) \;=\; \sum_{s=0}^{\mathbf{m}-n+1+r} (-1)^s \binom{\mathbf{m}-s+1}{n-r} (\mathbf{m}-s)! \left(\sum_{i+j=s} \frac{\binom{m_1}{i}}{(m_1-1-i)!} \cdot \frac{\binom{m_2}{j}}{(m_2-1-j)!} \right).$$

The polar degrees above serve as upper bounds for any particular Wasserstein distance problem. For a fixed model $\mathcal{M}$, the equality in Theorem 5 holds only when the data (ℓ, L) is generic. However, for the optimization problem in (11), the linear space $L = L_F$ and the linear functional $\ell = \ell_F$ are very specific. They depend on the Lipschitz polytope P_d and the type F of the optimal solution ν^*. For such specific scenarios, we only get an inequality.

Proposition 16. *Consider the problem (11) for the independence model $((m_1)_{d_1}, \ldots, (m_k)_{d_k})$ with a given face F of the Wasserstein ball $B = P_d^*$. The degree of the optimal solution ν^* as an algebraic function of the data μ is bounded above by the polar degree δ_{r-1} in (13).*

Proof. This follows from Theorem 5. The upper bound relies on general principles of algebraic geometry. Namely, the graph of the map $\mu \mapsto \nu^*(\mu)$ is an irreducible variety, and we are interested in its degree over μ. The map depends on the parameters (ℓ, L). When the coordinates of L and ℓ are independent transcendentals then the algebraic degree is the polar degree δ_{r-1}. That algebraic degree can only go down when these coordinates take on special values in the real numbers. That same semi-continuity argument holds for most polynomial optimization problems, including the Euclidean distance optimization in the last lecture. □

We now examine the drop in algebraic degree for the four models in Example 11. In the language of algebraic geometry, they are the Segre threefold $\mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1$ in $\mathbb{P}^7$, the variety $\mathbb{P}^2 \times \mathbb{P}^2$ of rank one 3×3 matrices in $\mathbb{P}^8$, the rational normal curve $\mathbb{P}^1$ in $\mathbb{P}^6 = \mathbb{P}(\mathrm{Sym}_6(\mathbb{R}^2))$, and the Segre-Veronese surface $\mathbb{P}^1 \times \mathbb{P}^1$ in $\mathbb{P}^5 = \mathbb{P}(\mathrm{Sym}_2(\mathbb{R}^2) \times \mathrm{Sym}_1(\mathbb{R}^2))$. The finite metrics d are specified in the fourth column of Table 1. The fifth column records the combinatorial complexity of our optimization problem, while the algebraic complexity is recorded in Table 3.

$\mathcal{M}$	Polar degrees	Maximal degree	Average degree
$(2,2,2)$	$(0,0,0,6,12,12,4)$	$(0,0,0,4,12,6,0)$	$(0,0,0,2.138,6.382,3.8,0)$
$(3,3)$	$(0,0,0,6,12,12,6,3)$	$(0,0,0,2,8,6,6,0)$	$(0,0,0,1.093,3.100,4.471,6.0,0)$
(2_6)	$(0,0,0,0,6,10)$	$(0,0,0,0,6,5)$	$(0,0,0,0,6,5)$
$(2_2,2)$	$(0,0,4,6,4)$	$(0,0,3,5,2)$	$(0,0,2.293,3.822,2.0)$

Table 3: The algebraic degrees of the problem (8) for the four models in Example 11.

The second column in Table 3 gives the vector $(\delta_0, \delta_1, \ldots, \delta_{n-2})$ of polar degrees. The third and fourth column are the results of a computational experiment. For each model, we take 1000 uniform samples μ with rational coordinates from Δ_{n-1}, and we solve the optimization problem (8). The output is an exact representation of the optimal solution ν^*. This includes the optimal face F that specifies ν^*, along with its maximal ideal over $\mathbb{Q}$. The algebraic degree of the optimal solution ν^* is computed as the number of complex zeros of that maximal ideal. This number is bounded above by the polar degree (cf. Proposition 16). The fourth column in Table 3 shows the average of the algebraic degrees we found. For example, for the 3-bit model $(2,2,2)$ we have $\delta_3 = 6$, corresponding to P_d^* touching $\mathcal{M}$ at a 3-face F, but the maximum degree we saw was 4, with an average degree of 2.138. For 4-faces F, we have $\delta_4 = 12$, and this degree was attained in some runs. The average was 6.382.

Such computational experiments are organized naturally into three stages: (1) combinatorial preprocessing, (2) numerical optimization, and (3) algebraic postprocessing. Our object of interest is a model $\mathcal{M}$ in the simplex Δ_{n-1}, typically one of the independence models $((m_1)_{d_1}, \ldots, (m_k)_{d_k})$ where $n = \prod_{i=1}^{k} \binom{m_i+d_i-1}{d_i}$. The state space $[n]$ is a metric space, with metric given by the matrix $d = (d_{ij})$. This matrix defines the Lipschitz polytope P_d and its dual, the Wasserstein ball P_d^*. Our first algorithm computes these combinatorial objects.

Algorithm 1 Combinatorial Preprocessing

Input: An $n \times n$ symmetric matrix $d = (d_{ij})$.
Output: A description of all facets F of the Wasserstein ball P_d^*.
Step 1: From the inequality presentation in (6), find all vertices of the Lipschitz polytope P_d. These vertices are the inner normal vectors ℓ_F to the facets F of P_d^*. Store them.
Step 2: Determine an inequality description of the cone C_F over each facet F.
Return: The list of pairs (ℓ_F, C_F), one for each vertex of the Lipschitz polytope P_d.

In [5], the software `Polymake` was used to run Algorithm 1. We next solve the optimization problem in (8), by examining each facet F of the Wasserstein ball. The problem is that in (11) but with the linear space L_F now replaced by the convex cone C_F that is spanned by F.

Algorithm 2 Numerical Optimization

Input: Model $\mathcal{M}$ and a point μ in the simplex Δ_{n-1}; complete output from Algorithm 1.
Output: The optimal solution ν^* in (8) along with its type G.
Step 1: For each facet F of the Wasserstein ball P_d^* do: Step 1.1: Apply global optimization methods to identify a point $\nu^* \in \mathcal{M}$ that minimizes $\ell_F = \ell_F(\nu)$ subject to $\nu \in (\mu + C_F) \cap \mathcal{M}$.
Step 1.2: Identify the unique face G of F whose span contains ν^* in its relative interior.
Step 1.3: Identify a basis of vectors $e_i - e_j \in C_G$ for the linear space L_G spanned by G.
Step 1.4: Store the optimal solution ν^* and a basis for the linear subspace L_G of $\mathbb{R}^n$.
Step 2: Among all candidate solutions found in Step 1, identify the solution ν^* for which the Wasserstein distance $W_d(\mu, \nu^*)$ to the given data point μ is smallest. Record its type G.
Return: The optimal solution ν^*, its associated linear space L_G, and the facet normal ℓ_G.

In [5], the software `SCIP` was used to run Algorithm **??**. `SCIP` employs sophisticated branch-and-cut strategies to solve constrained polynomial optimizaion problems via LP relaxation. The virtue of Algorithm 2 is that it is guaranteed to find the global optimum for our problem (8). Moreover, it furnishes an identification of the combinatorial type. This serves as the input to the symbolic computation in Algorithm 3 below.

Algorithm 3 Algebraic Postprocessing

Input: The optimal solution (ν^*, G) to (8) in the form found by Algorithm **??**.
Output: The maximal ideal in the polynomial ring $\mathbb{Q}[\nu_1, \ldots, \nu_n]$ which has the zero ν^*.
Step 1: Use Lagrange multipliers to give polynomial equations that characterize the critical points of the linear function ℓ_F on the subvariety $(\mu + L_G) \cap \mathcal{M}$ in the affine space $\mathbb{R}^n$.
Step 2: Eliminate all variables representing Lagrange multipliers from the ideal in Step 1.
Step 3: The ideal from Step 2 lives in $\mathbb{Q}[\nu_1, \ldots, \nu_n]$. If this ideal is maximal then call it M.
Step 4: If not, remove extraneous primary components to get the maximal ideal M of ν^*.
Step 5: Determine the degree of ν^*, which is the dimension of $\mathbb{Q}[\nu_1, \ldots, \nu_n]/M$ over $\mathbb{Q}$.
Return: Output generators for the maximal ideal M along with the degree found in Step 5.

Algorithm 3 can be carried out with `Macaulay2`. Steps 2 and 4 are the result of standard Gröbner basis calculations. The entire pipeline is illustrated with examples in [5, Section 6].

References

[1] D. Bates, J. Hauenstein, A. Sommese and C. Wampler: *Numerically solving polynomial systems with Bertini*, Software, Environments, and Tools, 25, SIAM, Philadelphia, 2013

[2] P. Breiding, K. Rose and S. Timme: Certifying zeros of polynomial systems using interval arithmetic, 2020, arXiv:2011.05000.

[3] P. Breiding and S. Timme: HomotopyContinuation.jl: A Package for Homotopy Continuation in Julia, *Math. Software – ICMS 2018*, 458–465, Springer, 2018.

[4] T. Ö. Çelik, A. Jamneshan, G. Montúfar, B. Sturmfels and L. Venturello: *Optimal transport to a variety* Mathematical Aspects of Computer and Information Sciences, MACIS 2019, Istanbul, Springer Lecture Notes in Computer Science 11989 (2020) 364–381.

[5] T. Ö. Çelik, A. Jamneshan, G. Montúfar, B. Sturmfels and L. Venturello: *Wasserstein distance to independence models*, Journal of Symbolic Computation **104** (2021) 855–873.

[6] L. Sodomaco: *The distance function from the variety of partially symmetric rank-one tensors*, PhD thesis, Universitá degli Studi di Firenze, 2020.