Optimal Transport to Independence Models

Guido Montúfar
UCLA and MPI MIS

joint work with T. Celik, A. Jamneshan, B. Sturmfels, L. Venturello

UC Davis MADDD seminar (online), December 2020

We seek to describe

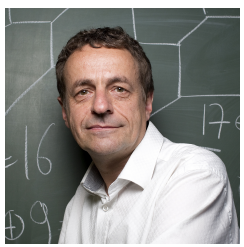
- The geometry and combinatorics of the optimization landscape in minimum optimal transport distance parameter estimation
- Where possible obtain the exact form of the estimator and devise better iterative optimization methods

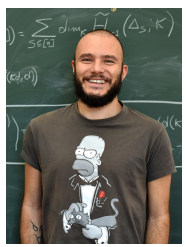
Parametric linear programming and duality of optimal transport / polyhedral norm distances to varieties.

Works [Celik et al., 2020a, Celik et al., 2020b] with


T. Celik


A. Jamneshan

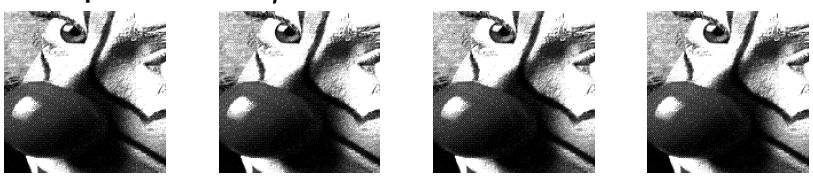

B. Sturmfels


L. Venturello

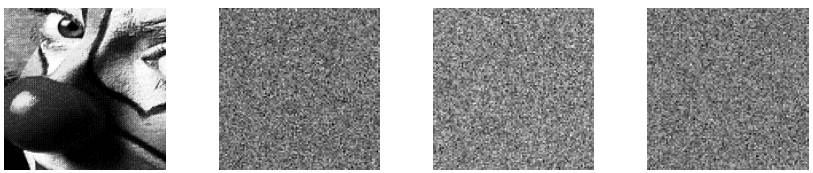
Estimating a distribution

- Given some data μ, we can select a hypothesis distribution $\nu \in \mathcal{M}$ by minimizing a discrepancy measure, as $\operatorname{arginf}_{\nu\in\mathcal{M}} D(\mu,\nu)$
- Negative log likelihood is good if we do not have any prior knowledge about the data. In general we may want to exploit its geometry

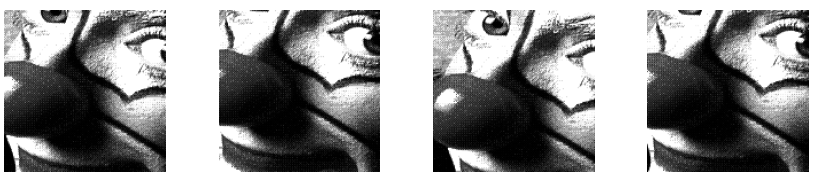
Samples from μ



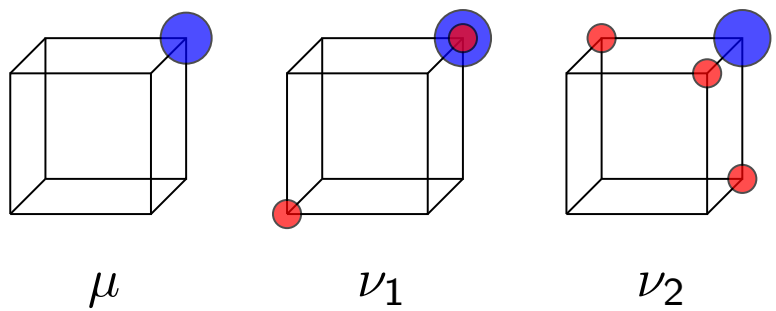
Samples from ν_1



Samples from ν_2



- Given $D\colon \mathcal{P}\times\mathcal{P}\to\mathbb{R}_{\geq 0}$.
  Estimation: $\operatorname{arginf}_{\nu\in\mathcal{M}} D(\mu,\nu)$
- Want: $D(\mu,\nu_1) > D(\mu,\nu_2)$.
  So D should account for distances between samples

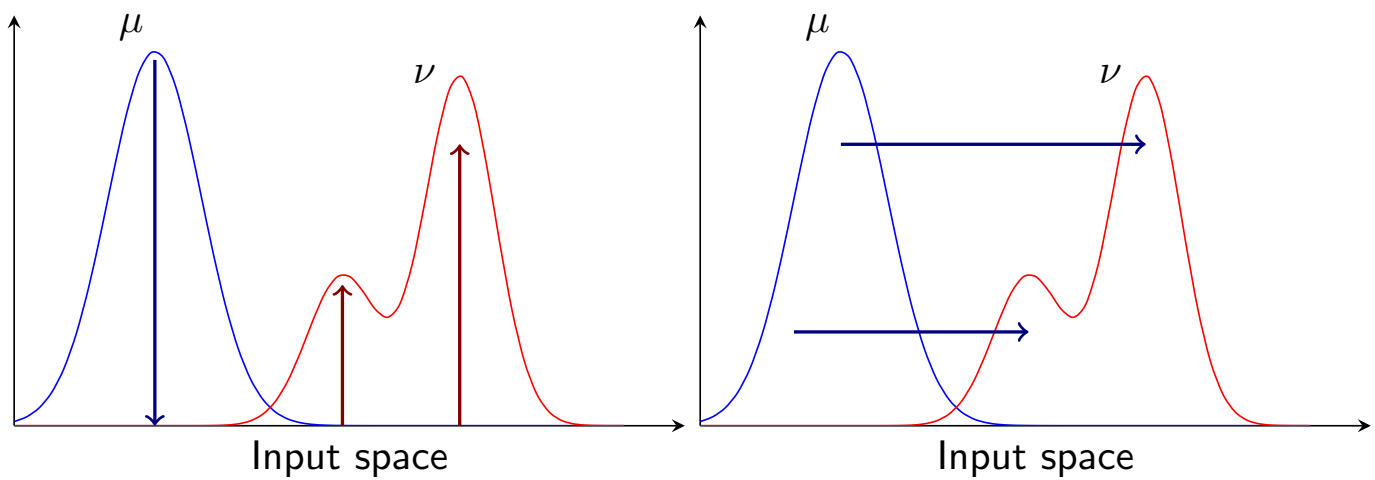


Wasserstein distance

- Consider a metric space $(\mathcal{X}, d_{\mathcal{X}})$ and the set $\mathcal{P}_p(\mathcal{X})$ of densities with finite p-th moment.
- The Wasserstein-p distance of a pair $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$ is

$$W_{p,d_{\mathcal{X}}}(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\pi}\left[d_{\mathcal{X}}(X,Y)^p\right]^{\frac{1}{p}},$$

  where $\Pi(\mu,\nu)$ is the set of distributions with margins $X\sim\mu$, $Y\sim\nu$.
- Note $W_{p,d_{\mathcal{X}}}$ depends on the *ground metric* $d_{\mathcal{X}}\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}$.

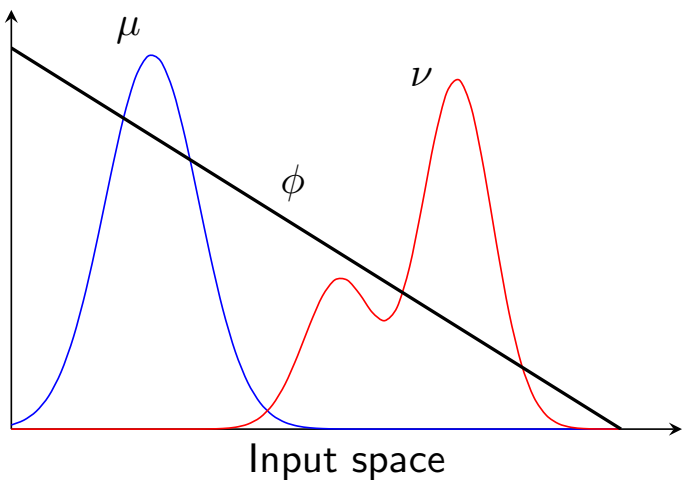


- In applications the dual formulation is useful. For $p = 1$ one has

$$W_{1,d_{\mathcal{X}}}(\mu,\nu) = \sup_{\phi \in \mathrm{Lip}_1} \mathbb{E}_\mu[\phi] - \mathbb{E}_\nu[\phi] = \sup_{\phi \in \mathrm{Lip}_1} \langle \phi, \mu - \nu \rangle,$$

  where $\mathrm{Lip}_1 = \{\phi \in \mathbb{R}^N / \mathbb{R}\mathbb{1} : |\phi_x - \phi_y| \leq d_{xy}\}$

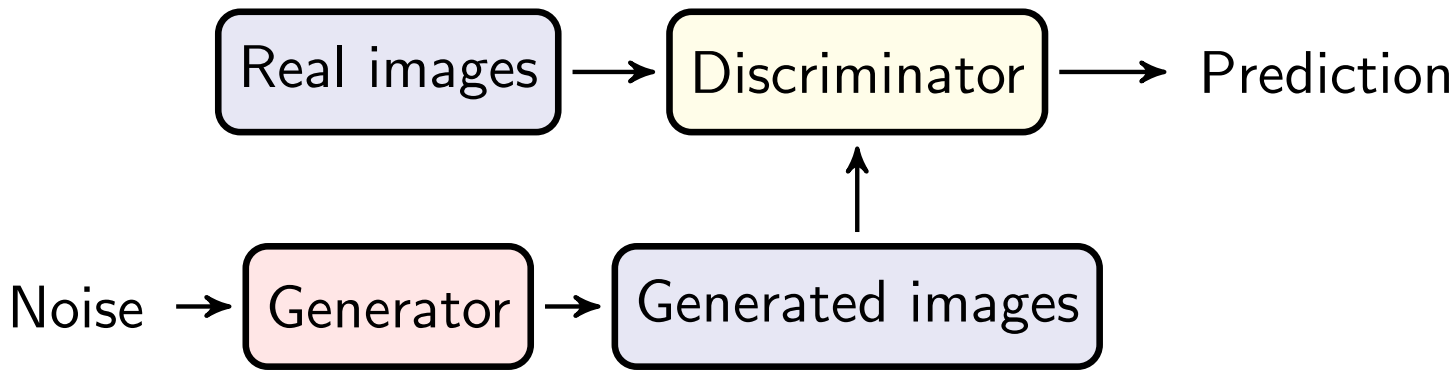
- Note the constraint $\phi \in \mathrm{Lip}_1$ depends on the ground metric $d_{\mathcal{X}}$



- The dual variable ϕ can be regarded as a *discriminator*

Generative Adversarial Networks (GANs)

- GANs can learn to generate samples that look like a given set of training examples
- Two neural networks are trained concurrently:
  the Generator $\theta \mapsto \nu(\theta)$ and the Discriminator $\vartheta \mapsto \phi(\vartheta)$



- Enormously successful in applications

What can GANs do?

Create new celebrity faces [Karras et al., 2018]

Text-to-image synthesis [Zhang et al., 2016]

Figure 3. Example results by our proposed StackGAN, GAWWN [20], and GAN-INT-CLS [22] conditioned on text descriptions from CUB test set. GAWWN and GAN-INT-CLS generate 16 images for each text description, respectively. We select the best one for each of them to compare with our StackGAN.

Image superresolution [Ledig et al., 2016]

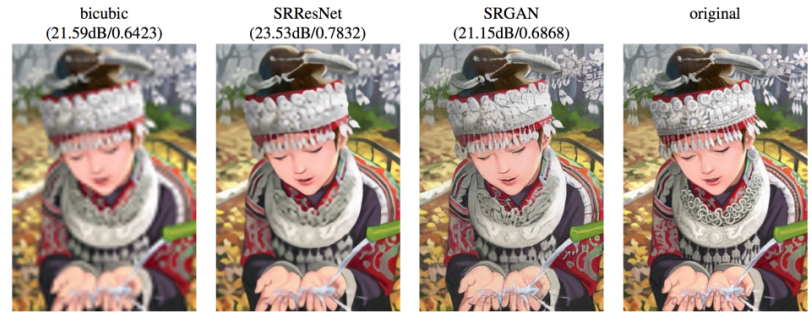
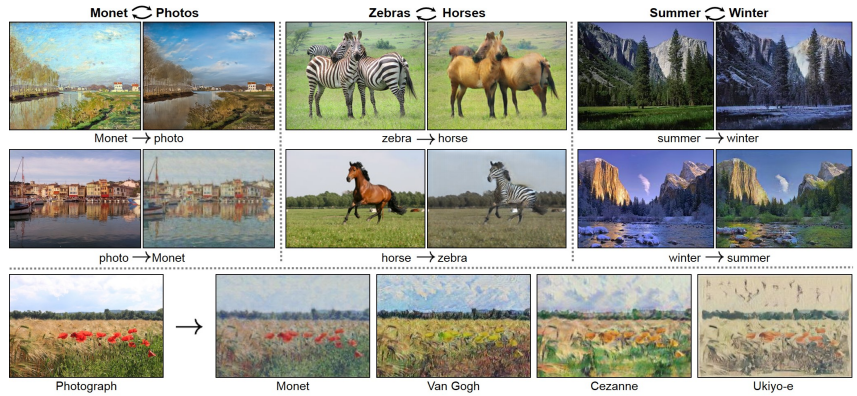


Figure 2: From left to right: bicubic interpolation, deep residual network optimized for MSE, deep residual generative adversarial network optimized for a loss more sensitive to human perception, original HR image. Corresponding PSNR and SSIM are shown in brackets. [4× upscaling]

Image-to-image translation [Zhu et al., 2017]]



Properties of the Wasserstein estimator?

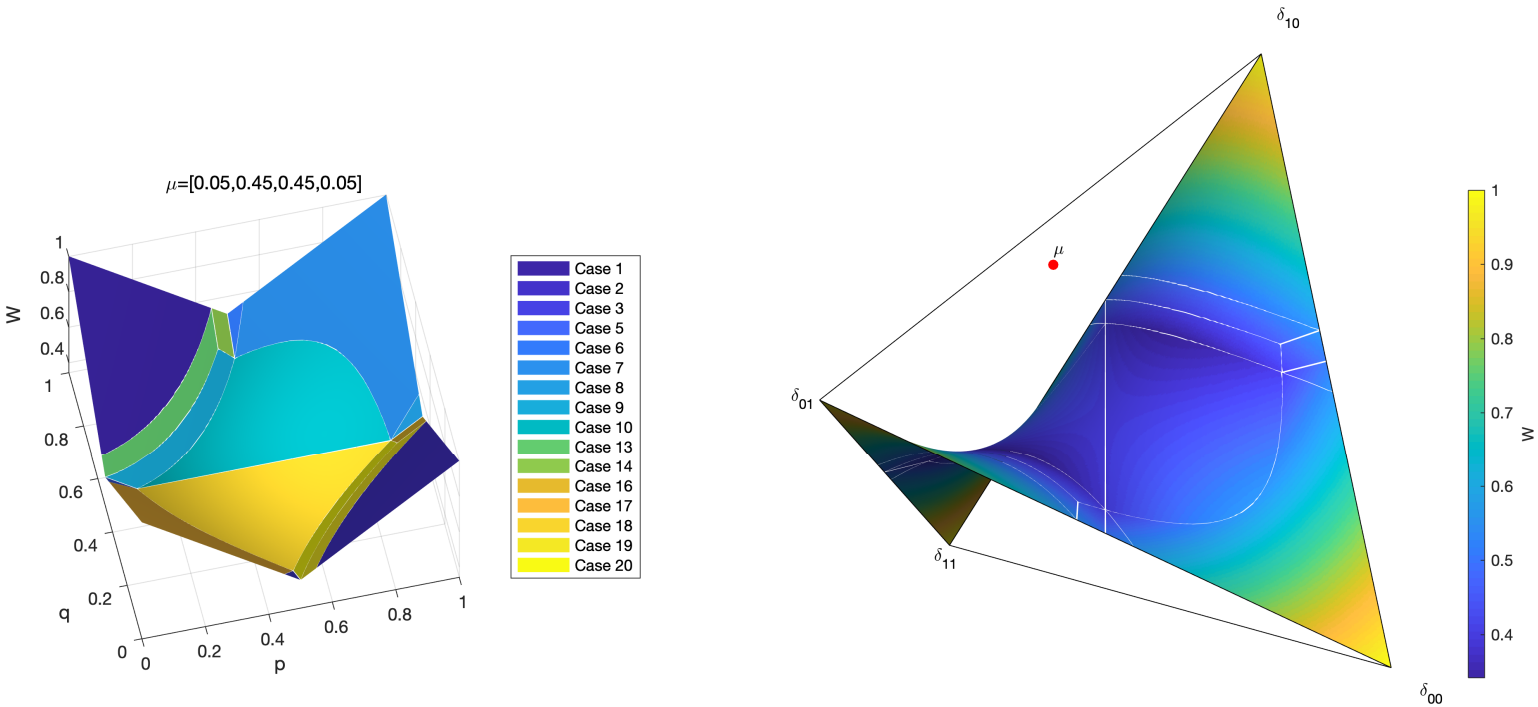
- We wish to understand the properties of the minimum Wasserstein distance estimator when $\mathcal{M}$ is an algebraic variety
- What is the structure of the functions

$$\theta \mapsto \nu \mapsto W_d(\mu, \nu) \qquad \text{(training objective)}$$
$$\mu \mapsto \nu^* = \operatorname{arginf}_{\nu \in \mathcal{M}} W_d(\mu, \nu) \qquad \text{(estimator)}$$
$$\mu \mapsto W_d(\mu, \nu^*) = W_d(\mu, \mathcal{M}) \qquad \text{(distance)}$$

- How do these depend on the metric d, target μ, model $\mathcal{M}$?
- Is the minimizer unique / can we obtain a closed formula?

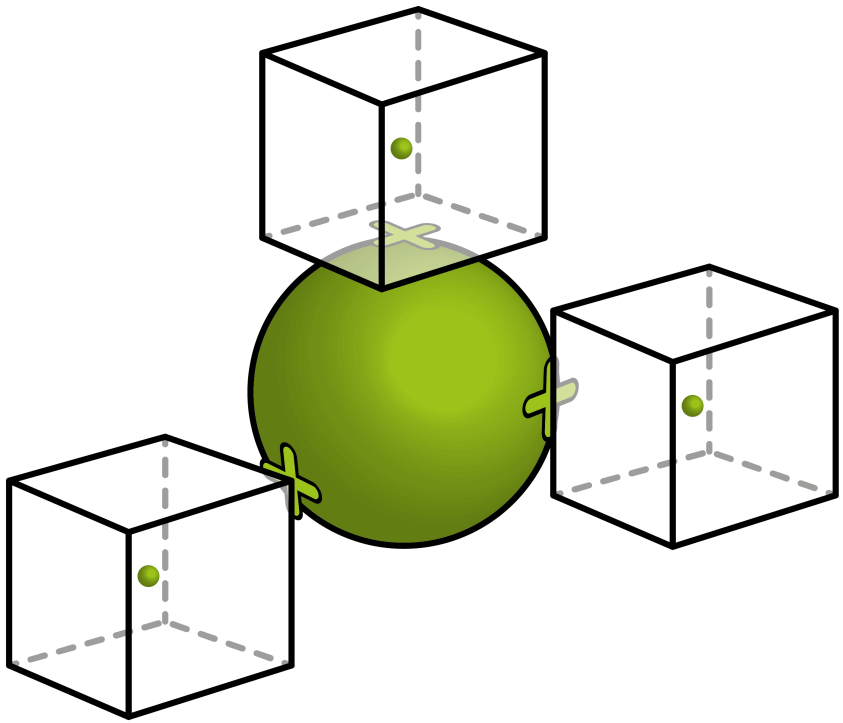
MLE for exponential families is convex in θ and in some cases we have closed formulas

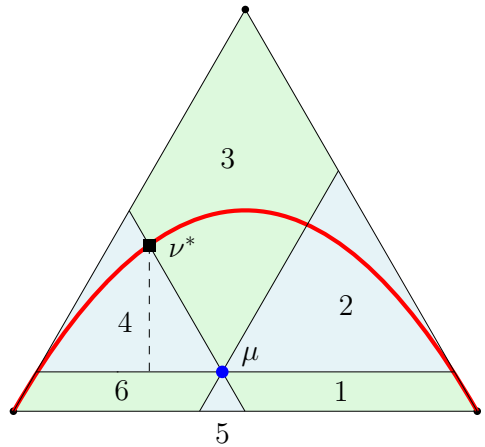
Optimal transport to a variety

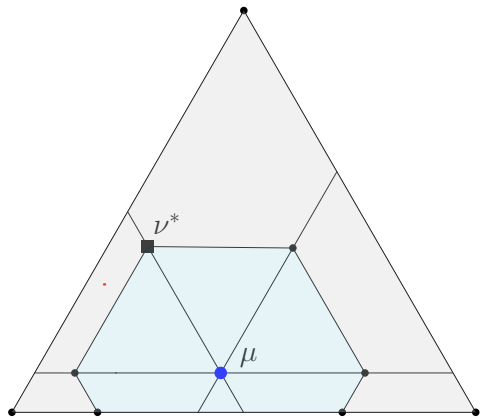


Optimization landscape in minimum optimal transport estimation. Here for the independence model of 2 binary variables. [Celik et al., 2020a]

Wasserstein distance to independence models







Combinatorics of critical points for minimum Wasserstein distance estimation. [Celik et al., 2020b]

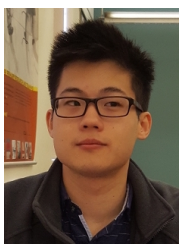
We have been developing Wasserstein Information Geometry to define

- Loss functions for parameter estimation (WGANs, WWGANs)
- Geometry of data space (e.g. images, adversarial robustness)
- Optimization methods (Wasserstein Natural Gradient)

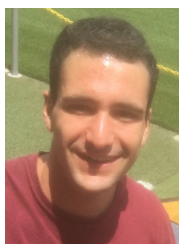
Dynamical formulation of optimal transport and Riemannian structures.

Works [Li and Montúfar, 2018, Lin et al., 2018, Li et al., 2019, Dukler et al., 2019, Lin et al., 2019, Arbel et al., 2020] with


W. Li


A. Lin

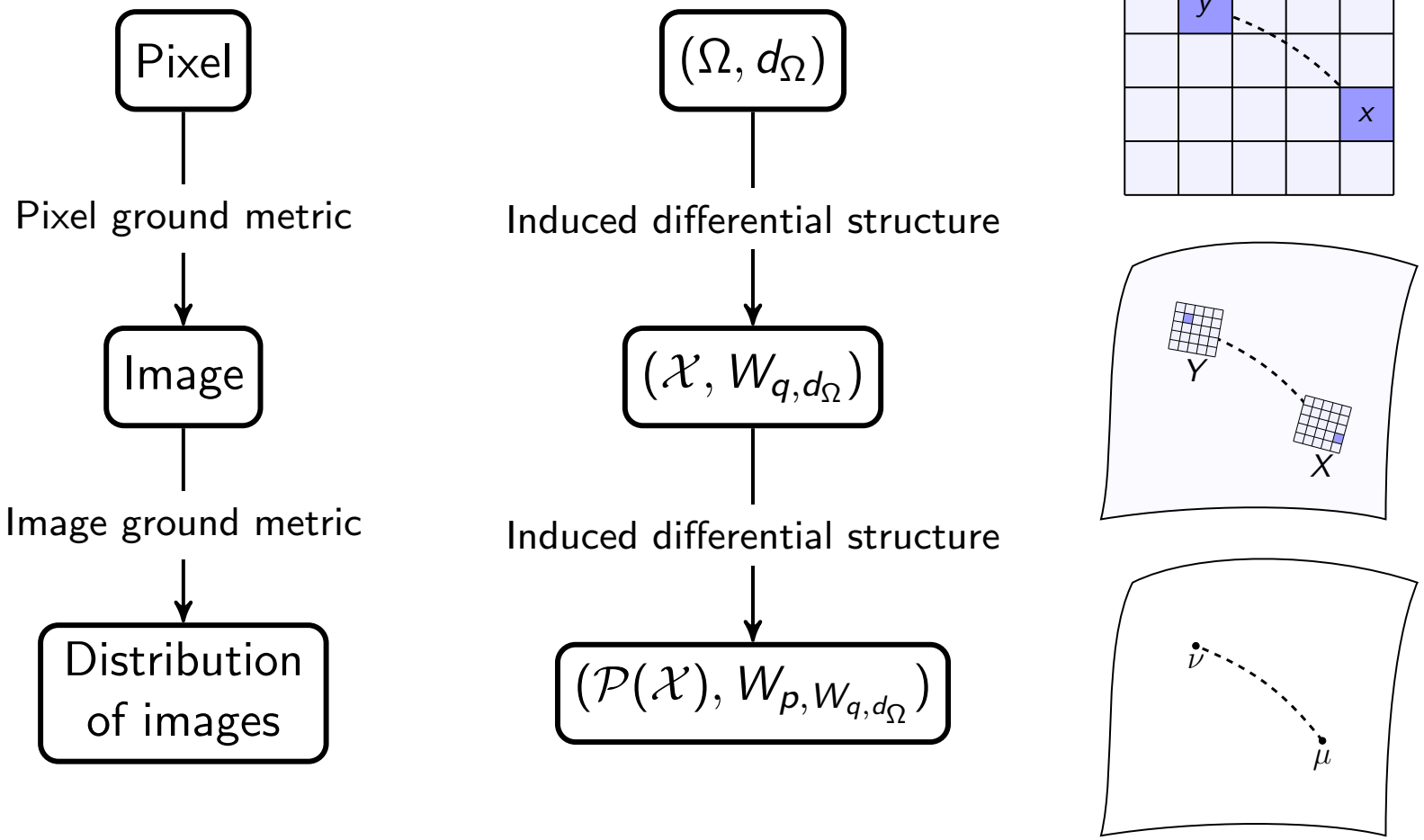

S. Osher


Y. Dukler


M. Arbel


A. Gretton

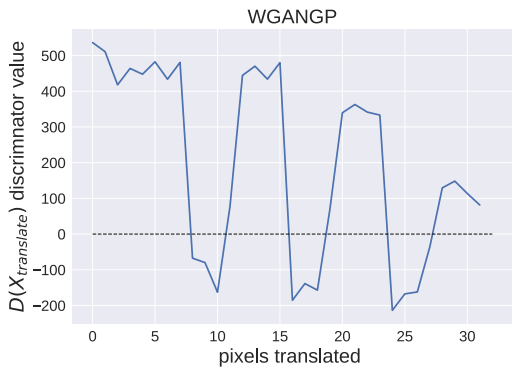
Wasserstein ground metric for WGANs

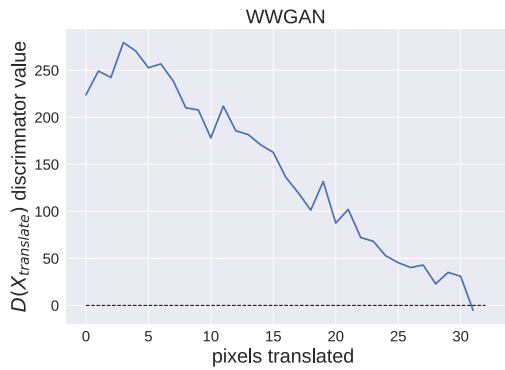


Wasserstein of Wasserstein loss [Dukler et al., 2019]

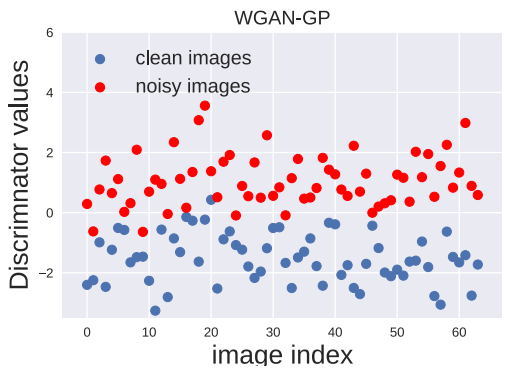
Improved interpretability

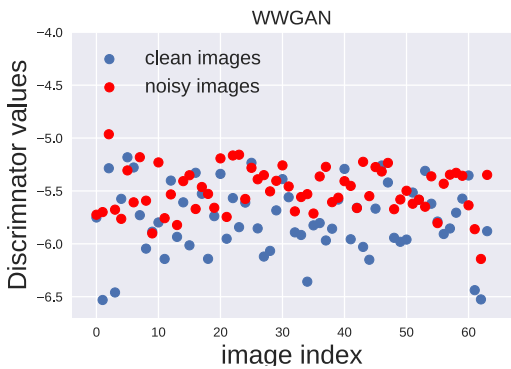
- Stability to natural data variability
  WGAN and WWGAN discriminators for CIFAR-10 images translated continuously.





- Stability to noise
  Discriminator values on CIFAR-10 images with RGB salt and pepper noise 15% of the pixels.

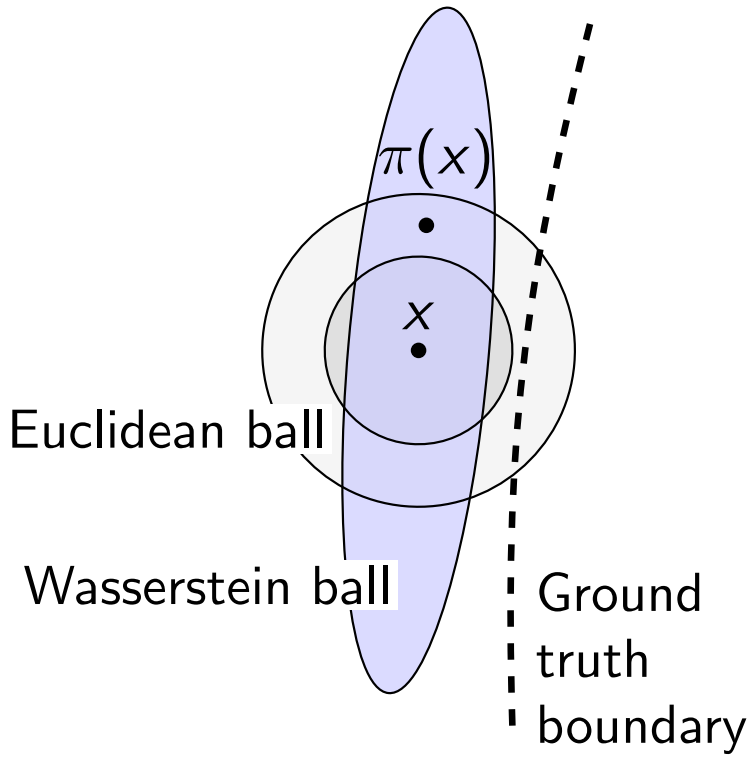




Wasserstein of Wasserstein loss [Dukler et al., 2019]

https://github.com/dukleryoni/WWGAN

Wasserstein ground metric for classification



- In this geometry, we may be able to apply higher levels of smoothness regularization (noise or gradient)

Wasserstein Diffusion Tikhonov Regularization [Lin et al., 2019]

Improved adversarial robustness

Robustness to adversarial attacks

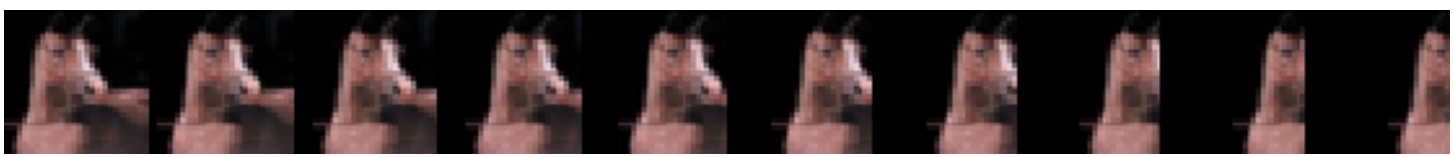
Test error under adversarial attacks on CIFAR-10

Test data	None	Euclidean grad.	Wassserstein grad.	SOTA '19
Natural	16.29	15.61	**15.35**	
FGSM $\epsilon = 8/255$	82.22	31.10	**30.20**	37.52
FGSM $\epsilon = 25/255$	89.72	66.83	**44.32**	
I-FGSM-20 $\alpha = 2/255$, $\epsilon = 8/255$	90.15	40.06	**32.12**	42.06

Stability to large in-class variations

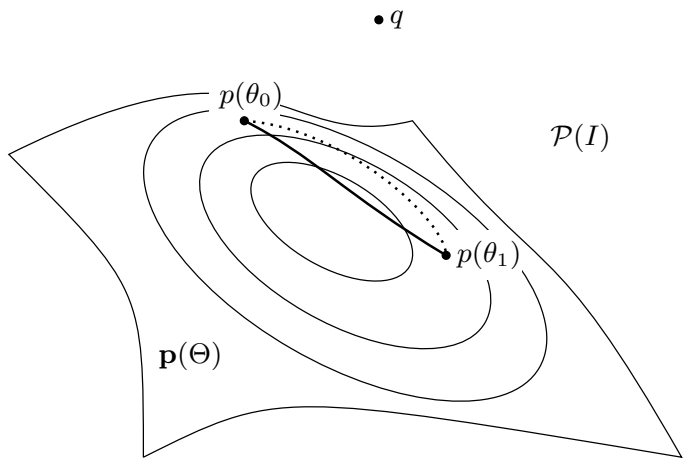
Average nr of prediction flips on sequences of translated test images CIFAR-10

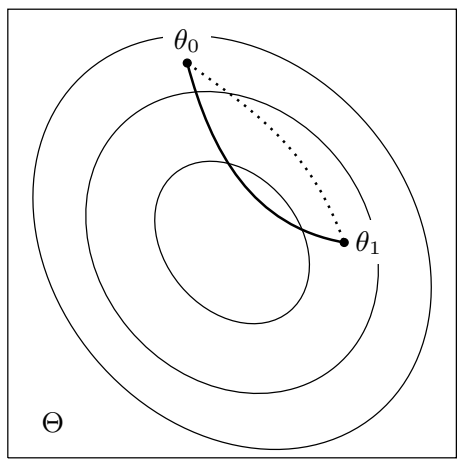
Perturbation \ Regularizer	None	Euclidean grad.	Wasserstein grad.
Horizontal translation	10.009	7.898	**6.488**
Vertical translation	9.920	9.437	**7.956**

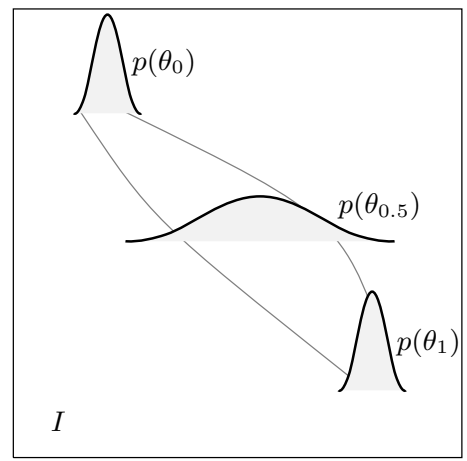


Wasserstein Diffusion Tikhonov Regularization [Lin et al., 2019]

Geometry of parameter space

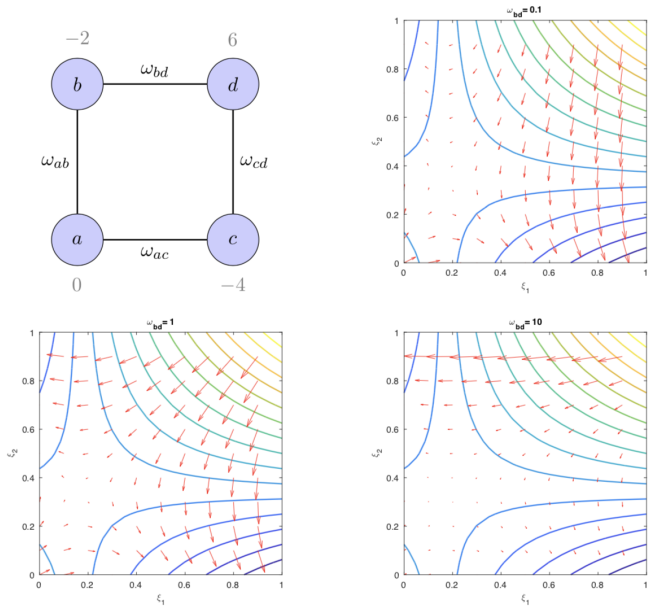




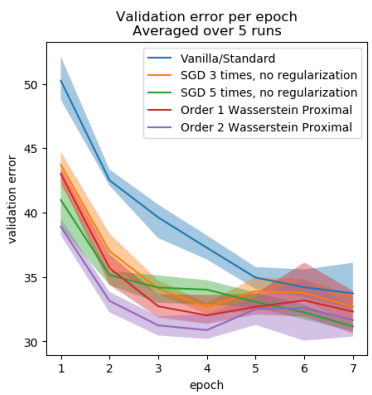


Parameter Space - Function Space - Data Space - Loss Function

Improved gradient optimization

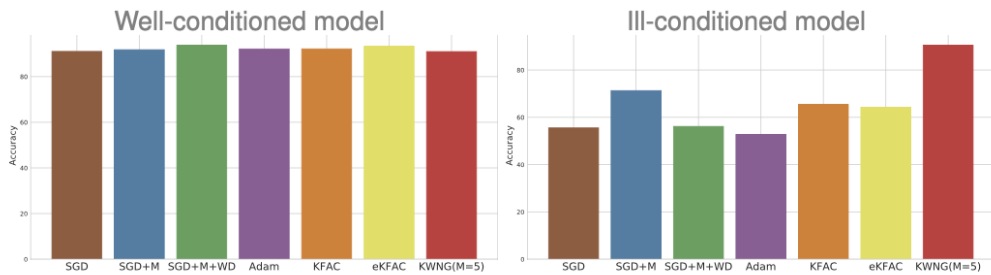


Natural Gradient via OT
[Li and Montúfar, 2018]



Affine natural proximal
[Li et al., 2019]

- Approximately invariant to parametrization
- Can be used as a drop-in optimizer
- Fast and scalable
- Comes with statistical guarantees



Cifar10 classification task using ResNet-18 networks.

Kernelized Wasserstein Natural Gradient
[Arbel et al., 2020]
M. Arbel's talk at ICLR 2020
https://github.com/MichaelArbel/KWNG

1 Introduction

2 Excursion

3 Parametric linear programming

4 Polyhedral norm distances

Linear programming

- We have a nested program

$$\min_{\nu \in \mathcal{M}} W_d(\mu, \nu) = \min_{\nu \in \mathcal{M}} \min_{\pi \in \Pi(\mu,\nu)} \sum_{1 \leq i < j \leq n} d_{ij}\, \pi_{ij}$$

- The inner problem is a linear program

$$\begin{aligned} \min_{\pi} \; & d \cdot \pi \\ \text{st} \;\; & A\pi = b \\ & \pi \geq 0 \end{aligned}$$

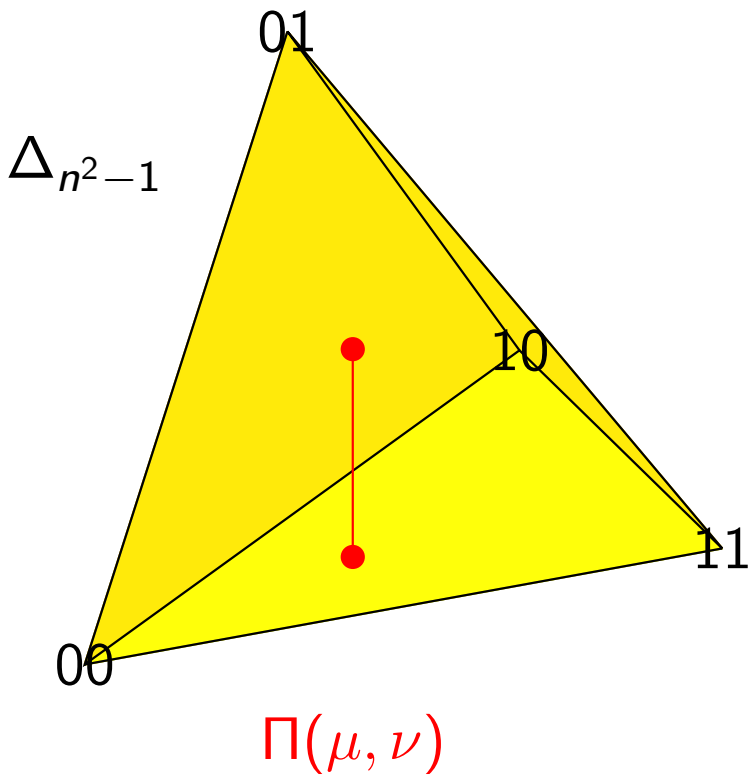
where A computes margins (columns are vertices of $\Delta_{n-1} \times \Delta_{n-1}$), and $b = [\mu_1, \ldots, \mu_n, \nu_1, \ldots, \nu_n]^\top$

Linear programming

- Write $A = [B, R]$ with full rank basis B
- Then $A\pi = B\pi^B + R\pi^R = b$ is solved by $\pi^B = B^{-1}b - B^{-1}R\pi^R$ and $d \cdot \pi = d^B(B^{-1}b - B^{-1}R\pi^R) + d^R\pi^R$
- The basic solution $\overline{\pi}^B = B^{-1}b$, $\overline{\pi}^R = 0$ satisfies eq constraint $A\overline{\pi} = b$
- The basis B is optimal if the basic solution is feasible $\overline{\pi}^B = B^{-1}b \geq 0$ and optimal (releasing $\overline{\pi}^R = 0$ doesn't help) $[d^B B^{-1}R - d^R] \leq 0$
- For a fixed B we have a linear map $(\mu, \nu) \mapsto [\overline{\pi}^B; 0]$
- The optimal bases, as μ, ν change, are the maximal simplices in a triangulation of the $2(n-1)$-polytope $\Delta_{n-1} \times \Delta_{n-1}$.

1 bit example

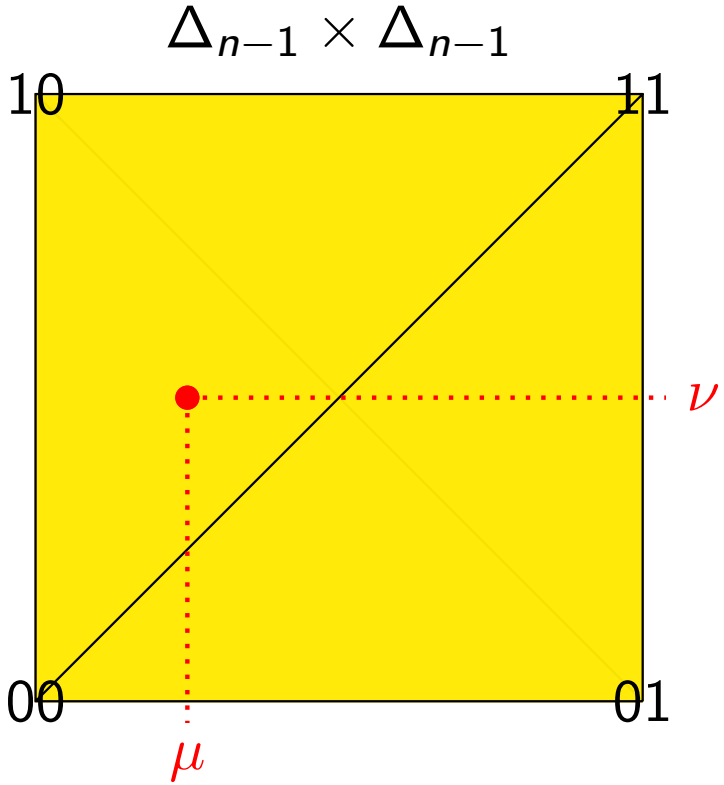
$$\dim \Pi(\mu,\nu) = (n^2-1) - 2(n-1)$$



- A basis B corresponds to a full-dim simplex of $\Delta_{n-1} \times \Delta_{n-1}$
  (a choice of columns of A)
- The transportation polytope $\Pi(\mu,\nu)$ is the fiber of the margins (μ,ν) in Δ_{n^2-1}
- A basic solution is a vertex of $\Pi(\mu,\nu)$
- $\pi \in \Pi(\mu,\nu)$ is a vertex iff $\mathrm{supp}(\pi)$ is a spanning forest in $K_{n,n}$,
  e.g. $\begin{smallmatrix} 1 & & 1 \\ 0 & & 0 \end{smallmatrix}$ gives $\pi = \left[\begin{smallmatrix} * & * \\ * & 0 \end{smallmatrix}\right]$
- For a given d, the optimal bases correspond to a triangulation Σ_d of $\Delta_{n-1} \times \Delta_{n-1}$

1 bit example

$$\dim \Pi(\mu,\nu) = (n^2-1) - 2(n-1)$$



- A basis B corresponds to a full-dim simplex of $\Delta_{n-1} \times \Delta_{n-1}$
  (a choice of columns of A)
- The transportation polytope $\Pi(\mu,\nu)$ is the fiber of the margins (μ,ν) in Δ_{n^2-1}
- A basic solution is a vertex of $\Pi(\mu,\nu)$
- $\pi \in \Pi(\mu,\nu)$ is a vertex iff $\mathrm{supp}(\pi)$ is a spanning forest in $K_{n,n}$,
  e.g. $\begin{smallmatrix} 1 & & 1 \\ 0 & & 0 \end{smallmatrix}$ gives $\pi = \left[\begin{smallmatrix} * & * \\ * & 0 \end{smallmatrix}\right]$
- For a given d, the optimal bases correspond to a triangulation Σ_d of $\Delta_{n-1} \times \Delta_{n-1}$

Linear programming perspective

- We have thus a piecewise linear map

$$(\mu,\nu) \mapsto W_d(\mu,\nu) = \sum_{1\leq i,j\leq n} d_{ij}\cdot(\tilde{\pi}_\sigma)_{ij}, \quad \text{for } (\mu,\nu)\in\sigma \qquad (1)$$

  An algorithm can be given as follows.

1. Compute the regular triangulation Σ_d of $\Delta_{n-1}\times\Delta_{n-1}$.
2. For a given data distribution μ, for each simplex $\sigma\in\Sigma_d$, minimize the linear function (1) over $(\mu\times\mathcal{M})\cap\sigma$.
3. Among all these solutions, select the one with smallest value.

Algorithm 1: A friendly description of the steps

Input: An $n \times n$ matrix $d = (d_{ij})$, a model $\mathcal{M} \subset \Delta_{n-1}$, and a distribution $\mu \in \Delta_{n-1}$.

Steps 1-3: Compute the triangulation of the polytope $\Delta_{n-1} \times \Delta_{n-1}$ that is given by d.

Step 4: Incorporate μ and express matrix entries $\tilde{\pi}_\sigma$ as linear functions in $\nu \in \mathcal{M}$.

Step 5: For each piece, minimize a linear function over the relevant part of the model $\mathcal{M}$.

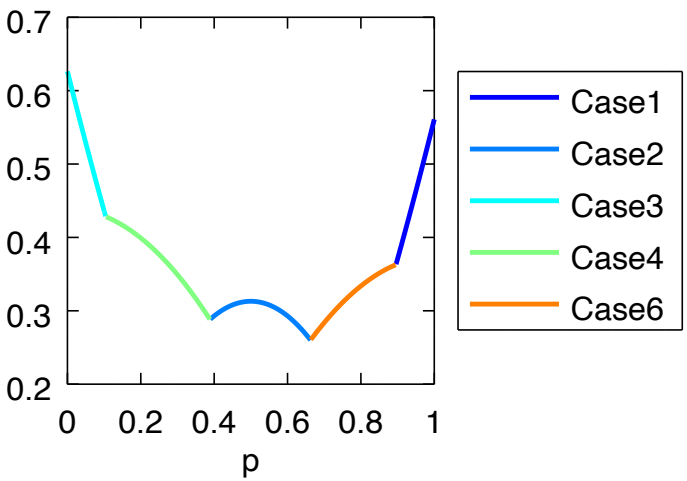
Steps 6-7: The smallest minimum found in Step 5 is the *Wasserstein distance* $W(\mu, \mathcal{M})$.

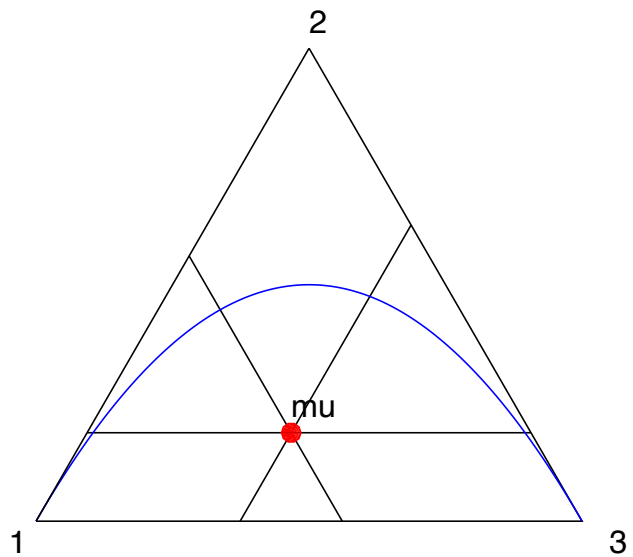
Hardy-Weinberg with discrete metric

- Let $n = 3$ with the discrete metric $d = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$
- Consider the Hardy-Weinberg model (2 trial binomial)

$$p \in [0,1] \mapsto \nu(p) = (p^2, 2p(1-p), (1-p)^2)$$

Quadratic curve inside Δ_2.



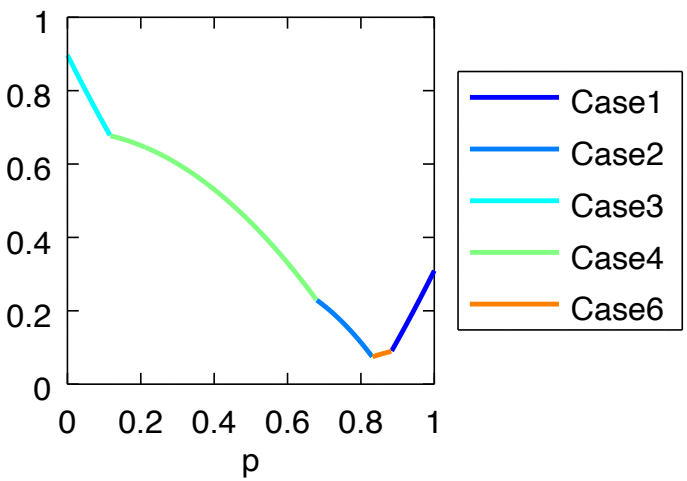


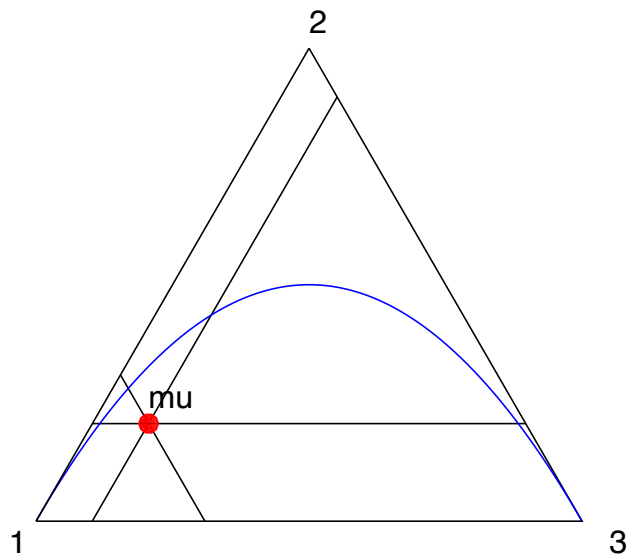
Hardy-Weinberg with discrete metric

- Let $n = 3$ with the discrete metric $d = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$
- Consider the Hardy-Weinberg model (2 trial binomial)

$$p \in [0,1] \mapsto \nu(p) = (p^2, 2p(1-p), (1-p)^2)$$

Quadratic curve inside Δ_2.



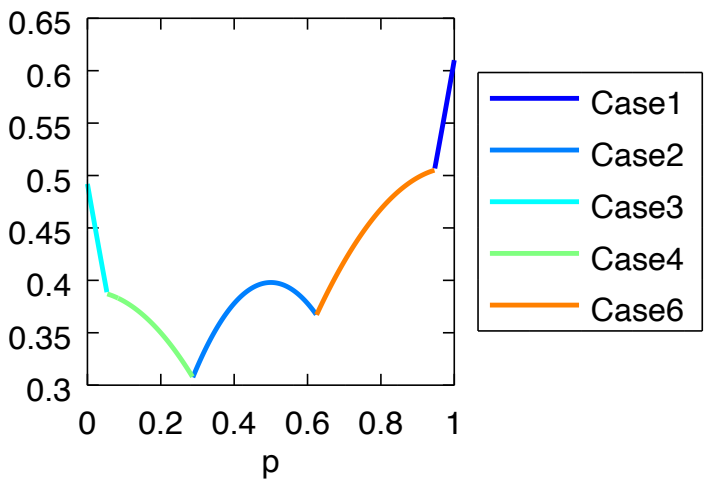


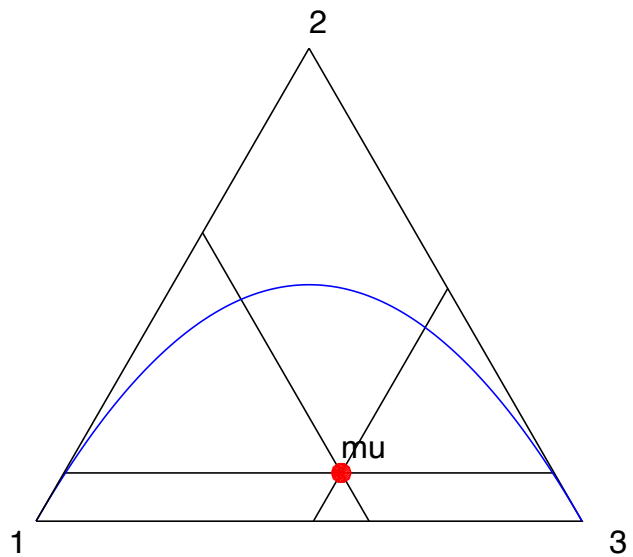
Hardy-Weinberg with discrete metric

- Let $n = 3$ with the discrete metric $d = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$
- Consider the Hardy-Weinberg model (2 trial binomial)

$$p \in [0,1] \mapsto \nu(p) = (p^2, 2p(1-p), (1-p)^2)$$

Quadratic curve inside Δ_2.



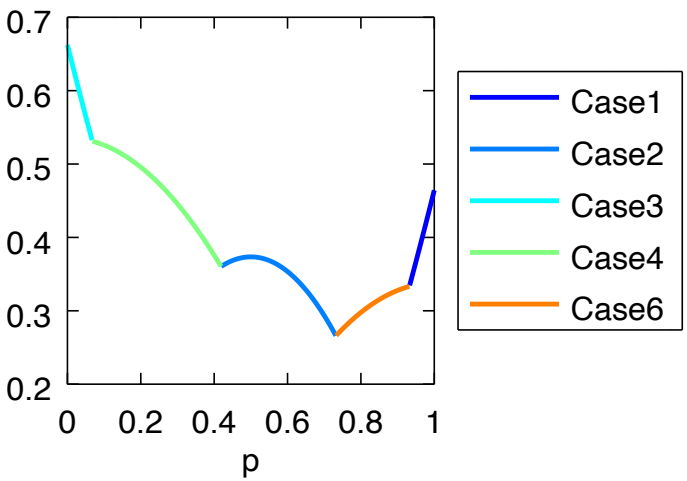


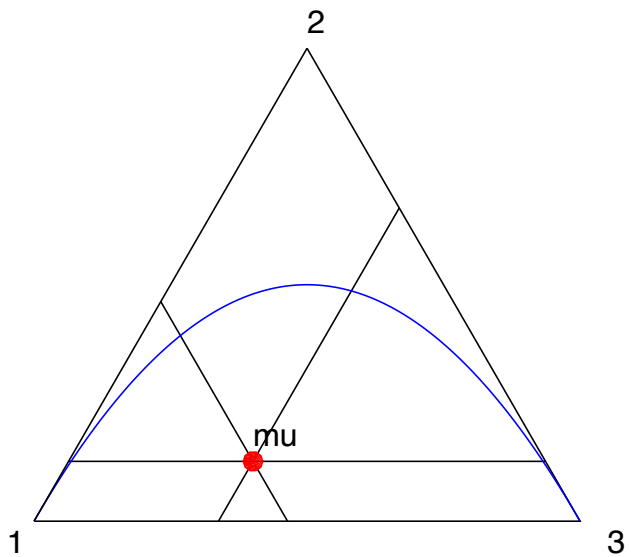
Hardy-Weinberg with discrete metric

- Let $n = 3$ with the discrete metric $d = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$
- Consider the Hardy-Weinberg model (2 trial binomial)

$$p \in [0,1] \mapsto \nu(p) = (p^2, 2p(1-p), (1-p)^2)$$

Quadratic curve inside Δ_2.



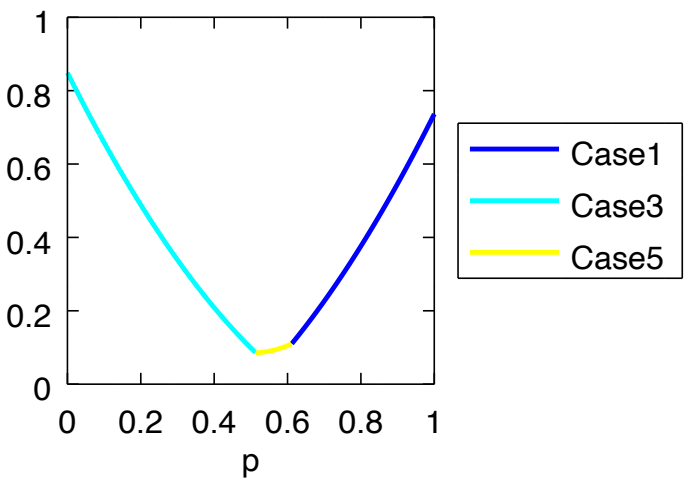


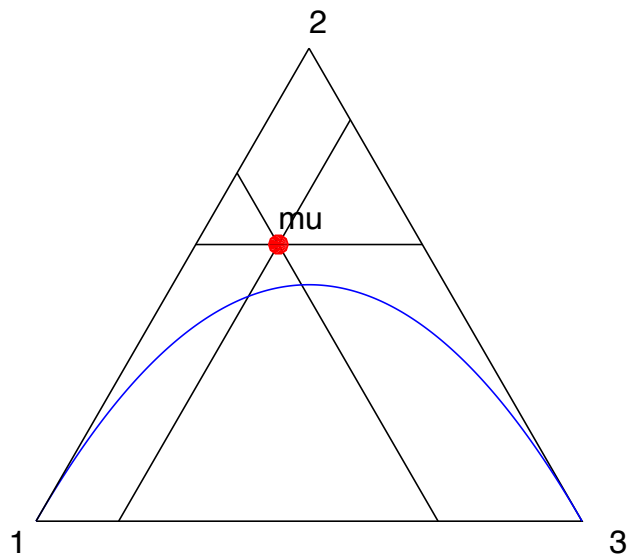
Hardy-Weinberg with discrete metric

- Let $n = 3$ with the discrete metric $d = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$
- Consider the Hardy-Weinberg model (2 trial binomial)

$$p \in [0,1] \mapsto \nu(p) = (p^2, 2p(1-p), (1-p)^2)$$

Quadratic curve inside Δ_2.





2 bit independence model

- The 2-*bit independence model*

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

  Quadratic surface defined by the equation $\nu_1\nu_4 = \nu_2\nu_3$.

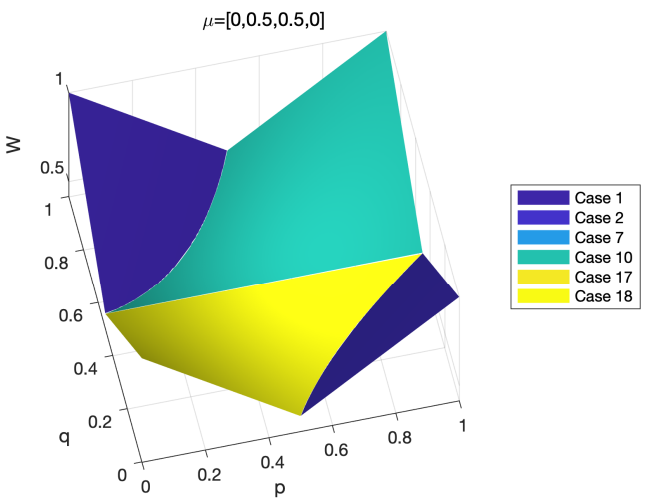
- We fix the L_0-metric (Hamming distance on $\{0,1\}^2$)

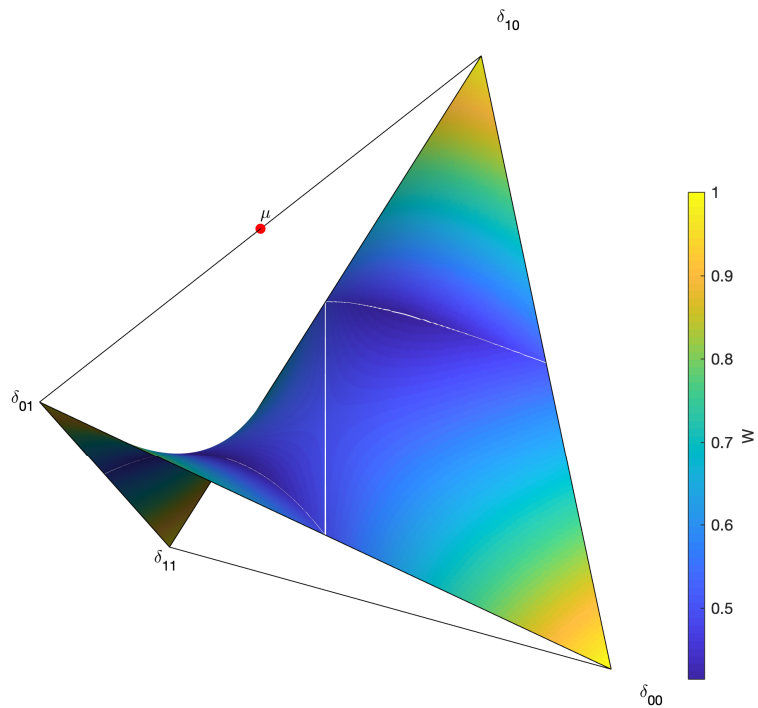
$$d = \begin{bmatrix} 0 & 1 & 1 & 2 \\ 1 & 0 & 2 & 1 \\ 1 & 2 & 0 & 1 \\ 2 & 1 & 1 & 0 \end{bmatrix}.$$

2 bit independence model

- The objective function can be written

$$\nu \mapsto W(\mu,\nu) = \max\{|\mu_1+\mu_2-p|+|\mu_1+\mu_3-q|, \\ |\mu_1+\mu_4-(pq+(1-p)(1-q))|\}$$

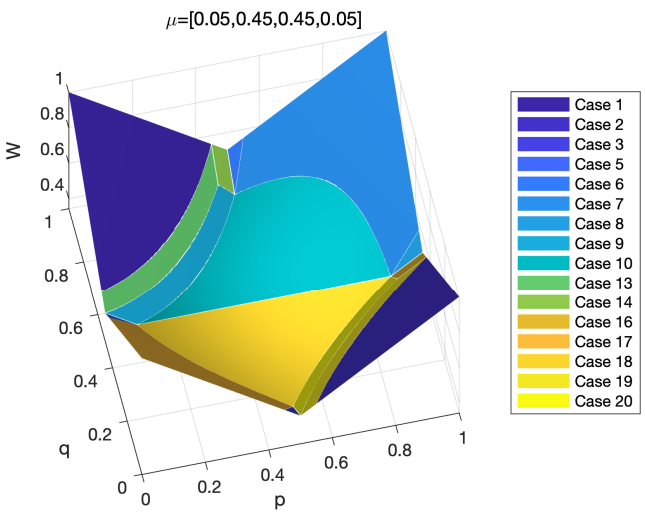


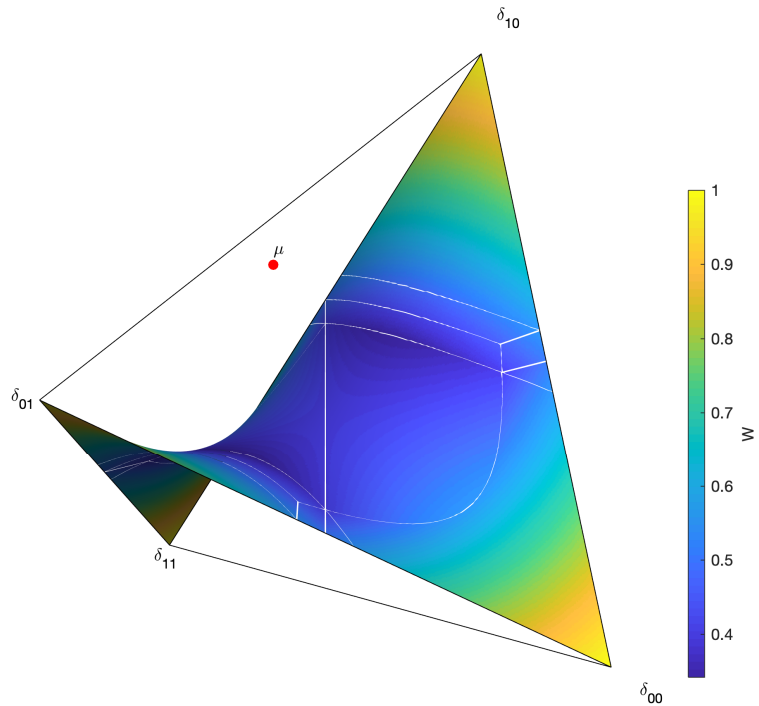


2 bit independence model

- The objective function can be written

$$\nu \mapsto W(\mu,\nu) = \max\{|\mu_1+\mu_2-p|+|\mu_1+\mu_3-q|, \\ |\mu_1+\mu_4-(pq+(1-p)(1-q))|\}$$

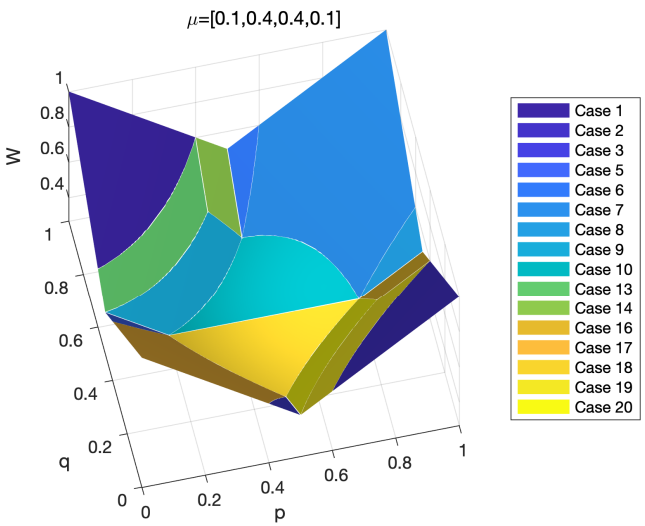


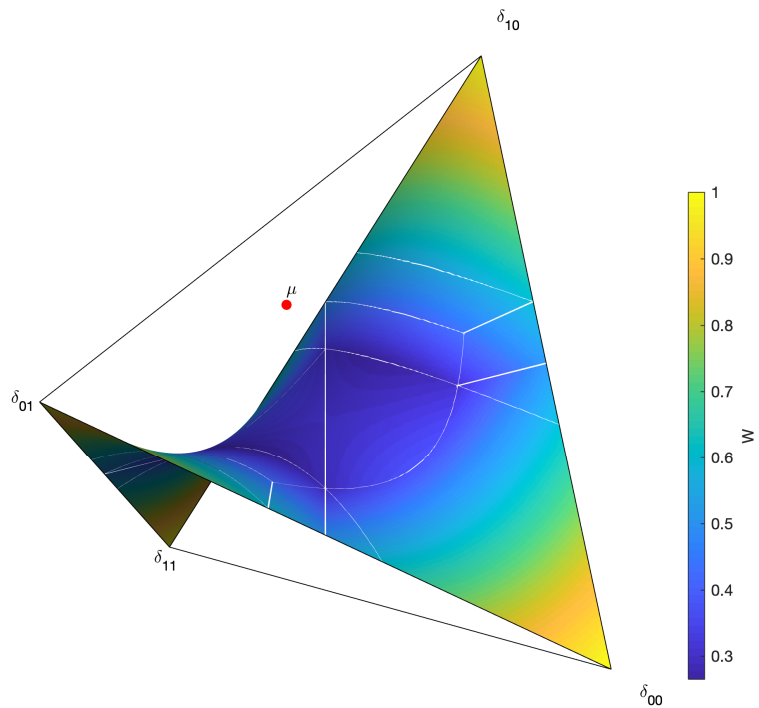


2 bit independence model

- The objective function can be written

$$\nu \mapsto W(\mu,\nu) = \max\{|\mu_1+\mu_2-p|+|\mu_1+\mu_3-q|, \\ |\mu_1+\mu_4-(pq+(1-p)(1-q))|\}$$

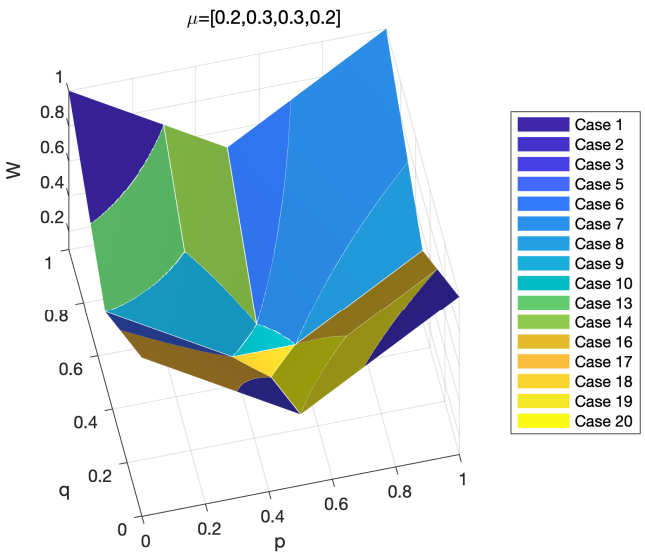


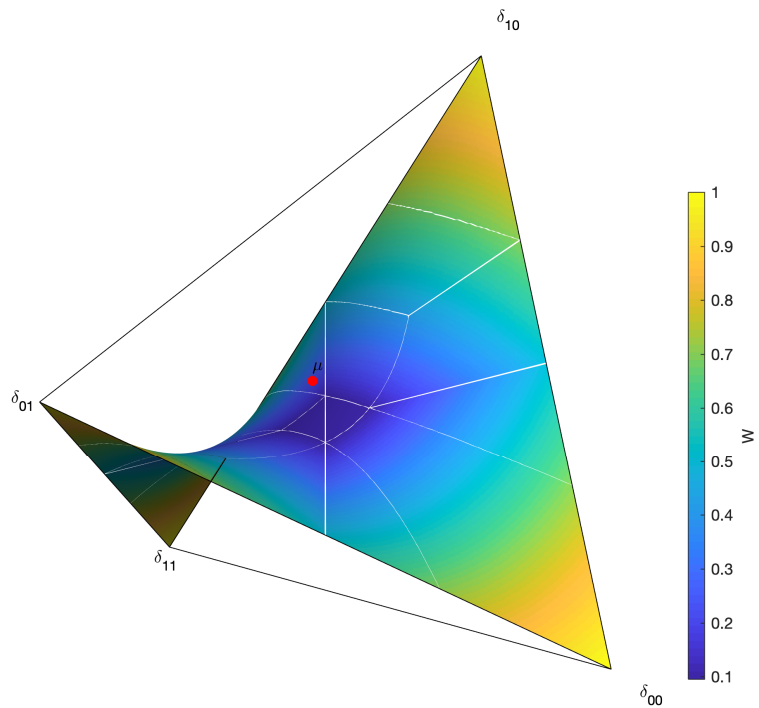


2 bit independence model

- The objective function can be written

$$\nu \mapsto W(\mu,\nu) = \max\{|\mu_1+\mu_2-p|+|\mu_1+\mu_3-q|, \\ |\mu_1+\mu_4-(pq+(1-p)(1-q))|\}$$

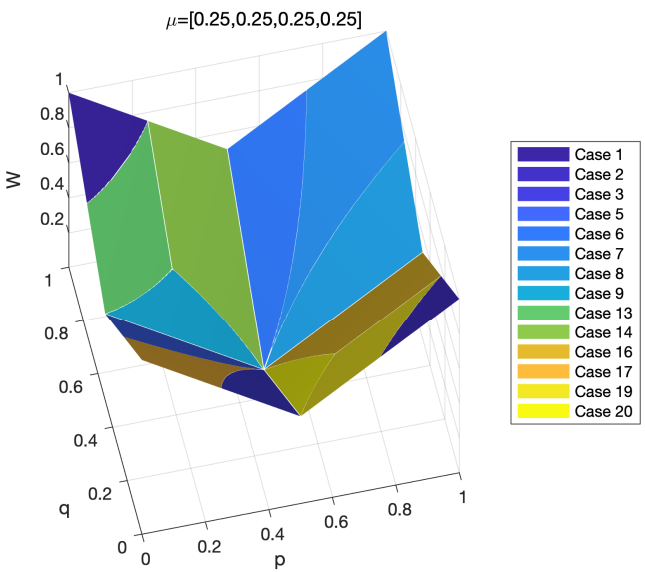


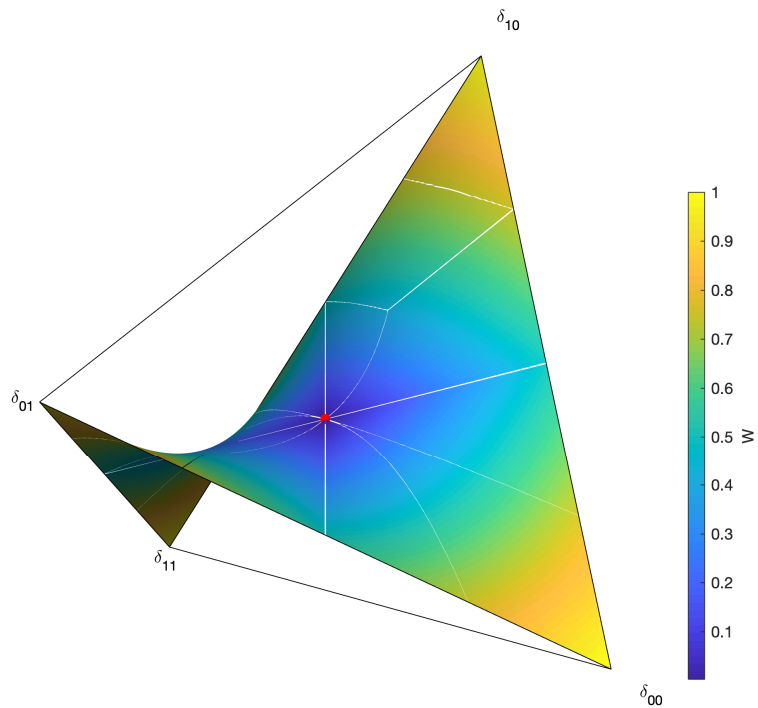


2 bit independence model

- The objective function can be written

$$\nu \mapsto W(\mu,\nu) = \max\{|\mu_1+\mu_2-p|+|\mu_1+\mu_3-q|, \\ |\mu_1+\mu_4-(pq+(1-p)(1-q))|\}$$

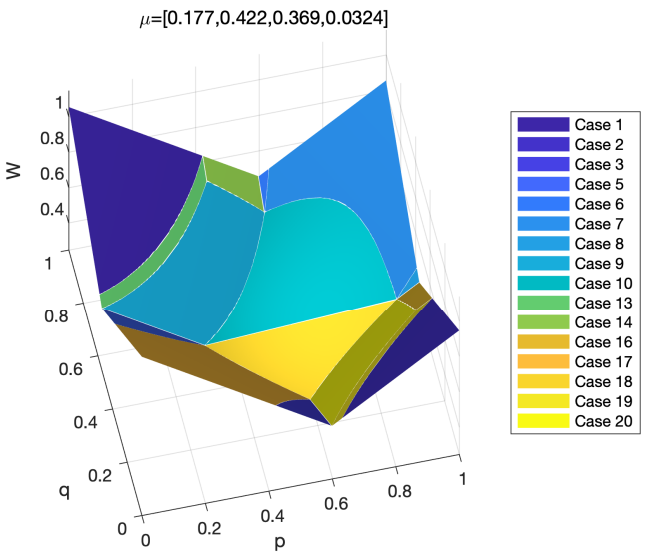


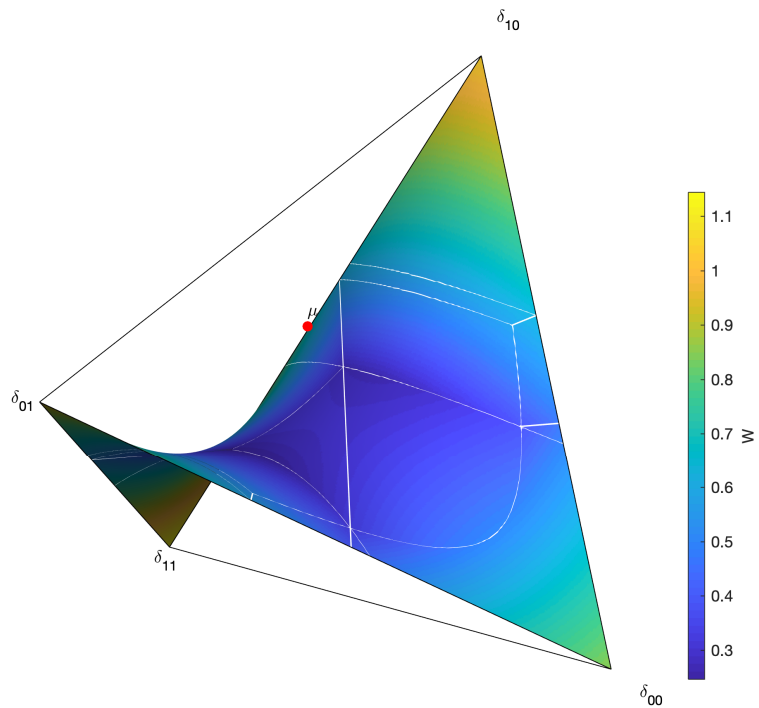


2 bit independence model

- The objective function can be written

$$\nu \mapsto W(\mu,\nu) = \max\{|\mu_1+\mu_2-p|+|\mu_1+\mu_3-q|, \\ |\mu_1+\mu_4-(pq+(1-p)(1-q))|\}$$

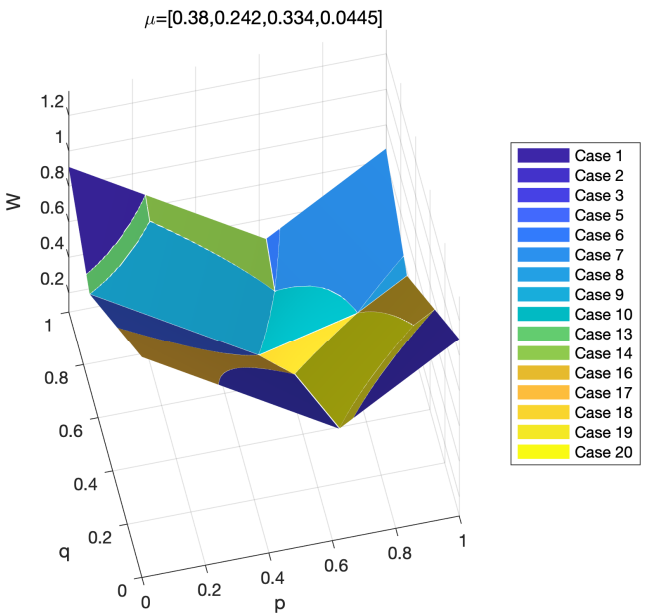


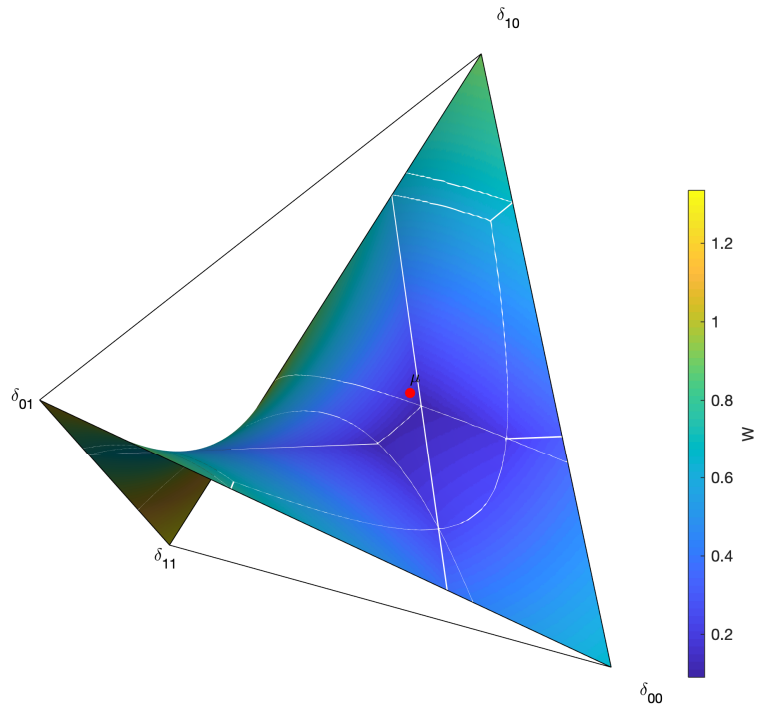


2 bit independence model

- The objective function can be written

$$\nu \mapsto W(\mu,\nu) = \max\{|\mu_1+\mu_2-p|+|\mu_1+\mu_3-q|, \\ |\mu_1+\mu_4-(pq+(1-p)(1-q))|\}$$

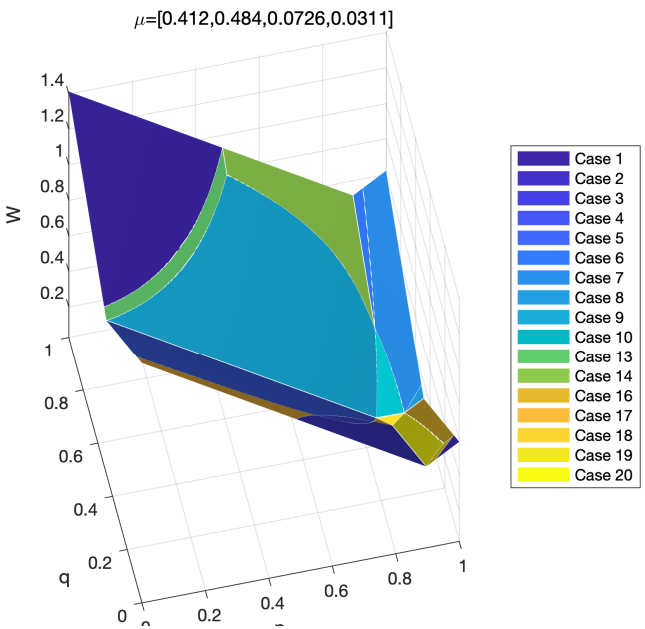


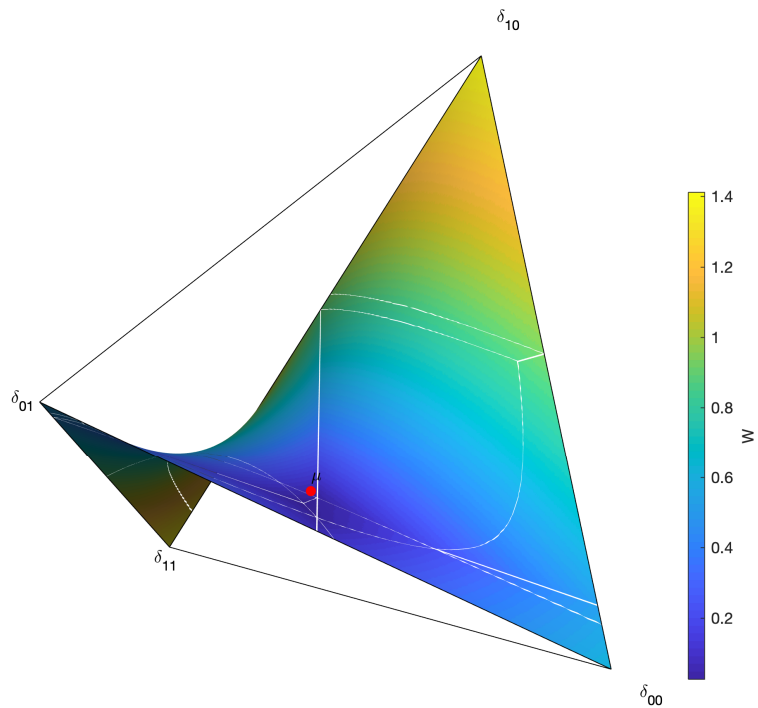


2 bit independence model

- The objective function can be written

$$\nu \mapsto W(\mu,\nu) = \max\{|\mu_1+\mu_2-p|+|\mu_1+\mu_3-q|, \\ |\mu_1+\mu_4-(pq+(1-p)(1-q))|\}$$

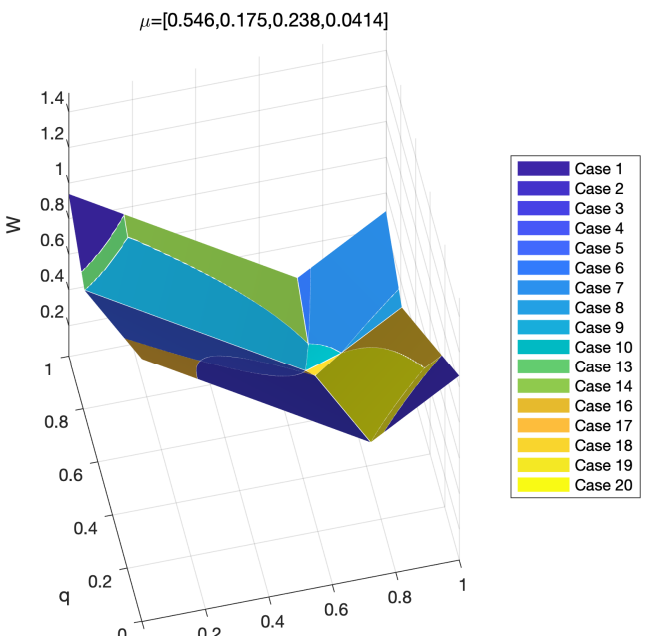


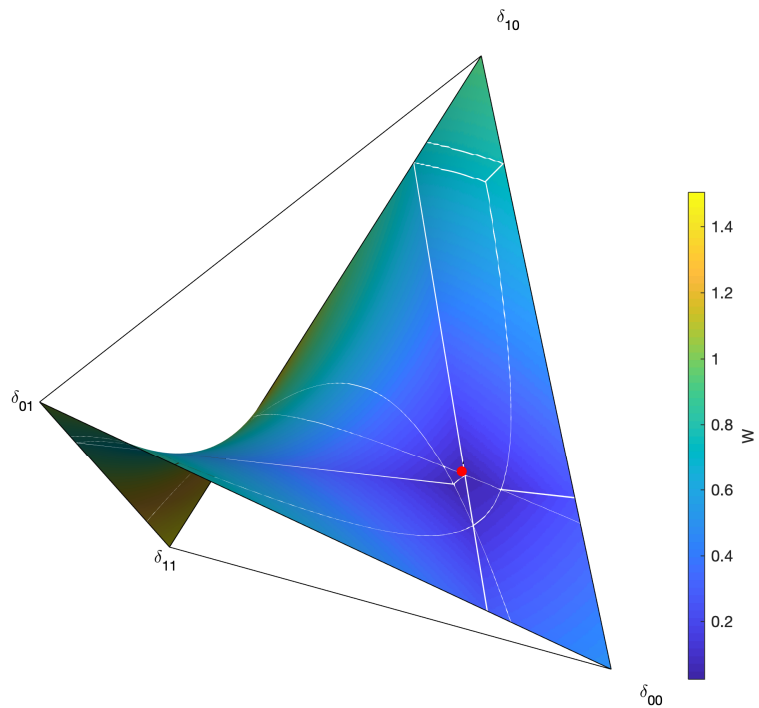


2 bit independence model

- The objective function can be written

$$\nu \mapsto W(\mu,\nu) = \max\{|\mu_1+\mu_2-p|+|\mu_1+\mu_3-q|, \\ |\mu_1+\mu_4-(pq+(1-p)(1-q))|\}$$

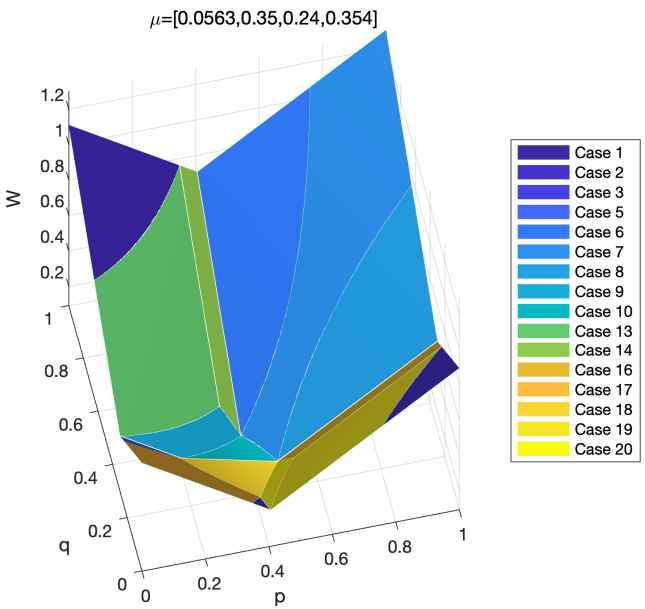


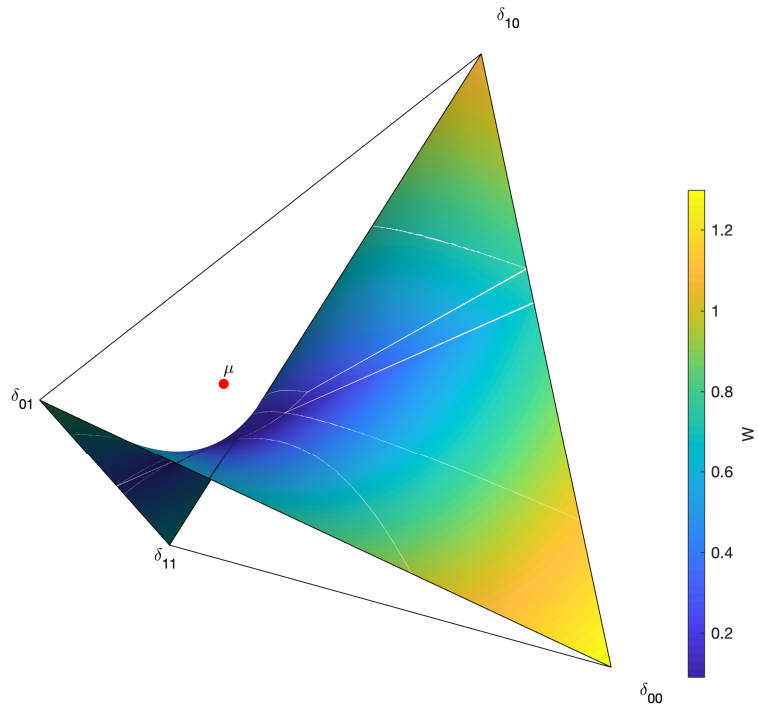


2 bit independence model

- The objective function can be written

$$\nu \mapsto W(\mu,\nu) = \max\{|\mu_1+\mu_2-p|+|\mu_1+\mu_3-q|, \\ |\mu_1+\mu_4-(pq+(1-p)(1-q))|\}$$





- The case distinction arises from (a triangulation near) the induced polyhedral subdivision of the 6-polytope $\Delta_3 \times \Delta_3$.
- The triangulation has 20 maximal simplices and restricts to a mixed subdivision on $\mu \times \Delta_3$.

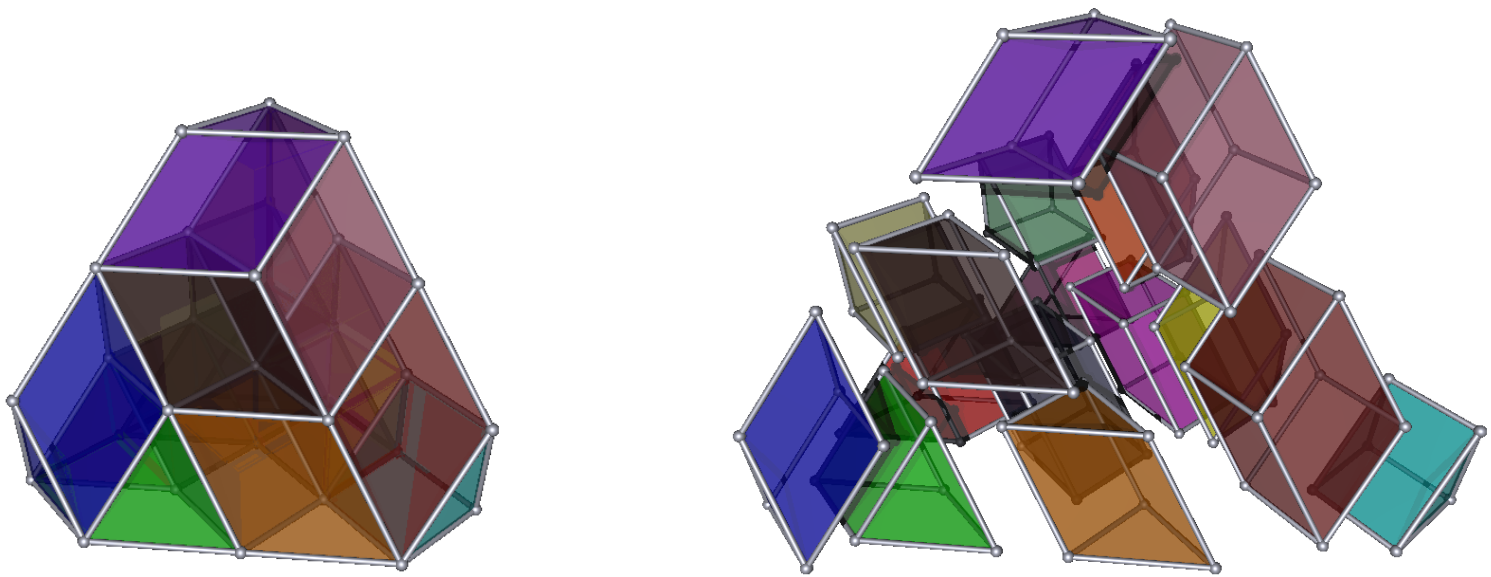


Figure 1: A mixed subdivision of a (truncated) tetrahedron into $(4+)12+4$ cells.

Case	Objective Function	Feasible Region, $0 \leq * \leq 1$	Solution	Minimum Value	Subdivision
Quadratic pieces					
10	$2pq - p - q + \mu_2 + \mu_3$	$q - \mu_1 - \mu_3$ $\mu_3 - (1-p)q$ $(1-p)(1-q) - \mu_4$	$(\frac{\mu_1}{\mu_1+\mu_3}, \mu_1 + \mu_3)$ $(\frac{\mu_2}{\mu_2+\mu_4}, \mu_1 + \mu_3)$ $(\mu_1 + \mu_2, \frac{\mu_3}{\mu_3+\mu_4})$ $(1-\sqrt{\mu_3}, \sqrt{\mu_3})$	$-\frac{\mu_1}{\mu_1+\mu_3} + \mu_1 + \mu_2$ $\frac{\mu_2}{\mu_2+\mu_4} - \mu_1 - \mu_2$ $\frac{\mu_3}{\mu_3+\mu_4} - \mu_1 - \mu_3$ $2\sqrt{\mu_3}(1-\sqrt{\mu_3}) - \mu_1 - \mu_4$	$\begin{bmatrix} * & * & * & 0 \\ 0 & * & 0 & 0 \\ 0 & 0 & * & 0 \\ 0 & * & 0 & * \end{bmatrix}$
18	$2pq - p - q + \mu_2 + \mu_3$	$pq - \mu_1$ $\mu_2 - p(1-q)$ $\mu_1 + \mu_3 - q$	$(\mu_1 + \mu_2, \frac{\mu_1}{\mu_1+\mu_2})$ $(\frac{\mu_2}{\mu_2+\mu_4}, \mu_1 + \mu_3)$ $(\frac{\mu_1}{\mu_1+\mu_3}, \mu_1 + \mu_3)$ $(\sqrt{\mu_2}, 1-\sqrt{\mu_2})$	$-\frac{\mu_1}{\mu_1+\mu_2} + \mu_1 + \mu_3$ $\frac{\mu_2}{\mu_2+\mu_4} - \mu_1 - \mu_2$ $-\frac{\mu_1}{\mu_1+\mu_3} + \mu_1 + \mu_2$ $2\sqrt{\mu_2}(1-\sqrt{\mu_2}) - \mu_1 - \mu_4$	$\begin{bmatrix} * & 0 & * & 0 \\ 0 & * & 0 & 0 \\ 0 & 0 & * & 0 \\ 0 & * & * & * \end{bmatrix}$
12	$-2pq + p + q - \mu_2 - \mu_3$	$\mu_1 - pq$ $p(1-q) - \mu_2$ $q - \mu_1 - \mu_3$	$(\mu_1 + \mu_2, \frac{\mu_1}{\mu_1+\mu_2})$ $(\frac{\mu_1}{\mu_1+\mu_3}, \mu_1 + \mu_3)$ $(\frac{\mu_2}{\mu_2+\mu_4}, \mu_1 + \mu_3)$ $(\sqrt{\mu_1}, \sqrt{\mu_1})$	$\frac{\mu_1}{\mu_1+\mu_2} - \mu_1 - \mu_3$ $\frac{\mu_1}{\mu_1+\mu_3} - \mu_1 - \mu_2$ $-\frac{\mu_2}{\mu_2+\mu_4} + \mu_1 + \mu_2$ $2\sqrt{\mu_1}(1-\sqrt{\mu_1}) - \mu_2 - \mu_3$	$\begin{bmatrix} * & 0 & 0 & 0 \\ 0 & * & 0 & * \\ * & 0 & * & * \\ 0 & 0 & 0 & * \end{bmatrix}$
15	$-2pq + p + q - \mu_2 - \mu_3$	$\mu_1 + \mu_3 - q$ $(1-p)q - \mu_3$ $\mu_4 - (1-p)(1-q)$	$(1-\sqrt{\mu_4}, 1-\sqrt{\mu_4})$ $(\frac{\mu_1}{\mu_1+\mu_3}, \mu_1 + \mu_3)$ $(\mu_1 + \mu_2, \frac{\mu_3}{\mu_3+\mu_4})$ $(\frac{\mu_2}{\mu_2+\mu_4}, \mu_1 + \mu_3)$	$2\sqrt{\mu_4}(1-\sqrt{\mu_4}) - \mu_2 - \mu_3$ $\frac{\mu_1}{\mu_1+\mu_3} - \mu_1 - \mu_2$ $-\frac{\mu_3}{\mu_3+\mu_4} + \mu_1 + \mu_3$ $-\frac{\mu_2}{\mu_2+\mu_4} + \mu_1 + \mu_2$	$\begin{bmatrix} * & 0 & 0 & 0 \\ * & * & 0 & * \\ * & 0 & * & 0 \\ 0 & 0 & 0 & * \end{bmatrix}$

Table 1: Quadratic cases for 2-bit independence model.
Objective is $|(\mu_1 + \mu_4) - (pq + ((1-p)(1-q))|$

Case	Objective Function	Feasible Region, $0 \le * \le 1$	Solution	Minimum Value	Subdivision
First affine piece					
2	$-p+q+\mu_2-\mu_3$	$(1-p)q+\mu_2-1$	$(1-\sqrt{1-\mu_2}, \sqrt{1-\mu_2})$	$2\sqrt{1-\mu_2}+\mu_2-\mu_3-1$	$\begin{bmatrix} 0&*&0&0\\ 0&*&0&0\\ *&*&*&*\\ 0&*&0&0 \end{bmatrix}$
9	$-p+q+\mu_2-\mu_3$	$\mu_1+\mu_3-(1-p)q$ $q-\mu_1-\mu_3$ $(1-p)q-\mu_3$ $(1-p)(1-q)-\mu_4$	$\left(\frac{\mu_2}{\mu_2+\mu_4}, \mu_1+\mu_3\right)$ $(1-\sqrt{\mu_3}, \sqrt{\mu_3})$ $\left(\frac{\mu_1}{\mu_1+\mu_3}, \mu_1+\mu_3\right)$ $\left(\mu_1+\mu_2, \frac{\mu_3}{\mu_3+\mu_4}\right)$	$-\frac{\mu_2}{\mu_2+\mu_4}+\mu_1+\mu_2$ $2\sqrt{\mu_3}+\mu_2-\mu_3-1$ $-\frac{\mu_1}{\mu_1+\mu_3}+\mu_1+\mu_2$ $\frac{\mu_3}{\mu_3+\mu_4}-\mu_1-\mu_3$	$\begin{bmatrix} *&*&0&0\\ 0&*&0&0\\ *&0&*&0\\ 0&*&0&* \end{bmatrix}$
13	$-p+q+\mu_2-\mu_3$	μ_2-p $(1-p)q-\mu_1-\mu_3$ $1-\mu_2-(1-p)q$	$(1-\sqrt{\mu_1+\mu_3}, \sqrt{\mu_1+\mu_3})$ $\left(\mu_2, \frac{\mu_1+\mu_3}{\mu_1+\mu_3+\mu_4}\right)$	$2\sqrt{\mu_1+\mu_3}+\mu_2-\mu_3-1$ $\frac{\mu_1+\mu_3}{\mu_1+\mu_3+\mu_4}-\mu_3$	$\begin{bmatrix} 0&*&0&0\\ 0&*&0&0\\ *&0&*&*\\ 0&*&0&* \end{bmatrix}$
14	$-p+q+\mu_2-\mu_3$	$p-\mu_2$ $\mu_2-p(1-q)$ $\mu_1+\mu_2-p$ $\mu_4-(1-p)(1-q)$	$\left(\mu_1+\mu_2, \frac{\mu_3}{\mu_3+\mu_4}\right)$ $\left(\mu_1+\mu_2, \frac{\mu_1}{\mu_1+\mu_2}\right)$ $\left(\frac{\mu_2}{\mu_2+\mu_4}, \mu_1+\mu_3\right)$	$\frac{\mu_3}{\mu_3+\mu_4}-\mu_1-\mu_3$ $\frac{\mu_1}{\mu_1+\mu_2}-\mu_1-\mu_3$ $-\frac{\mu_2}{\mu_2+\mu_4}+\mu_1+\mu_2$	$\begin{bmatrix} *&*&0&0\\ 0&*&0&0\\ *&0&*&*\\ 0&0&*&* \end{bmatrix}$
Second affine piece					
1	$p-q-\mu_2+\mu_3$	$p(1-q)+\mu_3-1$	$(\sqrt{1-\mu_3}, 1-\sqrt{1-\mu_3})$	$2\sqrt{1-\mu_3}-\mu_2+\mu_3-1$	$\begin{bmatrix} 0&0&*&0\\ *&*&*&*\\ 0&0&*&0\\ 0&0&*&0 \end{bmatrix}$
16	$p-q-\mu_2+\mu_3$	$q-\mu_3$ $\mu_3-(1-p)q$ $\mu_1+\mu_3-q$ $\mu_4-(1-p)(1-q)$	$\left(\mu_1+\mu_2, \frac{\mu_3}{\mu_3+\mu_4}\right)$ $\left(\frac{\mu_1}{\mu_1+\mu_3}, \mu_1+\mu_3\right)$ $\left(\frac{\mu_2}{\mu_2+\mu_4}, \mu_1+\mu_3\right)$	$-\frac{\mu_3}{\mu_3+\mu_4}+\mu_1+\mu_3$ $\frac{\mu_1}{\mu_1+\mu_3}-\mu_1-\mu_2$ $\frac{\mu_2}{\mu_2+\mu_4}-\mu_1-\mu_2$	$\begin{bmatrix} *&0&*&0\\ *&*&0&*\\ 0&0&*&0\\ 0&0&0&* \end{bmatrix}$
19	$p-q-\mu_2+\mu_3$	$p(1-q)-\mu_1-\mu_2$ μ_3-q $1-\mu_3-p(1-q)$	$(\sqrt{\mu_1+\mu_2}, 1-\sqrt{\mu_1+\mu_2})$ $\left(\frac{\mu_1+\mu_2}{\mu_1+\mu_2+\mu_4}, \mu_3\right)$	$2\sqrt{\mu_1+\mu_2}-\mu_2+\mu_3-1$ $\frac{\mu_1+\mu_2}{\mu_1+\mu_2+\mu_4}-\mu_2$	$\begin{bmatrix} 0&0&*&0\\ *&*&0&*\\ 0&0&*&0\\ 0&0&*&* \end{bmatrix}$
20	$p-q-\mu_2+\mu_3$	$\mu_1+\mu_2-p(1-q)$ $p(1-q)-\mu_2$ $p-\mu_1-\mu_2$ $(1-p)(1-q)-\mu_4$	$\left(\mu_1+\mu_2, \frac{\mu_3}{\mu_3+\mu_4}\right)$ $\left(\sqrt{\mu_2}, 1-\sqrt{\mu_2}\right)$ $\left(\mu_1+\mu_2, \frac{\mu_1}{\mu_1+\mu_2}\right)$ $\left(\frac{\mu_2}{\mu_2+\mu_4}, \mu_1+\mu_3\right)$	$-\frac{\mu_3}{\mu_3+\mu_4}+\mu_1+\mu_3$ $2\sqrt{\mu_2}-\mu_2+\mu_3-1$ $-\frac{\mu_1}{\mu_1+\mu_2}+\mu_1+\mu_3$ $\frac{\mu_2}{\mu_2+\mu_4}-\mu_1-\mu_2$	$\begin{bmatrix} *&0&*&0\\ *&*&0&0\\ 0&0&*&0\\ 0&0&*&* \end{bmatrix}$
Third affine piece					
3	$-p-q+\mu_1-\mu_4+1$	$p(1-q)-\mu_2$ $\mu_1+\mu_2-p$ $\mu_3-(1-p)q$	$\left(\mu_1+\mu_2, \frac{\mu_1}{\mu_1+\mu_2}\right)$ $\left(\mu_1+\mu_2, \frac{\mu_3}{\mu_3+\mu_4}\right)$	$-\frac{\mu_1}{\mu_1+\mu_2}+\mu_1+\mu_3$ $-\frac{\mu_3}{\mu_3+\mu_4}+\mu_1+\mu_3$	$\begin{bmatrix} *&0&0&0\\ *&*&0&0\\ 0&0&*&0\\ *&0&*&* \end{bmatrix}$
4	$-p-q+\mu_1-\mu_4+1$	$p(1-q)-\mu_2$ $(1-p)q-\mu_3$ $(1-p)(1-q)-\mu_4$	$\left(\gamma^+, 1-\frac{\mu_2}{\gamma^+}\right)$ $(1-\sqrt{\mu_4}, 1-\sqrt{\mu_4})$ $\left(\frac{\mu_2}{\mu_2+\mu_4}, \mu_1+\mu_3\right)$ $\left(\mu_1+\mu_2, \frac{\mu_3}{\mu_3+\mu_4}\right)$	$-\gamma^+ + \frac{\mu_2}{\gamma^+}+\mu_1-\mu_4$ $2\sqrt{\mu_4}+\mu_1-\mu_4-1$ $-\frac{\mu_2}{\mu_2+\mu_4}+\mu_1+\mu_2$ $-\frac{\mu_3}{\mu_3+\mu_4}+\mu_1+\mu_3$	$\begin{bmatrix} *&0&0&0\\ *&*&0&0\\ *&0&*&0\\ *&0&0&* \end{bmatrix}$
5	$-p-q+\mu_1-\mu_4+1$	$(1-p)q-\mu_3$ $\mu_1+\mu_3-q$ $\mu_2-p(1-q)$	$\left(\frac{\mu_1}{\mu_1+\mu_3}, \mu_1+\mu_3\right)$ $\left(\frac{\mu_2}{\mu_2+\mu_4}, \mu_1+\mu_3\right)$	$-\frac{\mu_1}{\mu_1+\mu_3}+\mu_1+\mu_2$ $-\frac{\mu_2}{\mu_2+\mu_4}+\mu_1+\mu_2$	$\begin{bmatrix} *&0&0&0\\ 0&*&0&0\\ *&0&*&0\\ *&*&0&* \end{bmatrix}$
17	$-p-q+\mu_1-\mu_4+1$	μ_1-pq $\mu_2-p(1-q)$ $\mu_3-(1-p)q$	$\left(\gamma^-, \frac{\mu_3}{1-\gamma^-}\right)$ $\left(\frac{\mu_1}{\mu_1+\mu_3}, \mu_1+\mu_3\right)$ $\left(\mu_1+\mu_2, \frac{\mu_1}{\mu_1+\mu_2}\right)$	$-\gamma^- - \frac{\mu_3}{1-\gamma^-}+\mu_1-\mu_4+1$ $-\frac{\mu_1}{\mu_1+\mu_3}+\mu_1+\mu_2$ $-\frac{\mu_1}{\mu_1+\mu_2}+\mu_1+\mu_3$	$\begin{bmatrix} *&0&0&0\\ 0&*&0&0\\ 0&0&*&0\\ *&*&*&* \end{bmatrix}$
Fourth affine piece					
6	$p+q-\mu_1+\mu_4-1$	$\mu_2-p(1-q)$ $p-\mu_1-\mu_2$ $(1-p)q-\mu_3$	$\left(\mu_1+\mu_2, \frac{\mu_3}{\mu_3+\mu_4}\right)$ $\left(\mu_1+\mu_2, \frac{\mu_1}{\mu_1+\mu_2}\right)$	$\frac{\mu_3}{\mu_3+\mu_4}-\mu_1-\mu_3$ $\frac{\mu_1}{\mu_1+\mu_2}-\mu_1-\mu_3$	$\begin{bmatrix} *&*&0&*\\ 0&*&0&0\\ 0&0&*&*\\ 0&0&0&* \end{bmatrix}$
7	$p+q-\mu_1+\mu_4-1$	$\mu_2-p(1-q)$ $\mu_3-(1-p)q$ $\mu_4-(1-p)(1-q)$	$\left(\gamma^+, \frac{\mu_3}{1-\gamma^+}\right)$ $\left(\frac{\mu_2}{\mu_2+\mu_4}, \mu_1+\mu_3\right)$ $\left(\mu_1+\mu_2, \frac{\mu_3}{\mu_3+\mu_4}\right)$	$\gamma^+ + \frac{\mu_3}{1-\gamma^+}-\mu_1+\mu_4-1$ $\frac{\mu_2}{\mu_2+\mu_4}-\mu_1-\mu_2$ $\frac{\mu_3}{\mu_3+\mu_4}-\mu_1-\mu_3$	$\begin{bmatrix} *&*&*&*\\ 0&*&0&0\\ 0&0&*&0\\ 0&0&0&* \end{bmatrix}$
8	$p+q-\mu_1+\mu_4-1$	$\mu_3-(1-p)q$ $q-\mu_1-\mu_3$ $p(1-q)-\mu_2$	$\left(\frac{\mu_2}{\mu_2+\mu_4}, \mu_1+\mu_3\right)$ $\left(\frac{\mu_1}{\mu_1+\mu_3}, \mu_1+\mu_3\right)$	$\frac{\mu_2}{\mu_2+\mu_4}-\mu_1-\mu_2$ $\frac{\mu_1}{\mu_1+\mu_3}-\mu_1-\mu_2$	$\begin{bmatrix} *&0&*&*\\ 0&*&0&*\\ 0&0&*&0\\ 0&0&0&* \end{bmatrix}$
11	$p+q-\mu_1+\mu_4-1$	$pq-\mu_1$ $p(1-q)-\mu_2$ $(1-p)q-\mu_3$	$\left(\gamma^-, \frac{\mu_3}{1-\gamma^-}\right)$ $\left(\sqrt{\mu_1}, \sqrt{\mu_1}\right)$ $\left(\mu_1+\mu_2, \frac{\mu_1}{\mu_1+\mu_2}\right)$ $\left(\frac{\mu_1}{\mu_1+\mu_3}, \mu_1+\mu_3\right)$	$\gamma^- + \frac{\mu_3}{1-\gamma^-}-\mu_1+\mu_4-1$ $2\sqrt{\mu_1}-\mu_1+\mu_4-1$ $\frac{\mu_1}{\mu_1+\mu_2}-\mu_1-\mu_3$ $\frac{\mu_1}{\mu_1+\mu_3}-\mu_1-\mu_2$	$\begin{bmatrix} *&0&0&*\\ 0&*&0&*\\ 0&0&*&*\\ 0&0&0&* \end{bmatrix}$
$\gamma^+ := (1+m_2-m_3)/2+\sqrt{(1+m_2-m_3)^2/4-m_2}$ and $\gamma^- := (1+m_2-m_3)/2-\sqrt{(1+m_2-m_3)^2/4-m_2}$					

Table 2: Affine cases for 2-bit independence model.
Objective is $|\mu_1+\mu_2-p|+|\mu_1+\mu_3-q|$

Wasserstein degree

- We can phrase our problem as a parametric optimization problem:

$$\text{minimize } c_1 x_1 + \cdots + c_n x_n \text{ subject to } x \in \mathcal{M} = X \cap \Delta_{n-1}. \qquad (2)$$

The c_i will be functions in d_{ij} and μ_k.

(Here assume min attained at a smooth point, or add equations)

- The optimal value of the problem (2) is a function

$$c_0^* = c_0^*(c_1, \ldots, c_n).$$

By [Rostalski and Sturmfels, 2010, Section 3], it is an algebraic function. This means that there exists a polynomial $\Phi(c_0, c_1, \ldots, c_n)$ in $n+1$ variables such that $\Phi(c_0^*, c_1, \ldots, c_n)=0$.

- The degree of Φ in its first argument c_0 measures the algebraic complexity of our optimization problem (2). We call this number the *Wasserstein degree* of our model $\mathcal{M}$.

Theorem 1

The polynomial $\Phi(-c_0, c_1, \ldots, c_n)$ is the defining equation of the hypersurface $\bar{X}^*$ that is dual (parametrizes hyperplanes in the ambient projective space of $\bar{X}$ that are tangent to $\bar{X}$) to the projective variety $\bar{X}$ that represents the model $\mathcal{M}$ in Δ_{n-1}. Hence the Wasserstein degree of $\mathcal{M}$ is the degree of Φ in its first argument. This is generically equal to the degree of $\bar{X}^*$.

For many natural classes of varieties $\bar{X}$, there are formulas for the degree of the dual $\bar{X}^*$.

Sodomaco gives a formula for the polar degrees of all Segre-Veronese varieties.

Example 2

Suppose that the model $\mathcal{M}$ is a hypersurface, namely, it is the zero set in the simplex Δ_{n-1} of a general polynomial of degree m. Then the Wasserstein degree of $\mathcal{M}$ equals $m(m-1)^{n-2}$.

E.g. for Hardy-Weinberg $n-1 = m = 2$ so it has Wasserstein degree 2.

Since the c_i depend on d and μ, we can consider the function that measures the Wasserstein distance:

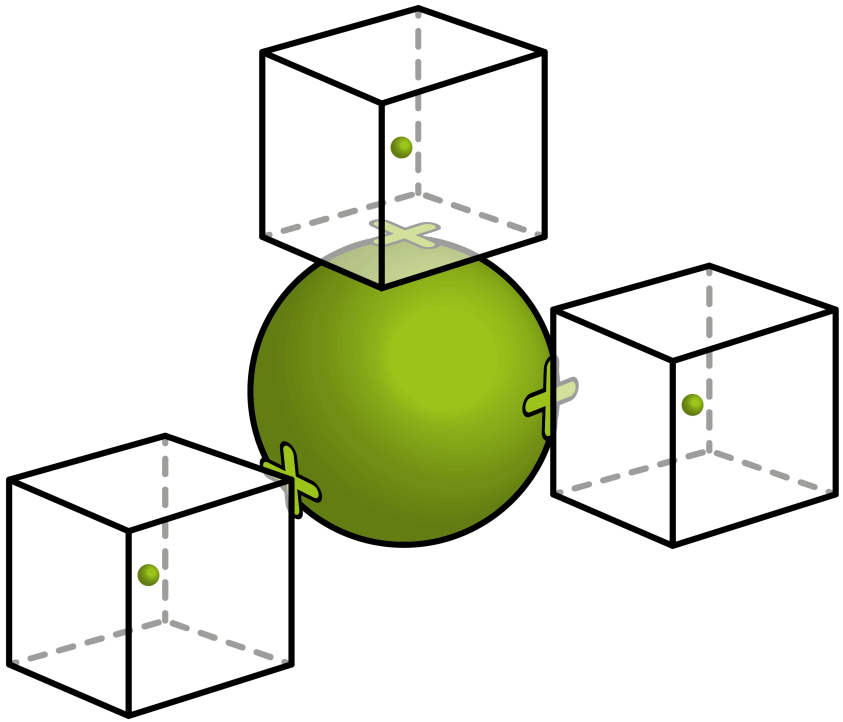
$$\mathbb{R}^{n^2} \times \Delta_{n-1} \to \mathbb{R}, \quad (d, \mu) \mapsto c_0^*(d, \mu) = W_d(\mu, \mathcal{M}).$$

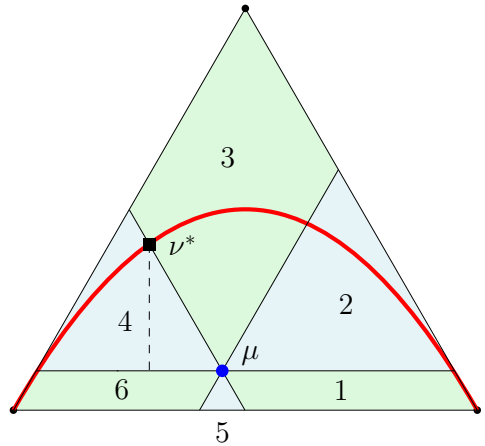
Our discussion establishes the following result about this function which depends only on $\mathcal{M}$.

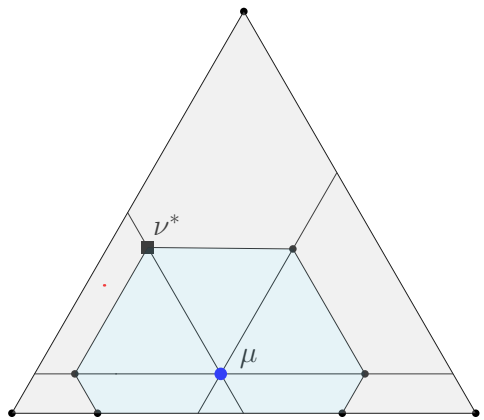
Corollary 3

The Wasserstein distance is a piecewise algebraic function of d and μ. Each piece is an algebraic function whose degree is bounded above by the degree of the hypersurface dual to $\mathcal{M}$.

1. Introduction
2. Excursion
3. Parametric linear programming
4. Polyhedral norm distances




3
ν^*
2
4
μ
6
1
5


ν^*
μ

- Consider the min-max formulation of the Wasserstein distance problem:

$$W_d(\mu, \mathcal{M}) \quad := \quad \min_{\nu \in \mathcal{M}} W_d(\mu, \nu) \quad = \quad \min_{\nu \in \mathcal{M}} \max_{x \in P_d} \langle \mu - \nu, x \rangle. \qquad (3)$$

- The Wasserstein metric on Δ_m defines a polyhedral norm on $\mathbb{R}^m$.
- Consider a unit Wasserstein ball B around the origin, which is a centrally symmetric m-dimensional polytope. It induces a norm on $\mathbb{R}^m$

$$\|y\|_B \quad := \quad \min\{\, \lambda \in \mathbb{R}_{\geq 0} \,:\, y \in \lambda B \,\}.$$

- In terms of the dual polytope $B^* \;=\; \{\, x \in \mathbb{R}^m \,:\, \sup_{z \in B} \langle x, z \rangle \leq 1 \,\}$, the polyhedral norm can be rewritten as

$$\|y\|_B \quad = \quad \min\{\, \lambda \in \mathbb{R}_{\geq 0} \,:\, \sup_{x \in B^*} \langle x, y \rangle \leq \lambda \,\} \quad = \quad \max_{x \in B^*} \langle x, y \rangle.$$

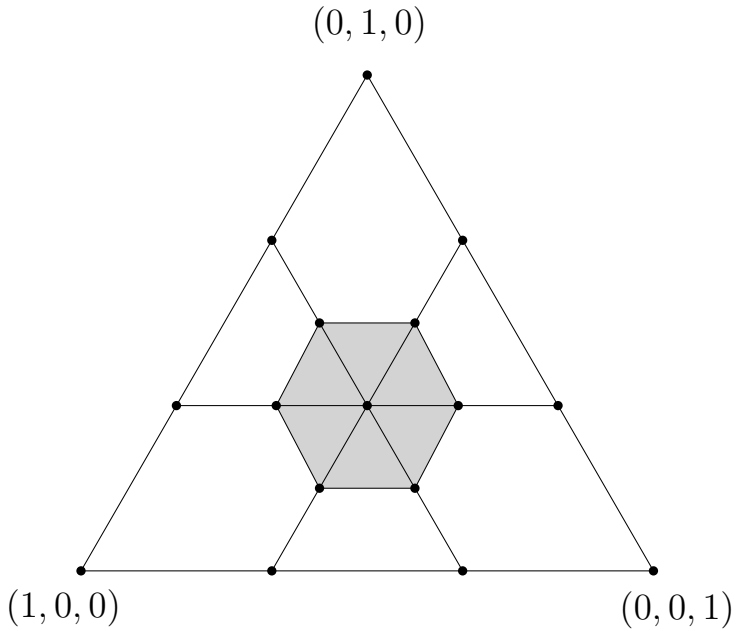
- The dual of the unit ball equals

$$B^* \;=\; P_d \;=\; \Big\{ x \in \mathbb{R}^n / \mathbb{R}\mathbf{1} \;:\; |x_i - x_j| \;\leq\; d_{ij} \;\text{ for all }\; 1 \leq i < j \leq n \Big\}.$$

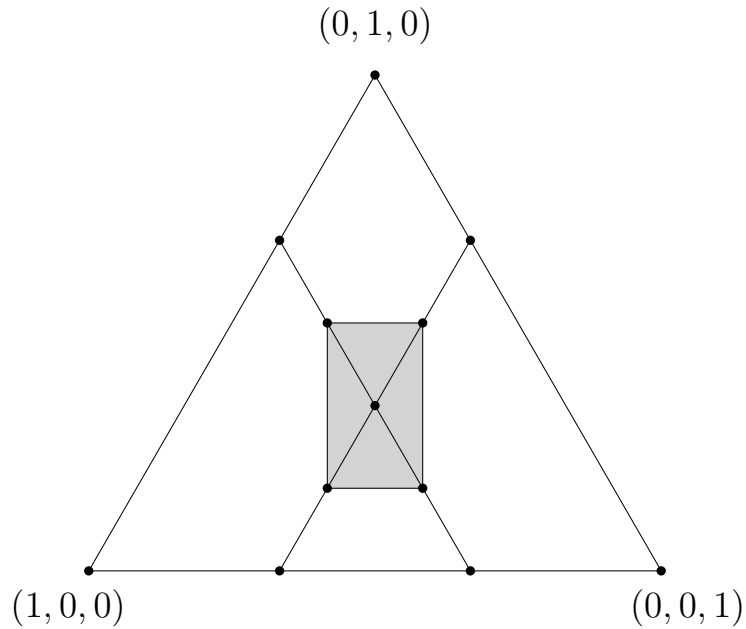
  This is the Lipschitz polytope and the unit ball $B = P_d^*$ is its dual.

- This means that the Wasserstein unit ball B is the convex hull of $n(n-1)$ vectors that lie on a hyperplane in $\mathbb{R}^n$:

$$B \;=\; P_d^* \;=\; \operatorname{conv}\left\{ \frac{1}{d_{ij}}(e_i - e_j) \;:\; 1 \leq i < j \leq n \right\}.$$



$$d = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$



$$d = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix}$$

Figure 2: Wasserstein balls centered in the uniform distribution associated to the discrete metric (left) and the L_1-metric (right) for $n = 3$.

Example 4

Fix $m = n - 1 = 3$ and let d be the 2-bit Hamming metric. We work in the linear space L that is defined by $x_1 + x_2 + x_3 + x_4 = 0$.
The Lipschitz polytope is the octahedron

$$\begin{aligned} P_d = B^* &= \{\, (x_1, x_2, x_3, x_4) \in L \,:\, |x_1 - x_2| \le 1,\ |x_1 - x_3| \le 1,\ |x_2 - x_4| \le 1,\ |x_3 - x_4| \le 1 \,\} \\ &= \operatorname{conv}\{(1,0,0,-1), (1,0,0,-1), (\tfrac{1}{2}, -\tfrac{1}{2}, -\tfrac{1}{2}, \tfrac{1}{2}), (-\tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}, -\tfrac{1}{2}), (0,1,-1,0), (0,-1,1,0)\}. \end{aligned}$$

The Wasserstein unit ball is the cube

$$\begin{aligned} B = P_d^* &= \big\{\, (y_1, y_2, y_3, y_4) \in L \,:\, |y_1 - y_4| \le 1,\ |y_2 - y_3| \le 1,\ |y_2 + y_3| \le 1 \big\} \\ &= \operatorname{conv}\big\{(1,-1,0,0), (1,0,-1,0), (0,1,0,-1), (0,0,1,-1) \\ &\qquad (-1,1,0,0), (-1,0,1,0), (0,-1,0,1), (0,0,-1,1)\big\}. \end{aligned}$$

For any point $u \in \mathbb{R}^m$, we are interested in its distance to the variety under our polyhedral norm:

$$\begin{aligned} D_B(u, \mathcal{M}) :=& \min\{ \|u - v\|_B \,:\, v \in \mathcal{M} \} \\ =& \min\{ \lambda \in \mathbb{R}_{\geq 0} \,:\, (u + \lambda B) \cap \mathcal{M} \neq \emptyset \}. \end{aligned}$$

Proposition 5

If the model $\mathcal{M}$ and the point u are in general position relative to the unit ball B then there is a unique optimal point $v \in \mathcal{M}$ for which $D_B(u, \mathcal{M}) = \|u - v\|_B = \lambda$ holds. The point $\frac{1}{\lambda}(v - u)$ is in the relative interior of a unique face F of the polytope B; we say that v has type F.

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

$$\text{Minimize } \ell_F = \ell_F(\nu) \text{ subject to } \nu \in (\mu + L_F) \cap \mathcal{M}. \tag{4}$$

- This is a polynomial optimization problem in a linear subspace $L_F \subseteq \mathbb{R}^m$.
- Here ℓ_F is any linear functional on $\mathbb{R}^m$ that attains its maximum over B at F. We work in the linear space

$$L_F \;=\; \Big\{ \sum_{(i,j)\in\mathcal{F}} \lambda_{ij}(e_i - e_j) \;:\; \lambda_{ij} \in \mathbb{R} \Big\}. \tag{5}$$

- For a given F, 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.
- We study the algebraic complexity of this problem (polar degrees of independence models) and the combinatorial complexity (facial structure of Wasserstein balls)

Theorem 6 (Hardy-Weinberg)

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

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

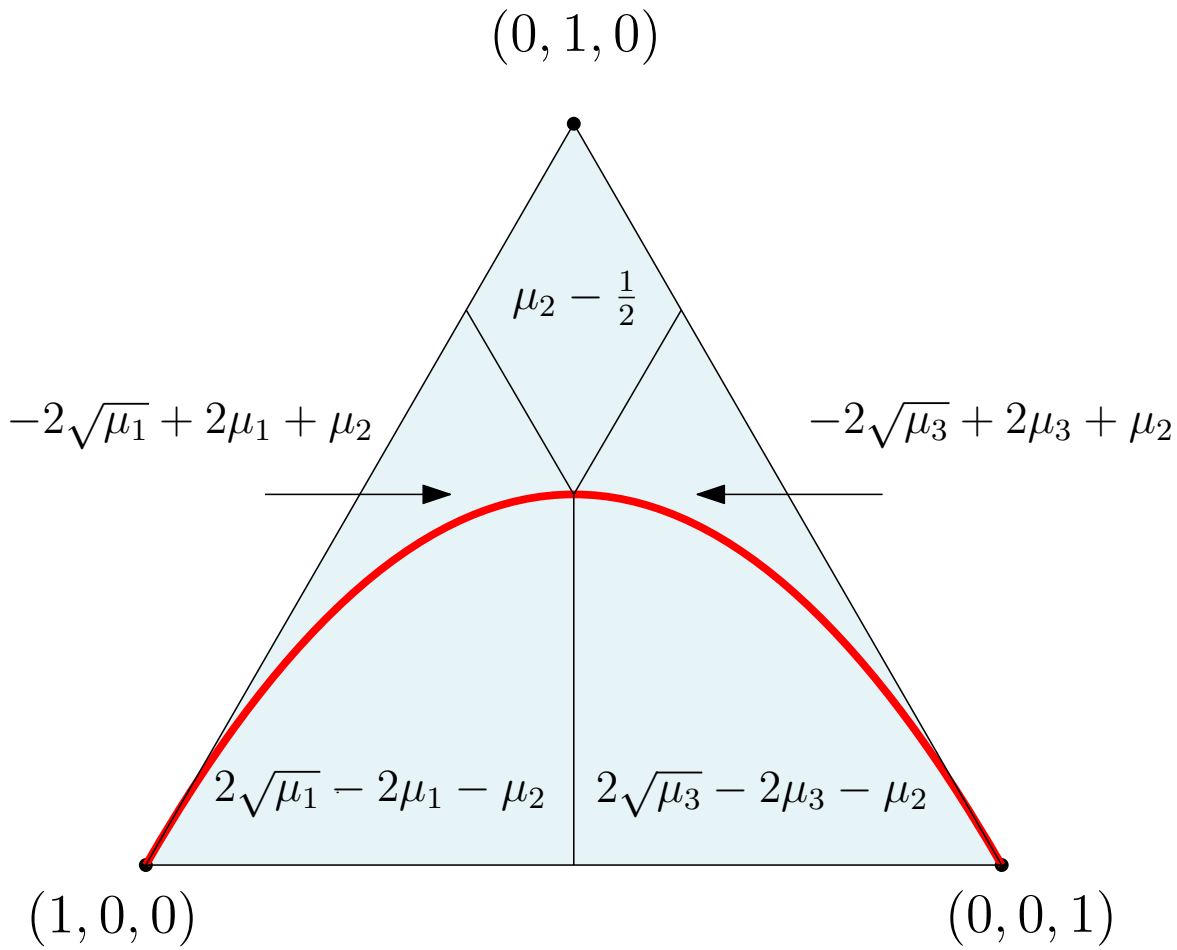


Figure 3: Hardy-Weinberg curve $\mathcal{M}$ is shown in red. The optimal value function for the Wasserstein distance to this curve is piecewise algebraic with five regions.

Theorem 7 (2-bit independence model)

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

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

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

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

The walls of indecision are 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
$\{\mu \in \Delta_3 : \mu_2 - \mu_3 = 0, \mu_1 + \mu_2 \geq \sqrt{\mu_2}, \mu_2 + \mu_4 \geq \sqrt{\mu_2}\}$.

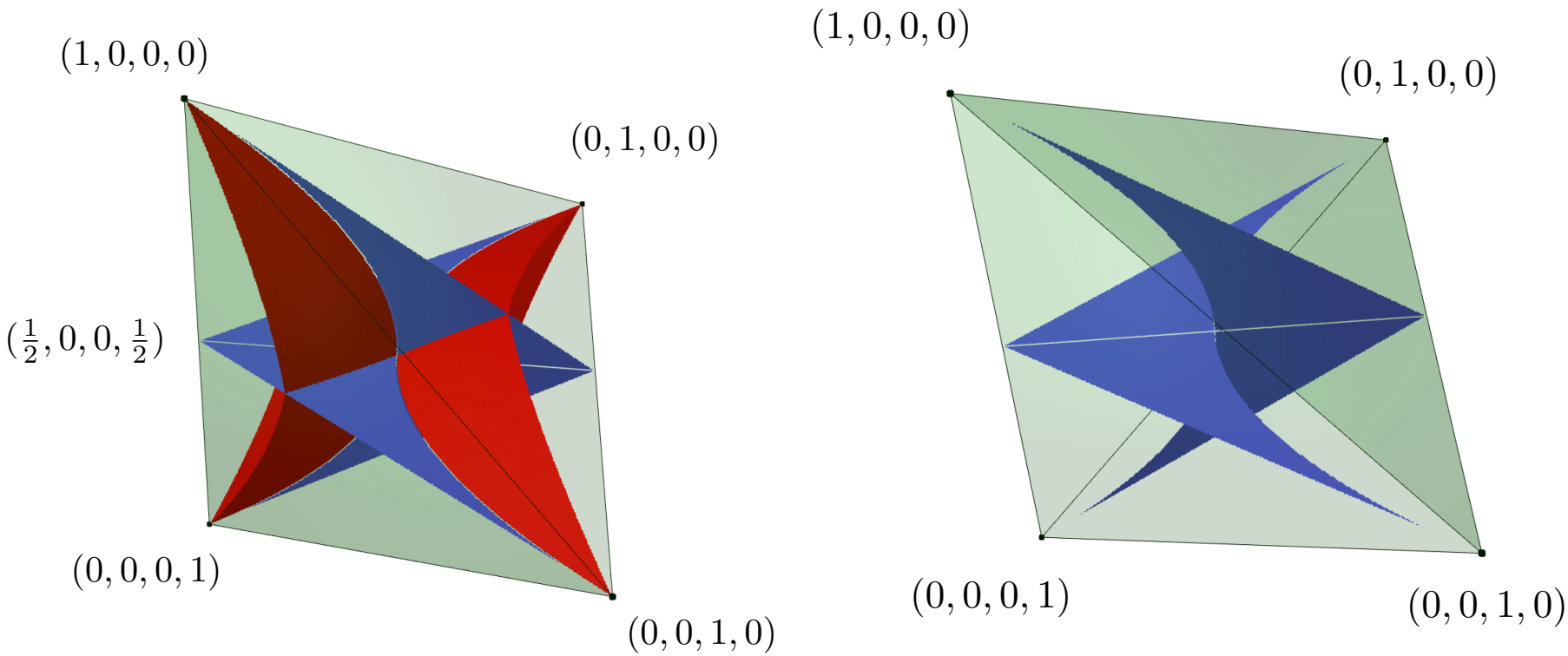
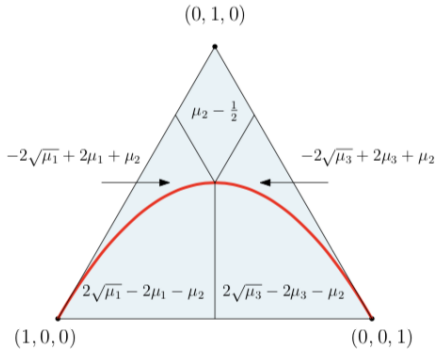
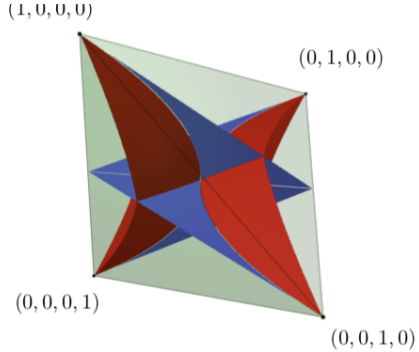


Figure 4: The optimal value function of (8) subdivides the tetrahedron of probability distributions μ (left). The walls of indecision are shown in blue (right).

Experiments



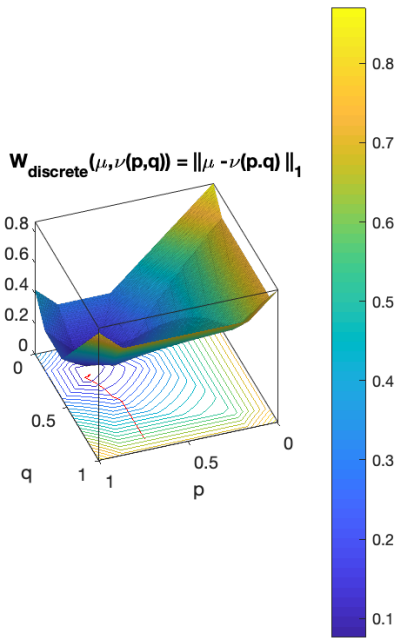


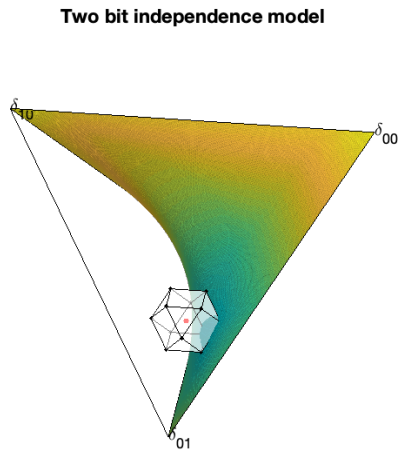
			% of opt. solutions of dim(type) $= i$						
$\mathcal{M}$	d	f-vector	0	1	2	3	4	5	6
$(2, 2)$	L_0	$(8, 12, 6)$	68.6	31.4	0	-	-	-	-
$(2, 2, 2)$	L_0	$(24, 192, 652, 1062, 848, 306, 38)$	0	0	0.1	70.9	27.5	1.5	0
$(2, 3)$	L_0	$(18, 96, 200, 174, 54)$	0	64.1	18.7	17.2	0	-	-
$(2, 3)$	L_1	$(14, 60, 102, 72, 18)$	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	-	-	-

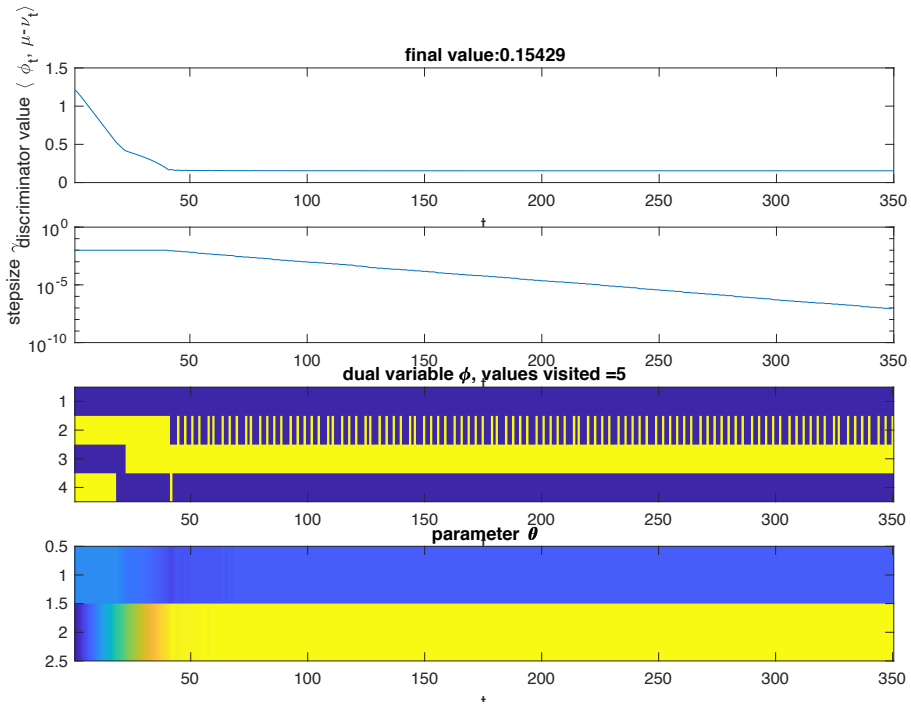
TABLE 6. Distribution of types among optimal solutions for a uniform sample of 1000 points.

Conclusion

- I discussed the problem of minimizing the Wasserstein distance from a given point μ to a variety $\mathcal{M}$
- One perspective is primal nested linear programming over joint distributions, another is dual as a polyhedral norm minimization
- We obtained descriptions of the loss function $\theta \mapsto W_d(\mu, \nu(\theta))$ and the solution function $\mu \mapsto W_d(\mu, \mathcal{M})$
- We discuss further the algebraic complexity in terms of the degree of the dual variety of independence models, and the combinatorial complexity, governed by the facial structure of the Wasserstein ball associated with the metric d.


$W_{discrete}(\mu,\nu(p,q)) = \|\mu - \nu(p.q)\|_1$
0.8
0.6
0.4
0.2
0
0
0.5
1
q
1
0.5
0
p
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1


Two bit independence model
δ_{10}
δ_{00}
δ_{01}


final value:0.15429
discriminator value $\langle \phi_t, \mu - \nu_t \rangle$
1.5
1
0.5
0
50
100
150
200
250
300
350
t
stepsize
10^0
10^{-5}
10^{-10}
50
100
150
200
250
300
350
t
dual variable ϕ, values visited =5
1
2
3
4
50
100
150
200
250
300
350
t
parameter θ
0.5
1
1.5
2
2.5
50
100
150
200
250
300
350
t

References I

Arbel, M., Gretton, A., Li, W., and Montúfar, G. (2020).
Kernelized Wasserstein natural gradient.
In *ICLR 2020 Eighth international conference on learning representations : Millennium Hall, Addis Ababa, Ethiopia*.

Celik, T. O., Jamneshan, A., Montúfar, G., Sturmfels, B., and Venturello, L. (2020a).
Optimal transport to a variety.
In Slamanig, D., Tsigaridas, E., and Zafeirakopoulos, Z., editors, *Mathematical aspects of computer and information sciences : 8th international conference, MACIS 2019, Gebze-Istanbul, Turkey, November 13-15, 2019 ; revised selected papers*, volume 11989 of *Lecture notes in computer science*, pages 364–381. Springer, Cham.

References II


Celik, T. O., Jamneshan, A., Montúfar, G., Sturmfels, B., and Venturello, L. (2020b).
Wasserstein distance to independence models.
Arxiv.


Dukler, Y., Li, W., Lin, A., and Montúfar, G. (2019).
Wasserstein of Wasserstein loss for learning generative models.
In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1716–1725, Long Beach, California, USA. PMLR.


Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2018).
Progressive growing of GANs for improved quality, stability, and variation.
In *International Conference on Learning Representations*.

References III


Ledig, C., Theis, L., Huszar, F., Caballero, J., Aitken, A. P., Tejani, A., Totz, J., Wang, Z., and Shi, W. (2016).
Photo-realistic single image super-resolution using a generative adversarial network.
CoRR, abs/1609.04802.


Li, W., Lin, A. T., and Montúfar, G. (2019).
Affine natural proximal learning.
In Nielsen, F. and Barbaresco, F., editors, *Geometric Science of Information*, pages 705–714, Cham. Springer International Publishing.


Li, W. and Montúfar, G. (2018).
Natural gradient via optimal transport.
Information geometry, 1(2):181–214.

References IV

Lin, A. T., Dukler, Y., Li, W., and Montúfar, G. (2019).
Wasserstein diffusion Tikhonov regularization.
https://www.researchgate.net/publication/335828271_Wasserstein_Diffusion_Tikhonov_Regularization.
Presented at Optimal Transport in Machine Learning Workshop, NeurIPS 2019.

Lin, A. T., Li, W., Osher, S., and Montúfar, G. (2018).
Wasserstein proximal of GANs.
https://www.researchgate.net/profile/Wuchen_Li/publication/327919922_Wasserstein_Proximal_of_GANs/links/5bad24f892851ca9ed2a504b/Wasserstein-Proximal-of-GANs.pdf.

References V


Rostalski, P. and Sturmfels, B. (2010).
Dualities in convex algebraic geometry.
Rendiconti di Matematica, (30):285–327.


Zhang, H., Xu, T., Li, H., Zhang, S., Huang, X., Wang, X., and Metaxas, D. N. (2016).
Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks.
CoRR, abs/1612.03242.


Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017).
Unpaired image-to-image translation using cycle-consistent adversarial networks.
In *Computer Vision (ICCV), 2017 IEEE International Conference on*.