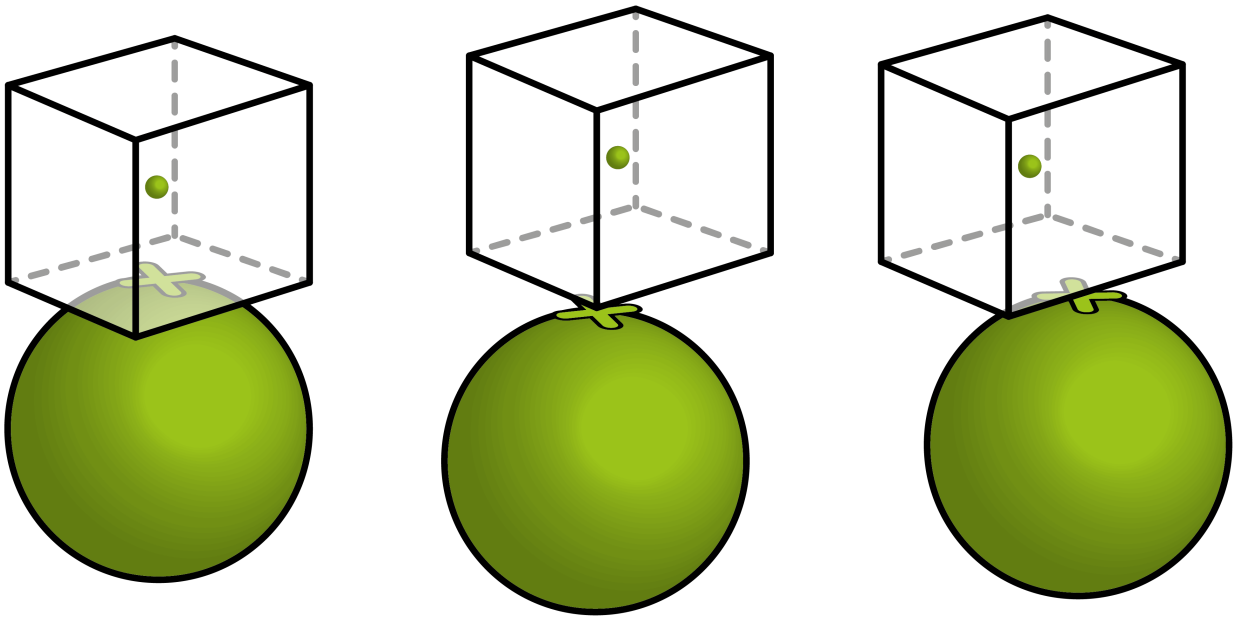
Wasserstein Distance to Independence Models

Bernd Sturmfels
MPI Leipzig and UC Berkeley



joint work with Türkü Özlüm Çelik, Asgar Jamneshan, Guido Montúfar, and Lorenzo Venturello

Wasserstein distance

Fix a symmetric $n \times n$ matrix $d = (d_{ij})$ with nonnegative entries that satisfy $d_{ii} = 0$ and $d_{ik} \leq d_{ij} + d_{jk}$ for all i, j, k.

This turns the set $[n] = \{1, 2, \ldots, n\}$ into a metric space.

Probability distributions on $[n]$ are points in the simplex

$$\Delta_{n-1} = \left\{ (\nu_1, \ldots, \nu_n) \in \mathbb{R}^n_{\geq 0} : \sum_{i=1}^n \nu_i = 1 \right\}.$$

Q: How to measure the distance between two distributions $\mu, \nu \in \Delta_{n-1}$?

Wasserstein distance

Fix a symmetric $n \times n$ matrix $d = (d_{ij})$ with nonnegative entries that satisfy $d_{ii} = 0$ and $d_{ik} \leq d_{ij} + d_{jk}$ for all i, j, k.

This turns the set $[n] = \{1, 2, \ldots, n\}$ into a metric space.

Probability distributions on $[n]$ are points in the simplex

$$\Delta_{n-1} = \left\{ (\nu_1, \ldots, \nu_n) \in \mathbb{R}^n_{\geq 0} : \sum_{i=1}^n \nu_i = 1 \right\}.$$

Q: How to measure the distance between two distributions $\mu, \nu \in \Delta_{n-1}$?

A: Solve the linear programming problem

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

Optimal value $W_d(\mu, \nu)$ is the *Wasserstein Distance* between μ and ν.

This turns the simplex Δ_{n-1} into a metric space.

$\longrightarrow$ Optimal Transport

Unit Balls

The unit ball of the Wasserstein metric is the polytope

$$B \;=\; \mathrm{conv}\left\{ \frac{1}{d_{ij}}(e_i - e_j) \;:\; 1 \le i < j \le n \right\}$$

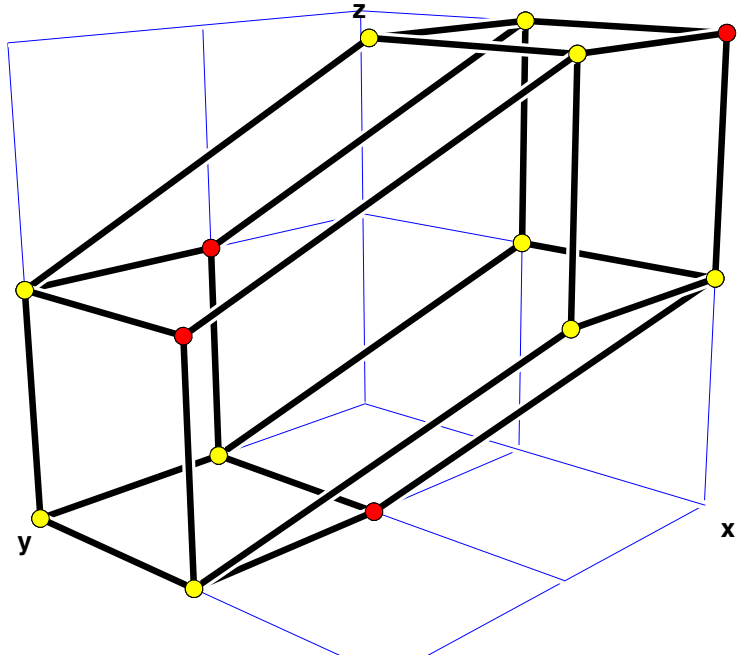
Its polar dual is the feasible region of our linear program:

$$B^* \;=\; \left\{ x \in \mathbb{R}^n/\mathbb{R}\mathbf{1} \;:\; |x_i - x_j| \;\le\; d_{ij} \;\text{ for all } i,j \right\}$$

Lipschitz polytope

polytrope
tropical polytope

alcoved polytope
of type A



Statistics

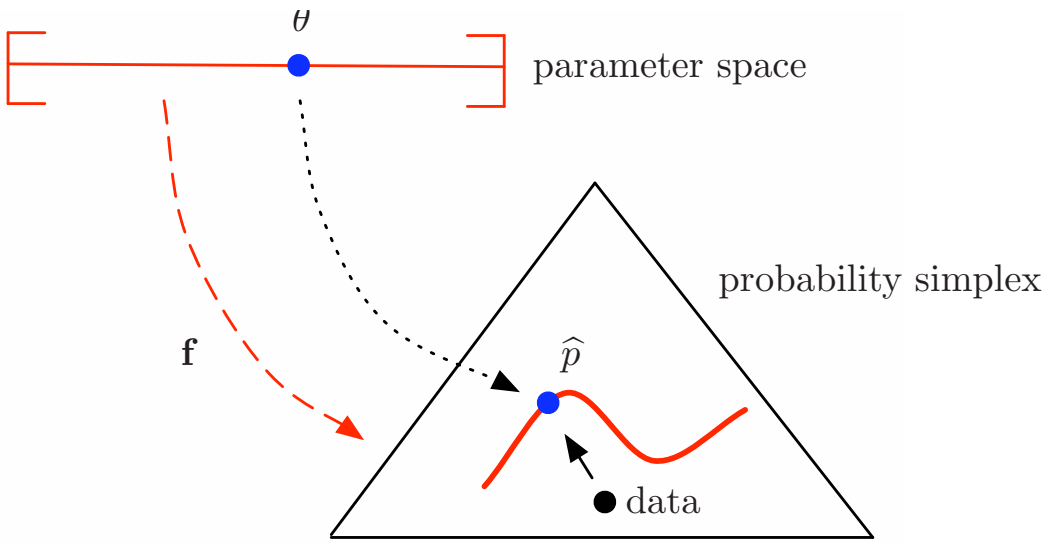


Fig. 3.2. The geometry of maximum likelihood estimation.

A *discrete statistical model* is any subset $\mathcal{M} \subset \Delta_{n-1}$.

We assume that $\mathcal{M}$ compact and defined by polynomials.

Statistics

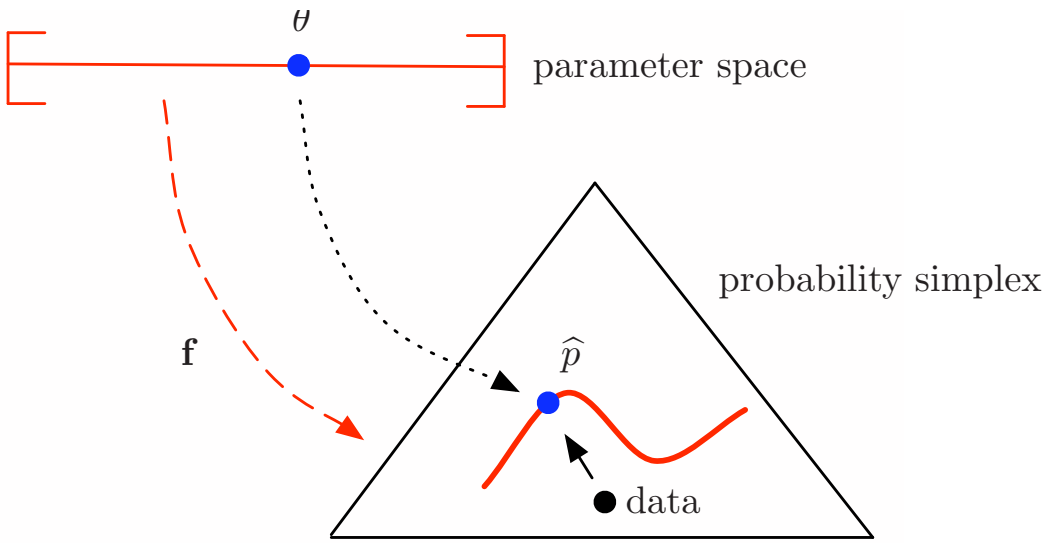


Fig. 3.2. The geometry of maximum likelihood estimation.

A *discrete statistical model* is any subset $\mathcal{M} \subset \Delta_{n-1}$.

We assume that $\mathcal{M}$ compact and defined by polynomials.

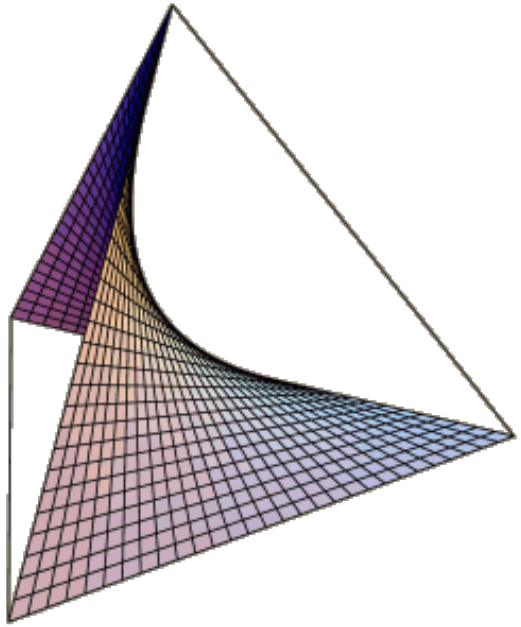
The *Wasserstein distance* from the data μ to the model $\mathcal{M}$ is

$$W_d(\mu, \mathcal{M}) \;:=\; \min_{\nu \in \mathcal{M}} W_d(\mu, \nu) \;=\; \min_{\nu \in \mathcal{M}} \max_{x \in B^*} \langle \mu - \nu, x \rangle.$$

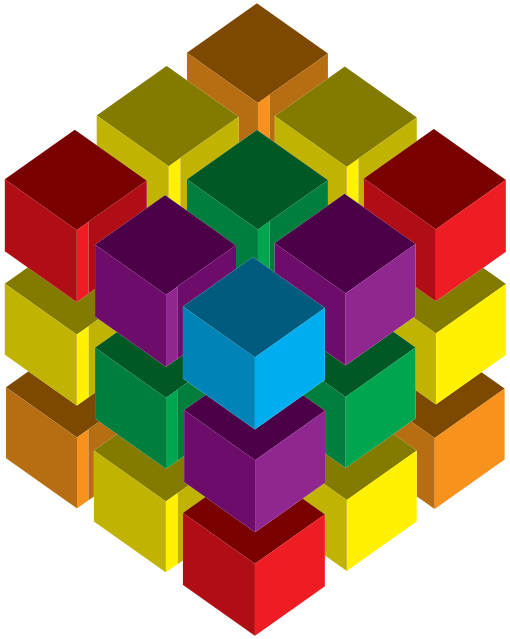
Computing this means solving a non-convex optimization problem.

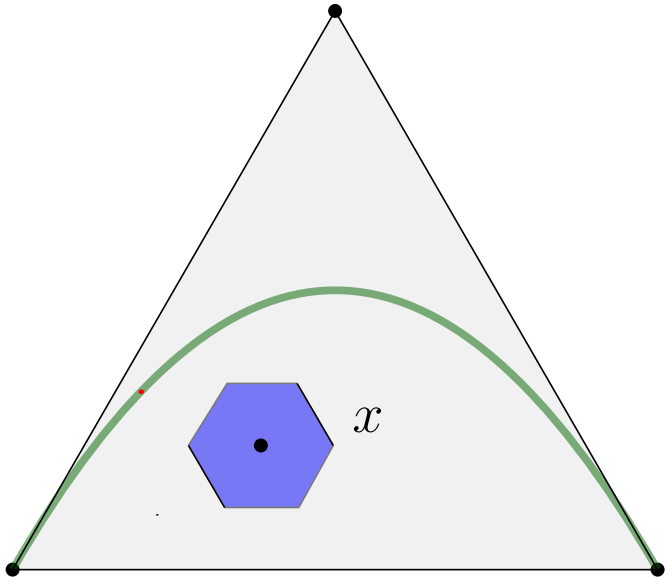
Independence

$$\begin{pmatrix} \nu_1 & \nu_2 \\ \nu_3 & \nu_4 \end{pmatrix} = \begin{pmatrix} pq & p(1-q) \\ (1-p)q & (1-p)(1-q) \end{pmatrix}$$



Independence models: $\mathcal{M} = \{$ matrices or tensors of rank one $\}$





2×2 matrices

We fix the L_0-metric d on the set of binary pairs $[2] \times [2]$. Under our identification (lexicographic order) of this state space with $[4] = \{1, 2, 3, 4\}$, the resulting metric on Δ_3 is given by the 4×4 matrix

$$d \;=\; \begin{pmatrix} 0 & 1 & 1 & 2 \\ 1 & 0 & 2 & 1 \\ 1 & 2 & 0 & 1 \\ 2 & 1 & 1 & 0 \end{pmatrix}. \tag{2.3}$$

We now present the optimal value function and the solution function for this independence model.

Theorem 2.2. *For the L_0-metric on the state space $[2] \times [2]$, the Wasserstein distance from a data distribution $\mu \in \Delta_3$ to the 2-bit independence surface $\mathcal{M}$ is given by*

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

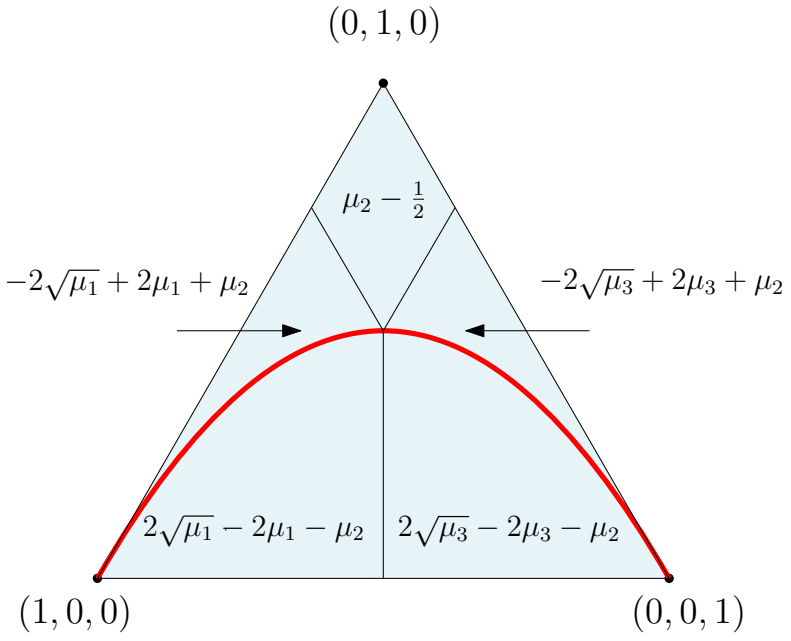
Symmetric 2×2 matrices

Theorem 2.1. *For the discrete metric and for the L_1-metric on the state space $[3] = \{1, 2, 3\}$, the Wasserstein distance from a data distribution $\mu \in \Delta_2$ to the Hardy-Weinberg curve $\mathcal{M}$ equals*

$$W_d(\mu, \mathcal{M}) = \begin{cases} |2\sqrt{\mu_1} - 2\mu_1 - \mu_2| & \text{if} \quad \mu_1 - \mu_3 \geq 0 \text{ and } \mu_1 \geq \frac{1}{4}, \\ |2\sqrt{\mu_3} - 2\mu_3 - \mu_2| & \text{if} \quad \mu_1 - \mu_3 \leq 0 \text{ and } \mu_3 \geq \frac{1}{4}, \\ \mu_2 - \frac{1}{2} & \text{if} \quad \mu_1 \leq \frac{1}{4} \text{ and } \mu_3 \leq \frac{1}{4}. \end{cases}$$

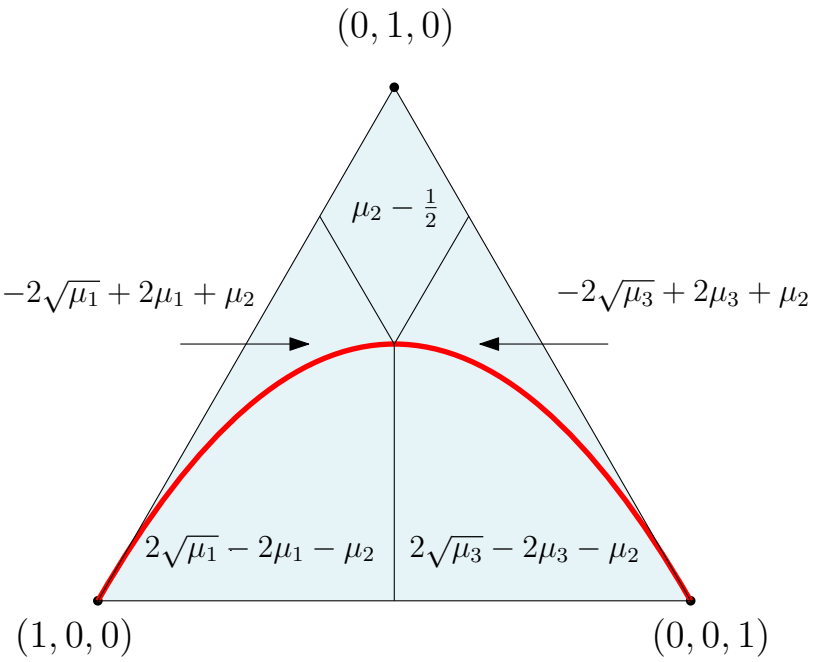
The solution function $\Delta_2 \to \mathcal{M}$, $\mu \mapsto \nu^*(\mu)$ is given (with the same case distinction) by

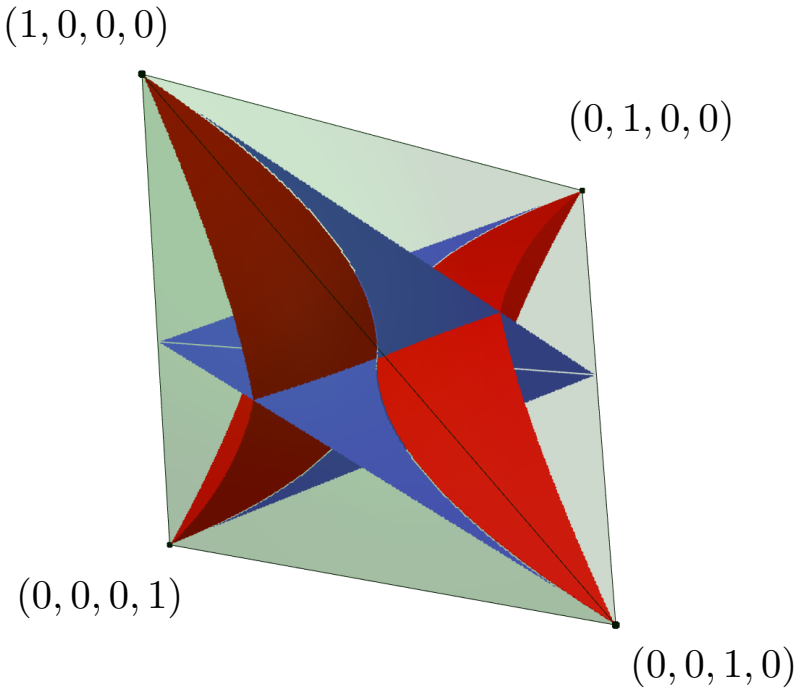
$$\nu^*(\mu) = \begin{cases} (\mu_1, 2\sqrt{\mu_1} - 2\mu_1, 1 + \mu_1 - 2\sqrt{\mu_1}), \\ (1 + \mu_3 - 2\sqrt{\mu_3}, 2\sqrt{\mu_3} - 2\mu_3, \mu_3), \\ (\frac{1}{4}, \frac{1}{2}, \frac{1}{4}). \end{cases}$$



Complexity of our Optimization Problem

The optimal value function and *solution function* are piecewise algebraic. This suggests a division of our problem into **two tasks**:
first identify all pieces, then find a formula for each piece.

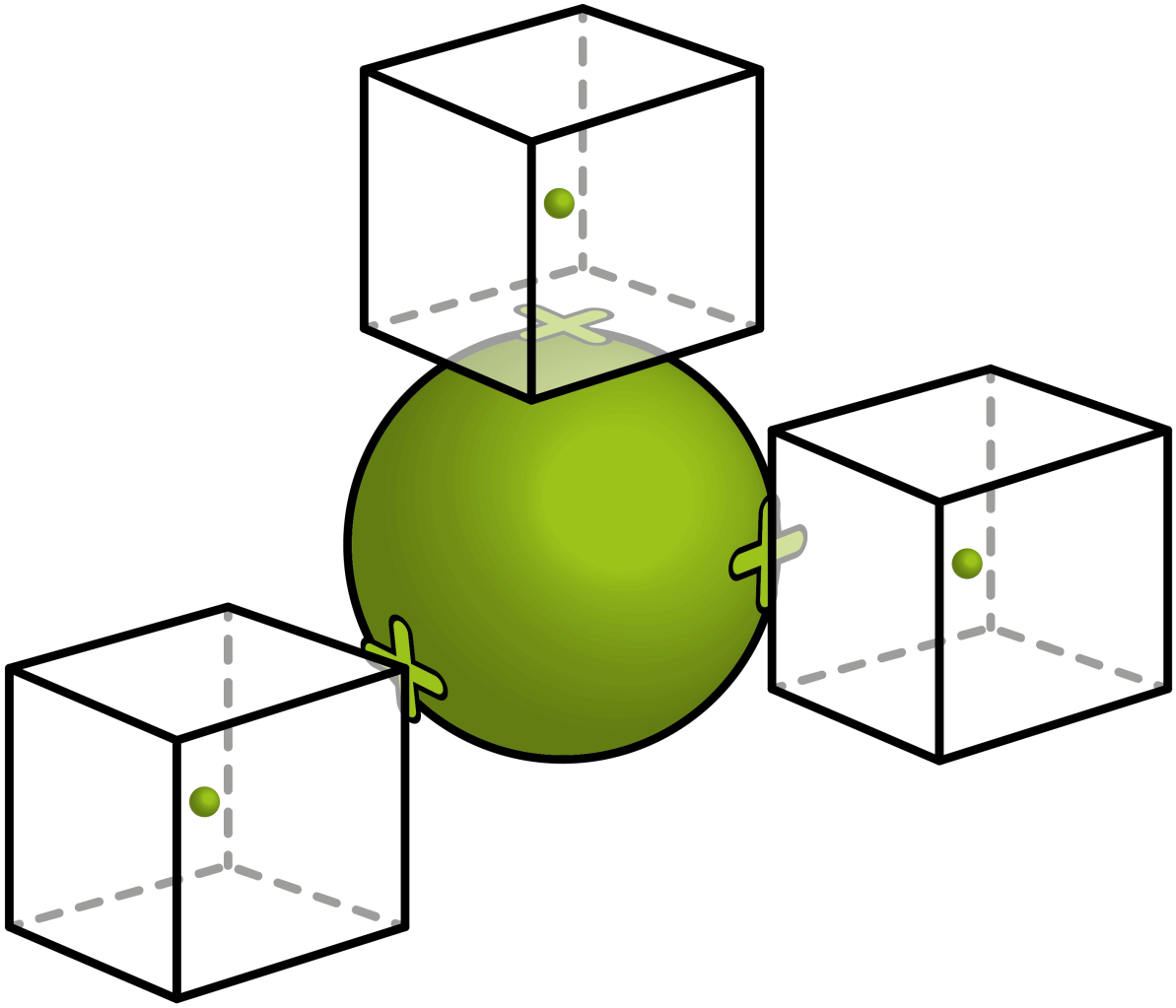




Both tasks have a high degree of complexity.

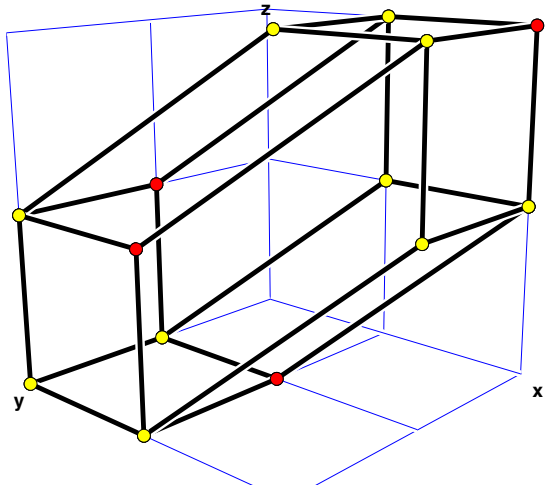
The **first task** pertains to *combinatorial complexity*,
the **second task** to *algebraic complexity*. *We address both.*

Complexity of our Optimization Problem



Combinatorial complexity: How many faces does the unit ball have?
Algebraic complexity: What is the degree of the critical variety?

Polytopes and their f-vectors



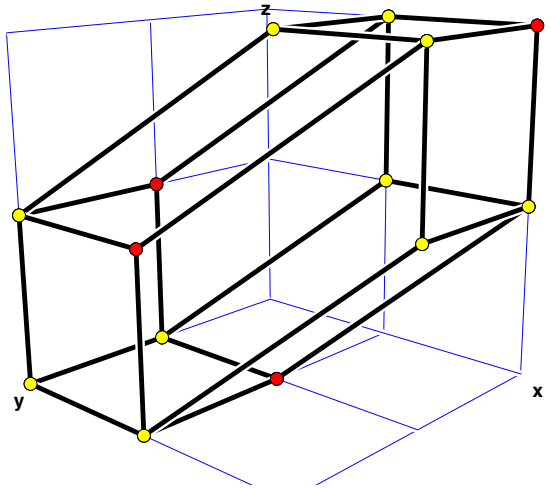
Proposition

Fix the graph metric on a graph G with vertex set $[n]$. The Lipschitz polytope is

$$B^* = \{\, x \in \mathbb{R}^n/\mathbb{R}\mathbf{1} \,:\, |x_i - x_j| \leq 1 \ \textit{for every edge}\ (i,j)\ \text{of}\ G \,\}.$$

These are the facets when G is bipartite.

Polytopes and their f-vectors



Proposition

Fix the graph metric on a graph G with vertex set $[n]$. The Lipschitz polytope is

$$B^* = \{\, x \in \mathbb{R}^n/\mathbb{R}\mathbf{1} \,:\, |x_i - x_j| \leq 1 \ \textit{for every edge}\ (i,j)\ \text{of}\ G \,\}.$$

These are the facets when G is bipartite.

Example

If G is the k-cube then the vertices of B^* are in bijection with the proper 3-colorings of G, with one vertex of fixed color. For $k = 2, 3, 4, 5, 6$, their number is 6, 38, 990, 395094, 33433683534.

We computed the unit balls B for small independence models. Their combinatorics is an interesting research direction.

Algebraic Geometry ...

Fix a smooth model $\mathcal{M}$, a linear functional ℓ, and an affine space L of dimension r, both in general position relative to $\mathcal{M}$.

Theorem

The polar degree δ_r of $\mathcal{M}$ is the algebraic degree of the problem

Minimize the linear functional ℓ over the intersection $L \cap \mathcal{M}$.

Algebraic Geometry ...

Fix a smooth model $\mathcal{M}$, a linear functional ℓ, and an affine space L of dimension r, both in general position relative to $\mathcal{M}$.

Theorem

The polar degree δ_r of $\mathcal{M}$ is the algebraic degree of the problem

Minimize the linear functional ℓ over the intersection $L \cap \mathcal{M}$.

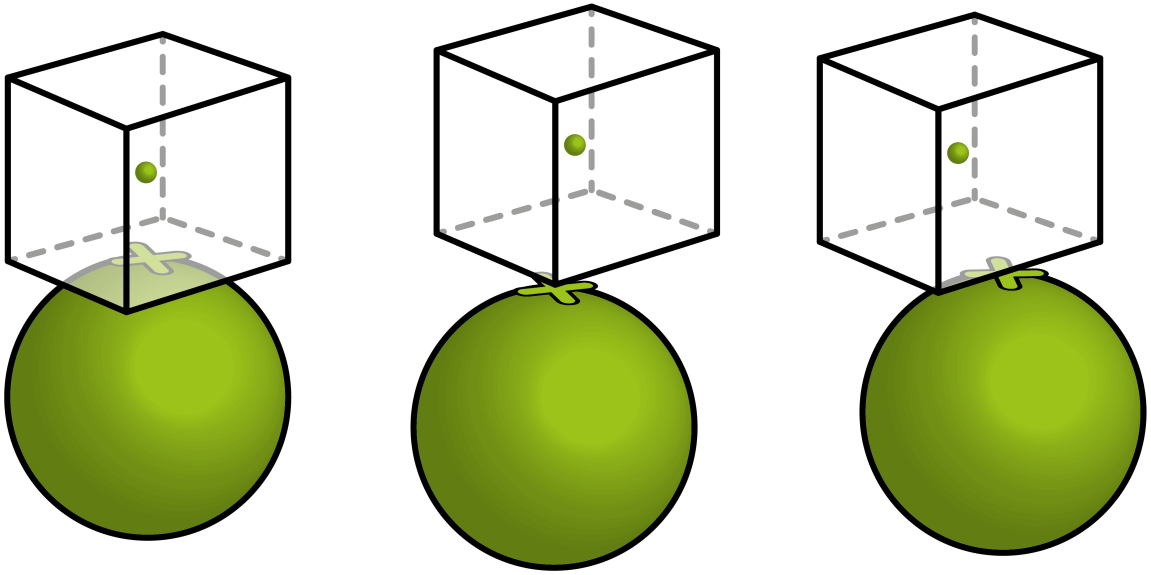
Recent work of Luca Sodomaco gives a (complicated) formula for the polar degrees of all Segre-Veronese varieties. For $(\mathbb{P}^1)^k$ we get

Corollary

If $\mathcal{M}$ is the k-bit independence model then

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

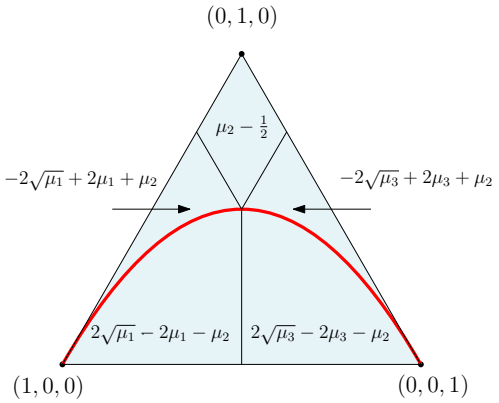
... meets Numerical Mathematics

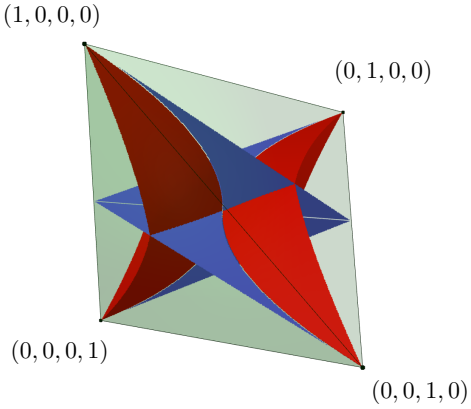


$r - \mathrm{codim}(\mathcal{M})$	$(2,3)$	$(2,4)$	$(2,5)$	$(2,6)$	$(3,3)$	$(3,4)$	$(3,5)$	$(3,6)$	$(4,4)$	$(4,5)$	$(4,6)$
0	3	4	5	6	6	10	15	21	20	35	56
1	4	6	8	10	12	24	40	60	60	120	210
2	3	4	5	6	12	27	48	75	84	190	360
3					6	16	30	48	68	176	360
4					3	6	10	15	36	105	228
5									12	40	90
6									4	10	20

TABLE 3. The polar degrees $\delta_{r-1}(\mathcal{M})$ of the independence model (m_1, m_2).

Experiments



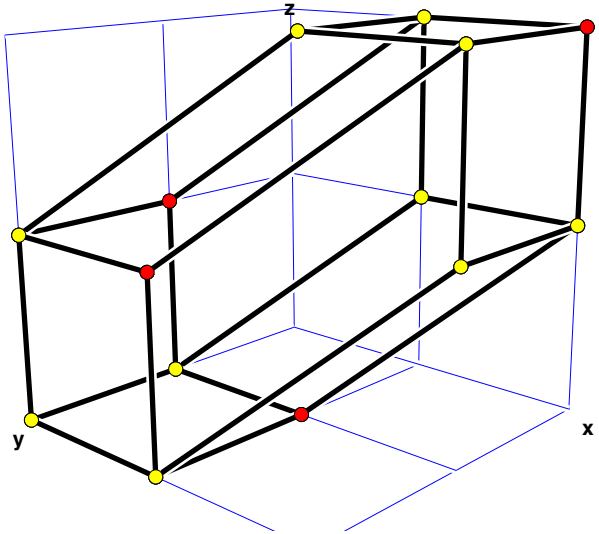


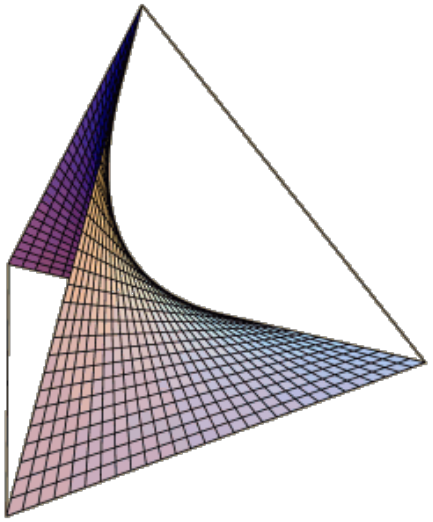
			% of opt. solutions of dim(type) $= i$						
$\mathcal{M}$	d	f-vector	0	1	2	3	4	5	6
$(2,2)$	L_0	$(8,12,6)$	68.6	31.4	0	-	-	-	-
$(2,2,2)$	L_0	$(24,192,652,1062,848,306,38)$	0	0	0.1	70.9	27.5	1.5	0
$(2,3)$	L_0	$(18,96,200,174,54)$	0	64.1	18.7	17.2	0	-	-
$(2,3)$	L_1	$(14,60,102,72,18)$	0	76.7	17.4	5.9	0	-	-
$(3,3)$	L_0	(36,468,2730,8010,12468,10200,3978,534)	0	0	0.1	58.3	28.2	4.6	8.8
$(3,3)$	L_1	$(24,216,960,2298,3048,2172,736,82)$	0	0	0	65.7	27.8	5.1	1.4
$(2,4)$	L_0	$(32,336,1464,3042,3168,1566,282)$	0	0.1	55.1	14.6	25.8	4.4	0
$(2,4)$	L_1	$(20,144,486,846,774,342,54)$	0	0	75.3	16.5	8.2	0	0
(2_3)	L_1	$(6,12,8)$	0	98.3	1.7	-	-	-	-
(2_3)	di	$(12,24,14)$	0.2	96.7	3.1	-	-	-	-
$(2_2,2)$	L_1	(14,60,102,72,18)	0	0	67.6	27.5	4.9	-	-
$(2_2,2)$	di	$(30,120,210,180,62)$	0	0.2	81.9	16.8	1.1	-	-
(3_2)	di	$(30,120,210,180,62)$	0	0.2	83.1	16.0	0.7	-	-
(2_4)	L_1	$(8,24,32,16)$	0	0.1	98.3	1.6	-	-	-
(2_4)	di	$(20,60,70,30)$	0	0	96.9	3.1	-	-	-

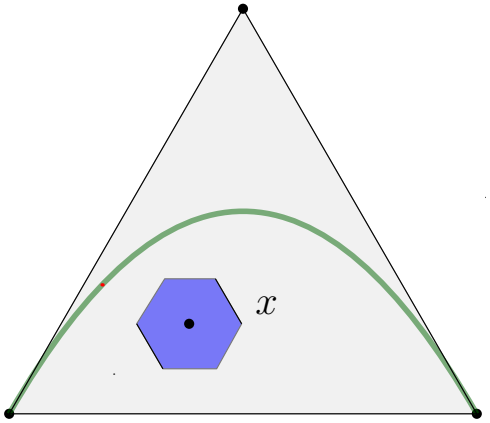
TABLE 6. Distribution of types among optimal solutions for a uniform sample of 1000 points.

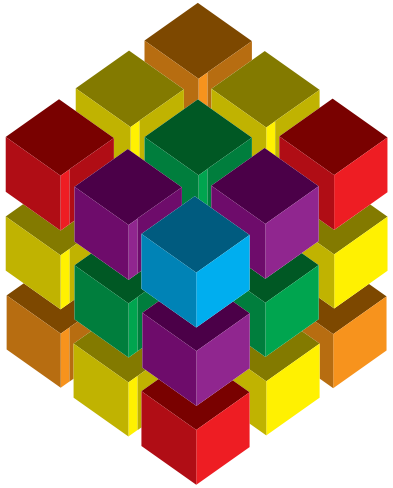
Conclusion

When Statistics, Optimization and Algebraic Geometry interact ...









... cool opportunities arise for **Algebraic Combinatorics**.