Clustering in Hilbert's Projective Geometry

— The Case Studies of the Probability Simplex and the Elliptope of Correlation Matrices —

Frank Nielsen[1] Ke Sun[2]

[1]Sony Computer Science Laboratories Inc, Japan

[2]CSIRO, Australia

September 2018

@FrnkNlsn

https://arxiv.org/abs/1704.00454

Motivation

- Which geometry and distance are appropriate for a space of probability distributions?
- Historically, information geometry [9] focused on the differential-geometric structures associated to spaces of parametric distributions
- But are there other interesting geometric structures?

Motivations: Probability simplex space

- Many tasks in text mining, machine learning, computer vision deal with multinomial distributions or mixtures thereof, living on the probability simplex

- Which **distance** and **underlying geometry** should we consider for handling the space of multinomials?
  - (L1): Total-variation distance and ℓ_1-norm geometry
  - (FRH): Fisher-Rao distance and spherical Riemannian geometry
  - (IG): Kullback-Leibler/f-divergences and information geometry
  - (HG): Hilbert cross-ratio metric and Hilbert projective geometry

- Benchmark experimentally these different geometries by considering k-means/k-center clusterings of multinomials

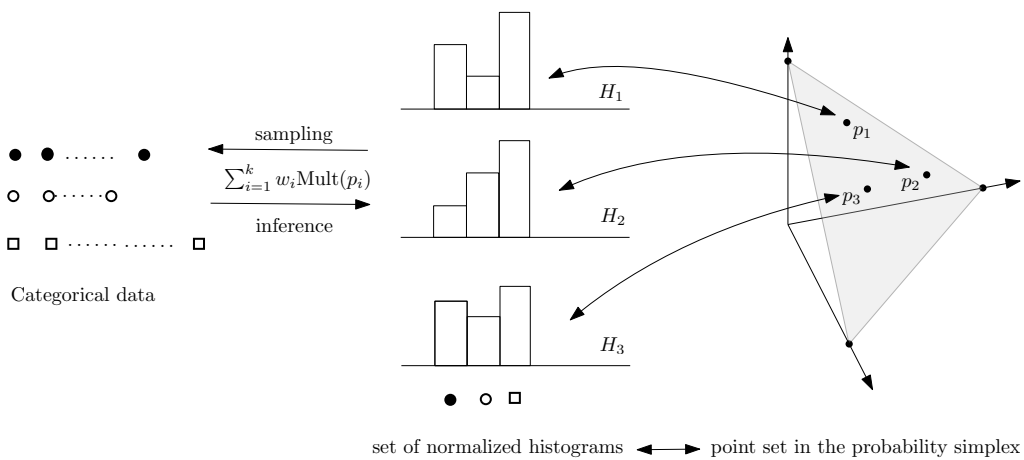
The space of multinomial distributions

- Probability simplex:
  $\Delta_d := \{p = (\lambda_p^0, \ldots, \lambda_p^d) \ : \ p_i > 0, \sum_{i=0}^d \lambda_p^i = 1\}$.

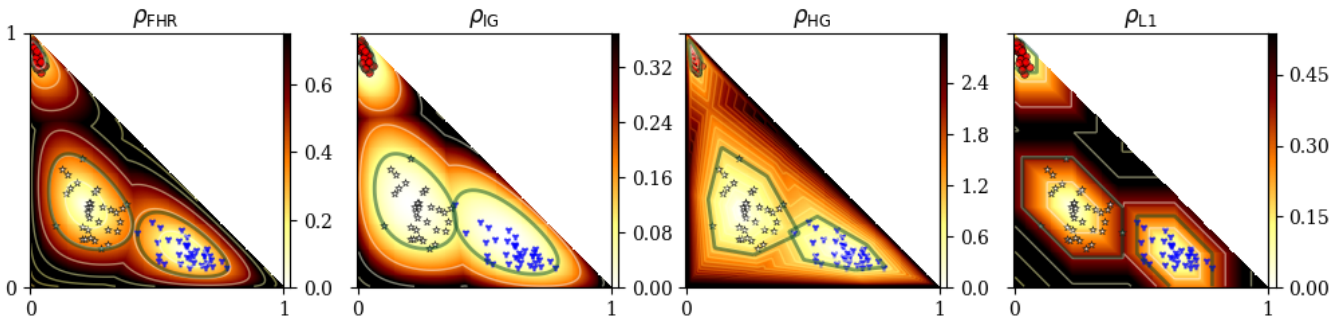
- Multinomial distribution
  $X = (X_0, \ldots, X_d) \sim \mathrm{Mult}(p = (\lambda_p^0, \ldots, \lambda_p^d), m)$ with probability mass function:

$$\Pr(X_0 = m_0, \ldots, X_d = m_d) = \frac{m!}{\prod_{i=0}^d m_i!} \prod_{i=0}^d \left(\lambda_p^i\right)^{m_i}, \sum_{i=0}^d m_i = m$$

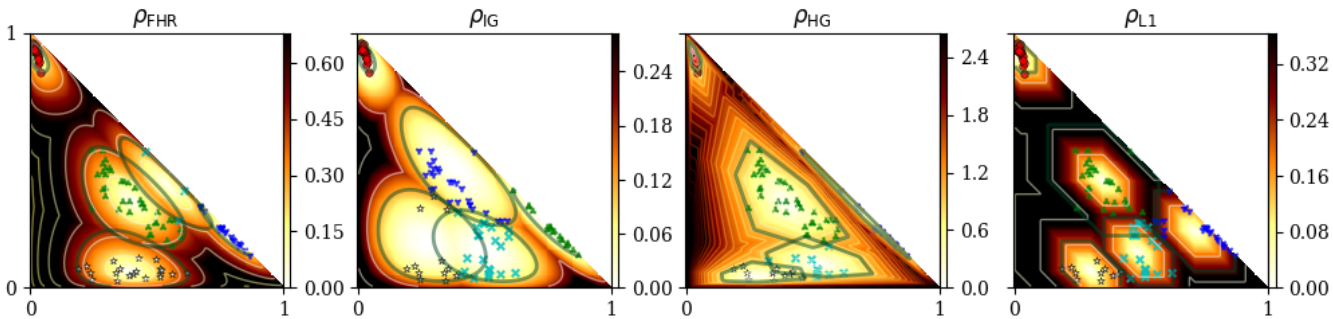
  Binomial ($d = 1$), Bernoulli ($d = 1$, $m = 1$), multinoulli/categorical/discrete ($m = 1$ and $d > 1$)



Teaser: k-center clustering in the probability simplex Δ_d



$k = 3$ clusters



$k = 5$ clusters

L1	:	Total-variation distance ρ_{L1} (ℓ_1-norm geometry)
FHR	:	Fisher-Hotelling-Rao distance ρ_{FHR} (spherical geometry)
IG	:	Kullback-Leibler divergence ρ_{IG} (information geometry)
HG	:	Hilbert metric ρ_{HG} (Hilbert projective geometry)

Review of four geometries for Δ_d

Fisher-Hotelling-Rao Riemannian geometry (1930)

- Fisher metric tensor $[g_{ij}]$ (constant positive curvature):

$$g_{ij}(p) = \frac{\delta_{ij}}{\lambda_p^i} + \frac{1}{\lambda_p^0}.$$

  Statistical invariance: unique Fisher metric [2]

- Distance is a geodesic length distance (Levi-Civita connection ∇^{LC} with metric-compatible parallel transport)

- Rao metric distance between two multinomials on Riemannian manifold (Δ_d, g):

$$\rho_{\mathrm{FHR}}(p, q) = 2 \arccos\left(\sum_{i=0}^{d} \sqrt{\lambda_p^i \lambda_q^i}\right).$$

- Can be embedded in the positive orthant of an Euclidean d-sphere of $\mathbb{R}^{d+1}$ by using the *square root representation* $\lambda \mapsto \sqrt{\lambda} = (\sqrt{\lambda^0}, \ldots, \sqrt{\lambda^d})$

Kullback-Leibler divergence and information geometry

- Kullback-Leibler (KL) divergence ρ_{IG}:

$$\rho_{\mathrm{IG}}(p,q) = \sum_{i=0}^{d} \lambda_p^i \log \frac{\lambda_p^i}{\lambda_q^i}.$$

- Satisfies the information monotonicity (f-divergence with $f(u) = -\log u$) when coarse-graining $\Delta_d \to \Delta_{d'}$ (with $d' < d$): $0 \leq I_f(p' : q') \leq I_f(p : q)$. Fisher-Rao ρ_{FHR} does not satisfy the information monotonicity

- Dualistic structure: dually flat manifold $(\Delta_d, g, \nabla^m, \nabla^e)$ (exponential connection ∇^e and mixture connection ∇^m, with $\nabla^{\mathrm{LC}} = \frac{\nabla^m + \nabla^e}{2}$), or expected α-geometry $(\Delta_d, g, \nabla^{-\alpha}, \nabla^{\alpha})$ for $\alpha \in \mathbb{R}$. Dual parallel transport that is metric compatible.

- Δ_d is both an exponential family and a mixture family (dually flat space)

Total variation distance and ℓ_1-norm geometry

- Total Variation metric distance (L1) divergence ρ_{L1}:

$$\rho_{L1}(p,q) = \sum_{i=0}^{d} |\lambda_p^i - \lambda_q^i|$$

- Satisfies information monotonicity (f-divergence for $f(u) = \frac{1}{2}|u-1|$):

$$0 \leq \rho_{L1}(p',q') \leq \rho_{L1}(p,q)$$

- ℓ_1-norm-induced distance:

$$\rho_{L1}(p,q) = \|\lambda_p - \lambda_q\|_1$$

  Total variation is $\rho_{\mathrm{TV}}(p,q) = \frac{1}{2}\|\lambda_p - \lambda_q\|_1 \leq 1$

- Chentsov [4], Theorem 6.3 p. 88: TV unique metric distance invariant in Markov geometry

The shape of ℓ_1-balls in Δ_d

- For trinomials in Δ_2, the ρ_{L1} distance is given by:

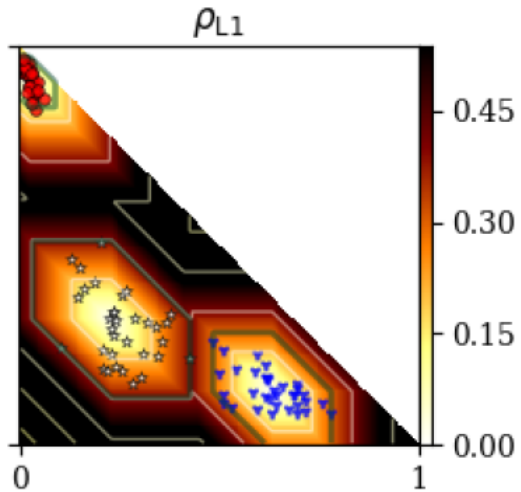
$$\rho_{L1}(p,q) = |\lambda_p^0 - \lambda_q^0| + |\lambda_p^1 - \lambda_q^1| + |\lambda_q^0 - \lambda_p^0 + \lambda_q^1 - \lambda_p^1|.$$

- ℓ_1-distance is polytopal: cross-polytope $\mathcal{Z} = \mathrm{conv}(\pm e_i :\in\in [d])$ also called co-cube (orthoplex)

- Shape:

$$\mathrm{Ball}_{L_1}(p,r) = (\lambda_p \oplus r\mathcal{Z}) \cap H_{\Delta^d}$$

In 2D, regular octahedron cut by H_{Δ^2} = hexagonal shapes.



Shape of L_1 balls in Δ^d

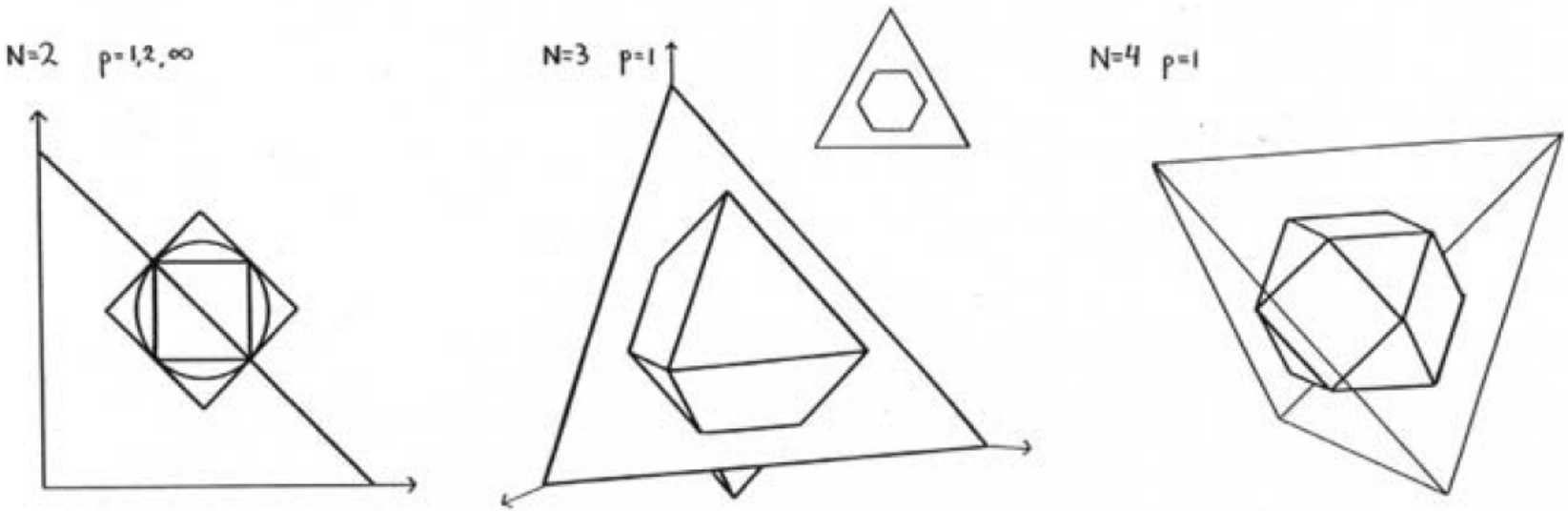


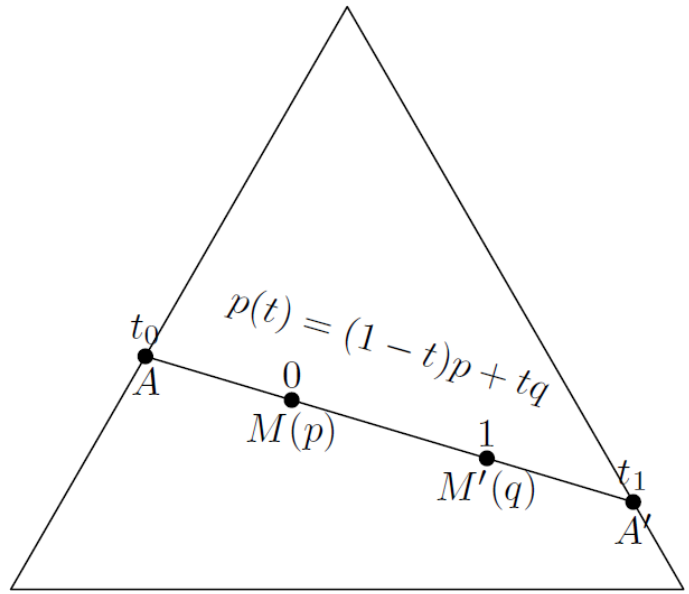
Figure 1.14. For $N = 2$ we show why all the l_p-distances agree when the definition (Eq. 1.55) is used. For $N = 3$ the l_1-distance gives hexagonal 'spheres', arising as the intersection of the simplex with an octahedron. For $N = 4$ the same construction gives an Archimedean solid known as the cuboctahedron.

From 'Geometry of Quantum States', [1]

Hilbert metric distance and projective Hilbert geometry

- Log cross ratio distance:

$$\rho_{\mathrm{HG}}(M, M') = \begin{cases} \left| \log \frac{|A'M||AM'|}{|A'M'||AM|} \right|, & M \neq M', \\ 0 & M = M'. \end{cases}$$

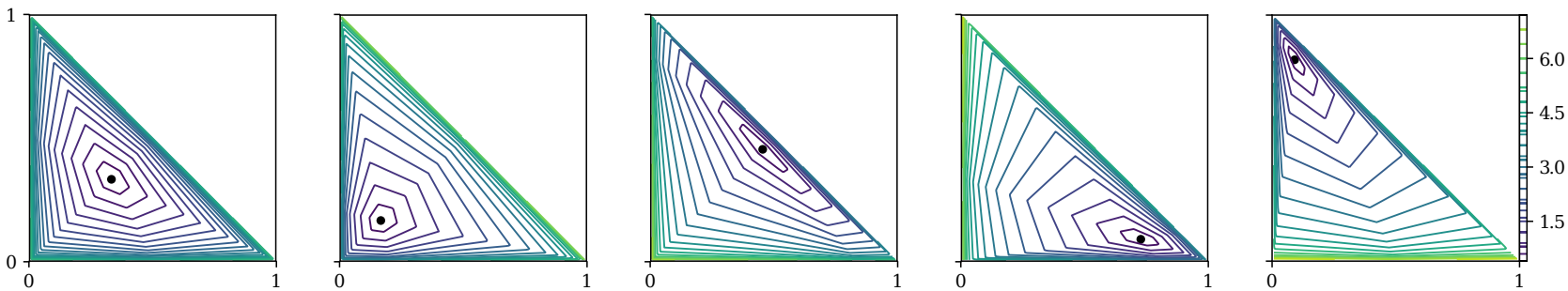


- Geodesics are Euclidean line segments

- Hilbert Geometry (HG) holds for any **convex compact subset** domain (e.g., Δ_d, elliptope of correlation matrices)

Euclidean shape of balls in Hilbert projective geometry

- Hilbert balls in Δ_d: hexagons shapes (2D) [14], rhombic dodecahedra (3D), polytopes [7] with $d(d+1)$ facets in dimension d.

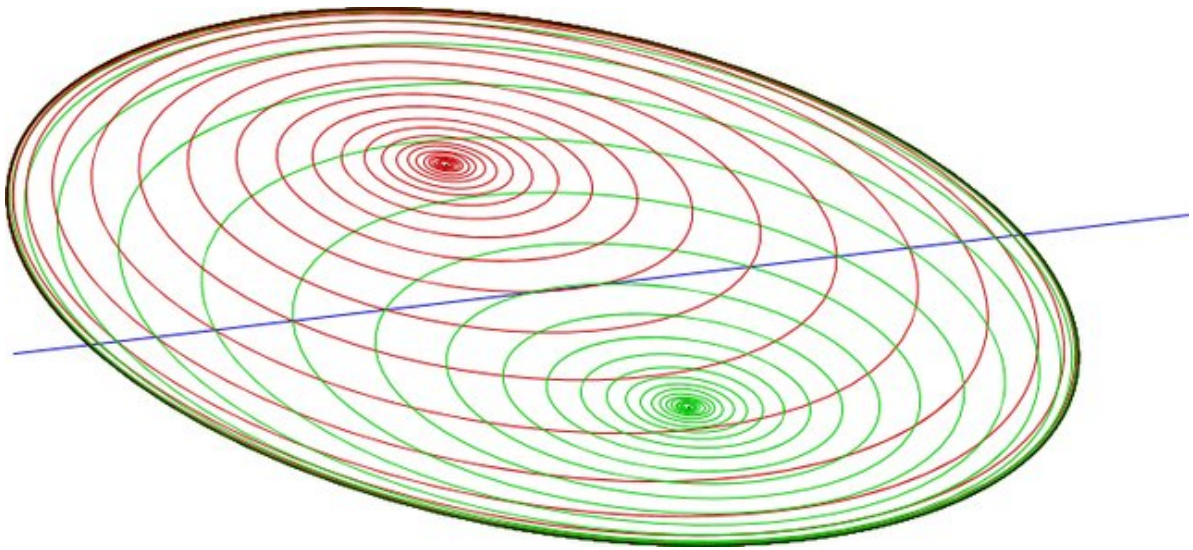


- Not Riemannian/differential-geometric geometry because infinitesimally small balls have not ellipsoidal shapes. (Tissot indicatrix in Riemannian geometry)

- A video explaining the shapes of Hilbert balls: `https://www.youtube.com/watch?v=XE5x5rAK8Hk&t=1s`

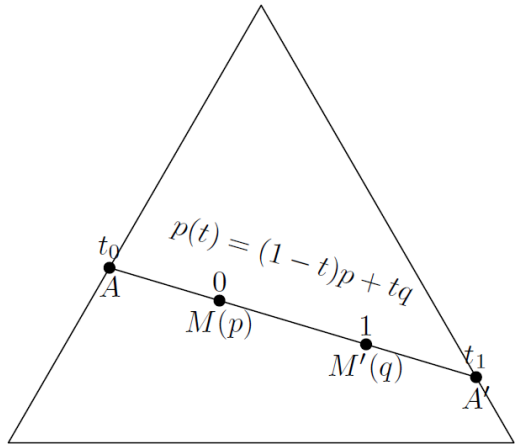
Hilbert geometry generalized Cayley-Klein hyperbolic geometry

- Defined for a quadric domain (curved Mahalanobis distance [10])



- Video: `https://www.youtube.com/watch?v=YHJLq3-RL58`

Computing Hilbert metric distance in 1D



- M ($t_0 \leq 0$) and M' ($t_1 \geq 1$): intersection points of the line $(1-t)p + tq$ with $\partial\Delta^d$.

- Expression using t_0 and t_1:

$$\rho_{\mathrm{HG}}(p,q) = \left|\log \frac{(1-t_0)t_1}{(-t_0)(t_1-1)}\right| = \log\left(1-\frac{1}{t_0}\right) - \log\left(1-\frac{1}{t_1}\right).$$

- Hilbert distance in the simplex of Bernoulli distributions:

$$\rho_{\mathrm{HG}}(p,q) = \left|\log \frac{\lambda_q(1-\lambda_p)}{\lambda_p(1-\lambda_q)}\right| = \left|\log \frac{\lambda_p}{1-\lambda_p} - \log \frac{\lambda_q}{1-\lambda_q}\right|.$$

Comparisons of four main distances for 1D case

- Rao distance (circle arc length) of Riemanian geometry:

$$\rho_{\mathrm{FHR}}(p,q) = 2\arccos\left(\sqrt{\lambda_p\lambda_q} + \sqrt{(1-\lambda_p)(1-\lambda_q)}\right).$$

- KL divergence of the dually flat ± 1-geometry:

$$\rho_{\mathrm{IG}}(p,q) = \lambda_p \log\frac{\lambda_p}{\lambda_q} + (1-\lambda_p)\log\frac{1-\lambda_p}{1-\lambda_q}.$$

- L1 distance for norm geometry:

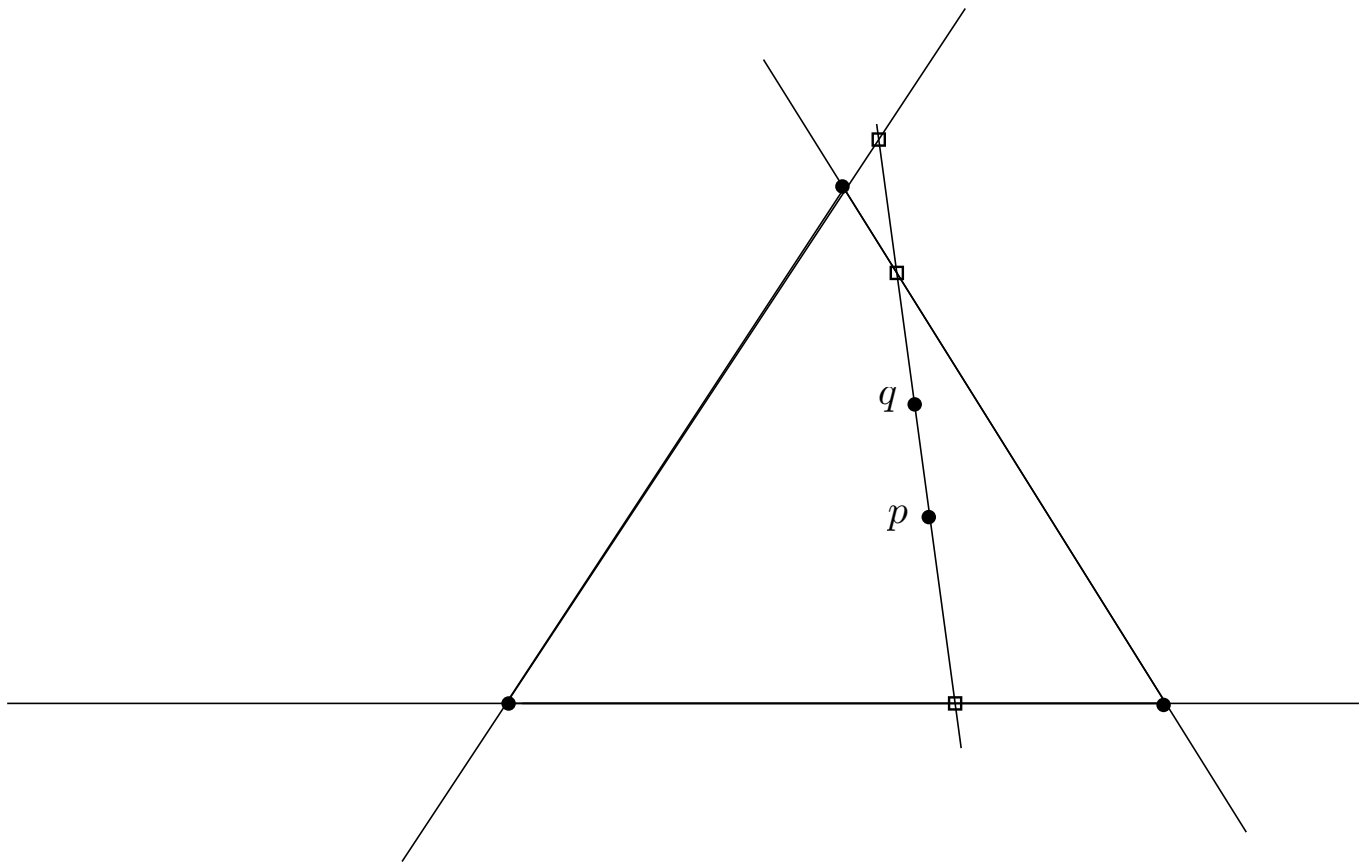
$$\rho_{\mathrm{IG}}(p,q) = |\lambda_p - \lambda_q|$$

- Hilbert distance of projective geometry:

$$\rho_{\mathrm{HG}}(p,q) = \left|\log\frac{\lambda_p}{1-\lambda_p} - \log\frac{\lambda_q}{1-\lambda_q}\right|.$$

Computing Hilbert metric distance in Δ_d

- Naive algorithm: Compute the intersection of line (pq) with each **hyperplane** supporting one of the $d+1$ facets of the probability simplex Δ_d, and check whether this intersection point falls inside this facet



Computing Hilbert metric distance in Δ_d

- Compute intersection of line $x(t) = (1-t)p + tq$ with each supporting hyperplane $H_{\setminus j}$ of a facet $f_{\setminus j}$ $(j = 0, \cdots, d)$. These intersection points are represented using the real values t. Any intersection point with $H_{\setminus j}$ must satisfy either $t \leq 0$ or $t \geq 1$.

- The two intersection points of (pq) with $\partial\Delta^d$:

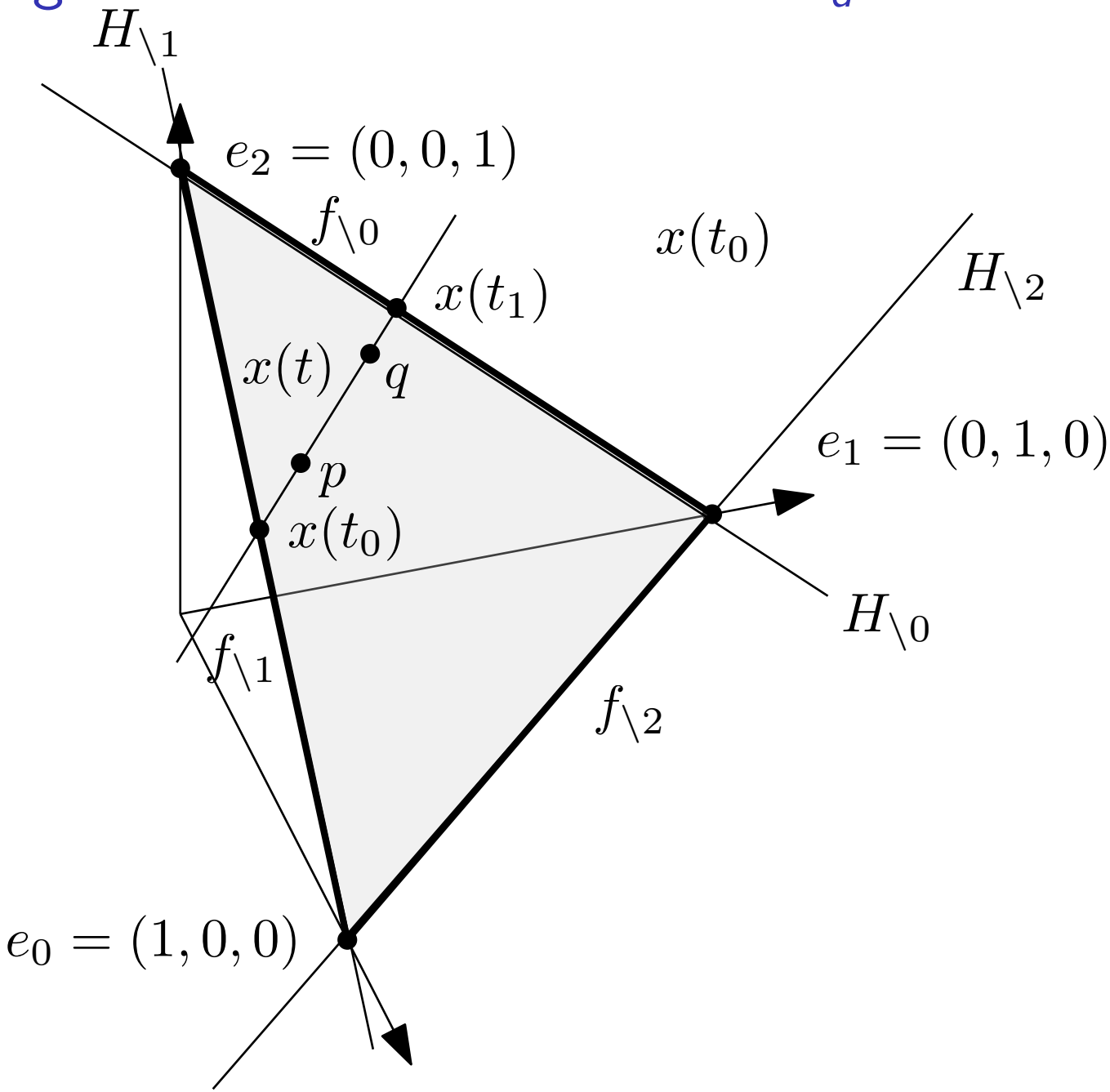
$$t_0 = \max\{t : \exists j,\ x(t) \in H_{\setminus j} \text{ and } t \leq 0\}$$

$$t_1 = \min\{t : \exists j,\ x(t) \in H_{\setminus j} \text{ and } t \geq 1\}$$

- Apply formula:
$\rho_{\mathrm{HG}}(p,q) = \left|\log \frac{(1-t_0)t_1}{(-t_0)(t_1-1)}\right| = \log\left(1 - \frac{1}{t_0}\right) - \log\left(1 - \frac{1}{t_1}\right)$

- Cost $O(d)$ time

Computing Hilbert metric distance in Δ_d: Illustration



Computing Hilbert metric distance in Δ_d

Algorithm 1: Computing the Hilbert distance

Data: Two points $p = (\lambda_p^0, \cdots, \lambda_p^d)$, $q = (\lambda_q^0, \cdots, \lambda_q^d)$ in the d-dimensional simplex Δ^d
Result: Their Hilbert distance $\rho_{\mathrm{HG}}(p,q)$

```
begin
    t_0 ← −∞; t_1 ← +∞;
    for i = 0 ··· d do
        if λ_p^i ≠ λ_q^i then
            t ← λ_p^i / (λ_p^i − λ_q^i);
            if t_0 < t ≤ 0 then
                t_0 ← t;
            else if 1 ≤ t < t_1 then
                t_1 ← t;
    if t_0 = −∞ or t_1 = +∞ then
        Output ρ_HG(p,q) = 0;
    else if t_0 = 0 or t_1 = 1 then
        Output ρ_HG(p,q) = ∞;
    else
        Output ρ_HG(p,q) = |log(1 − 1/t_0) − log(1 − 1/t_1)|;
```

This algorithm requires $O(d)$ time and $O(1)$ memory.

Shape of Hilbert balls in an open simplex

- A ball in the Hilbert simplex geometry has a Euclidean polytope shape with $d(d+1)$ facets [7]
- In 2D, Hilbert balls have $2(2+1) = 6$ hexagonal shapes varying with center locations. In 3D, the shape of balls are rhombic-dodecahedron
- When the domain is not simplicial, Hilbert balls can have varying complexities [14]
- No metric tensor in Hilbert geometry because infinitesimally the ball shapes are polytopes and not ellipsoids

Isometry of Hilbert simplex geometry with a normed vector space $(\Delta^d, \rho_{\mathrm{HG}}) \cong (V^d, \|\cdot\|_{\mathrm{NH}})$

- $V^d = \{v \in \mathbb{R}^{d+1} \ : \ \sum_i v^i = 0\} \subset \mathbb{R}^{d+1}$
- Map $p = (\lambda^0, \ldots, \lambda^d) \in \Delta^d$ to $v(x) = (v^0, \ldots, v^d) \in V^d$:

$$v^i = \frac{1}{d+1}\left(d \log \lambda^i - \sum_{j \neq i} \log \lambda^j\right) = \log \lambda^i - \frac{1}{d+1}\sum_j \log \lambda^j.$$

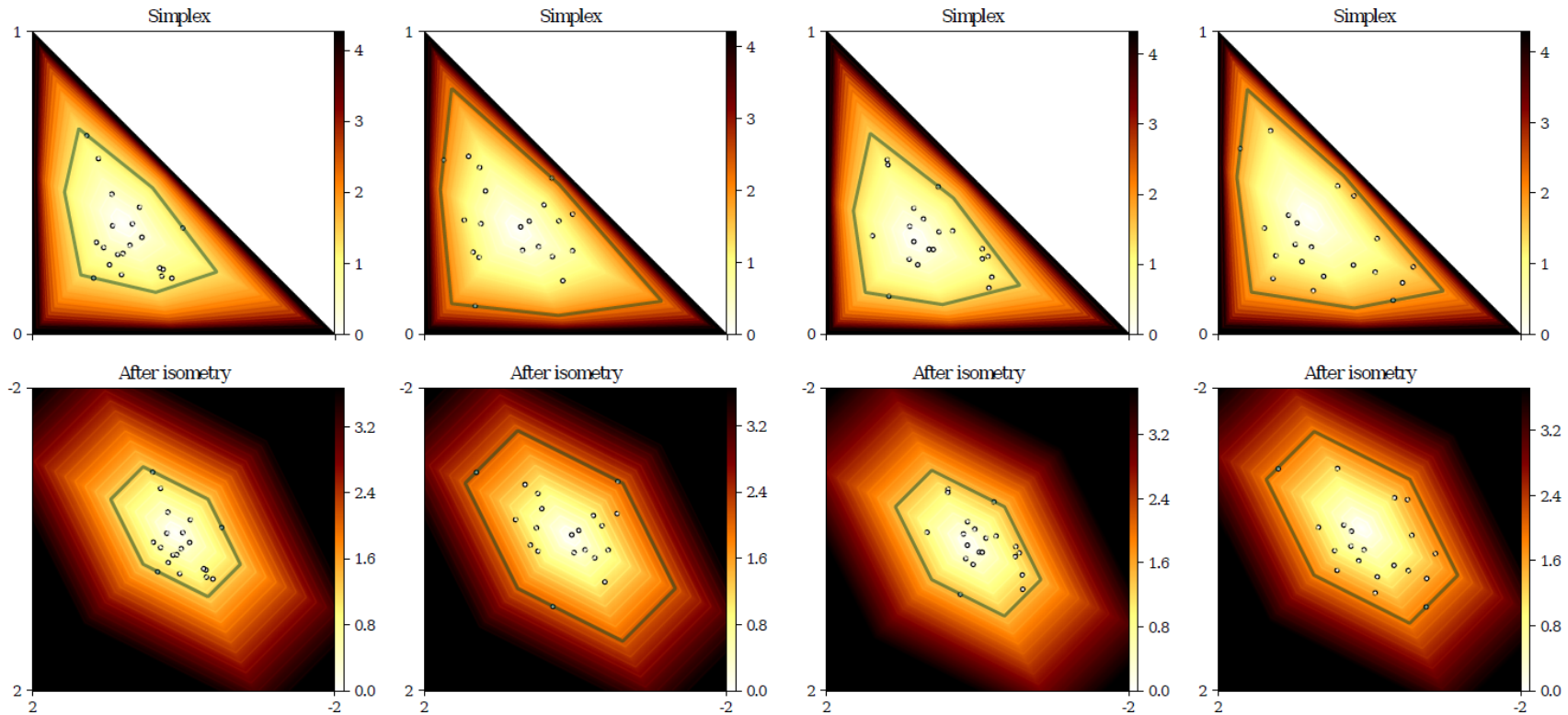
$$\lambda^i = \frac{\exp(v^i)}{\sum_j \exp(v^j)}.$$

- Norm $\|\cdot\|_{\mathrm{NH}}$ in V^d defined by the shape of its unit ball $B_V = \{v \in V^d \ : \ |v^i - v^j| \leq 1, \forall i \neq j\}$.
- Polytopal norm-induced distance:

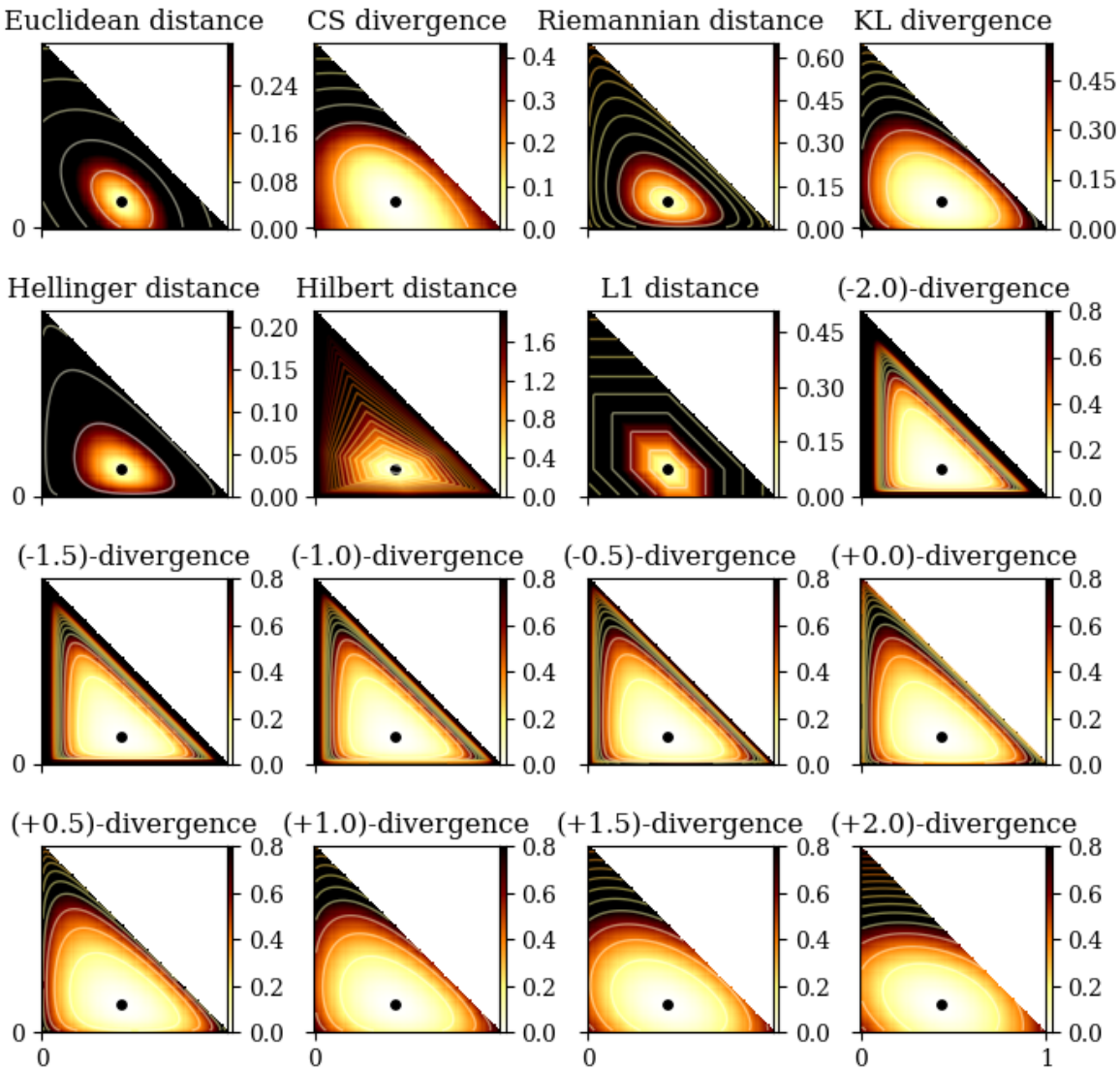
$$\rho_V(v, v') = \|v - v'\|_{\mathrm{NH}} = \inf\left\{\tau \ : \ v' \in \tau(B_V \oplus \{v\})\right\},$$

- Norm does not satisfy parallelogram law (no inner product)

Visualizing the isometry: $(\Delta^d, \rho_{\mathrm{HG}}) \cong (V^d, \|\cdot\|_{\mathrm{NH}})$

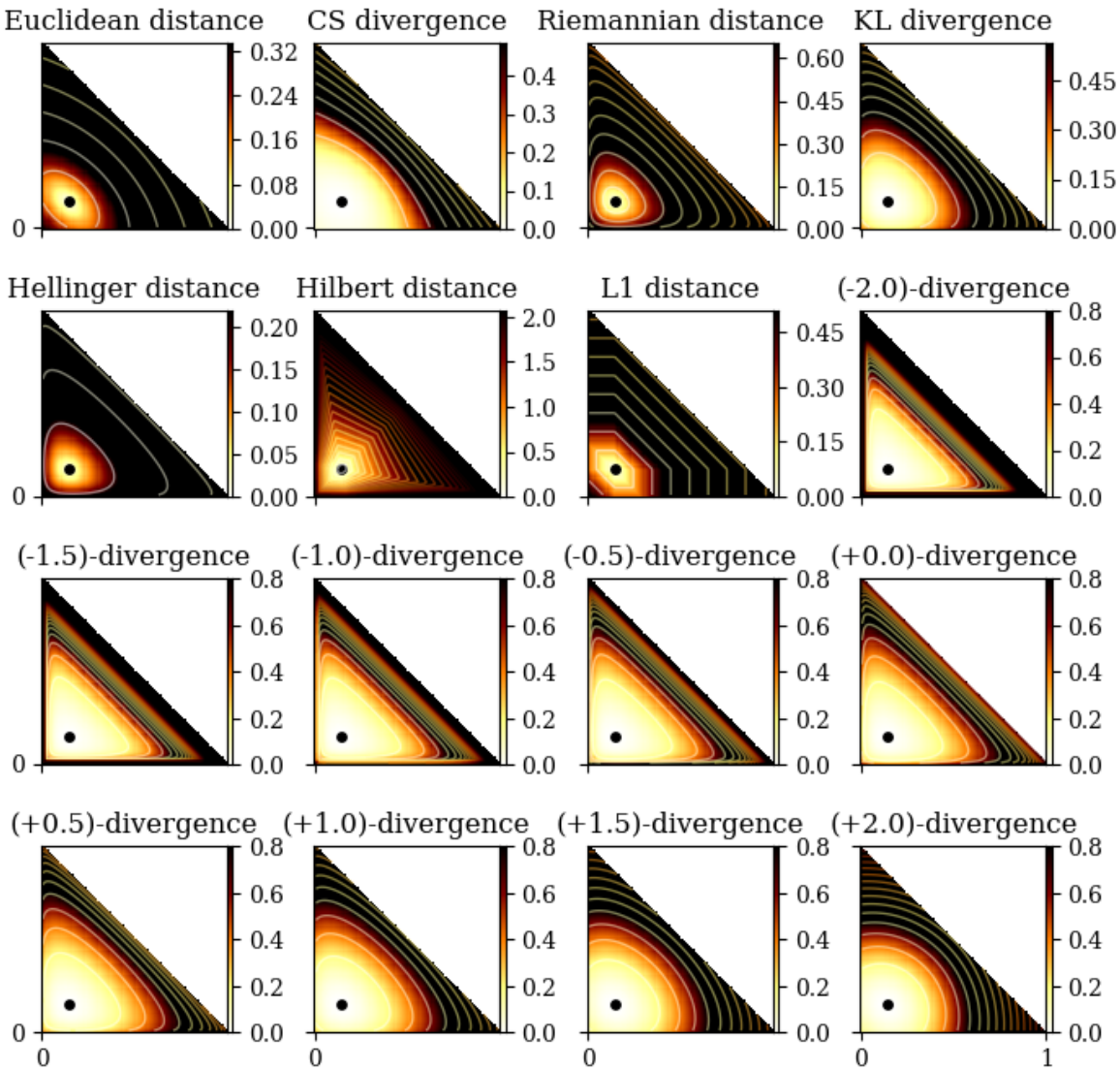


Some distance profiles in Δ_d



Reference point (3/7,3/7,1/7)

Some distance profiles in Δ_d



Reference point (5/7,1/7,1/7)

Benchmarking these geometries with center-based clustering

k-means++ clustering (seeding initialization)

- For an arbitrary distance D, define the k-means objective function (NP-hard to minimize):

$$\boxed{E_D(\Lambda, C) = \frac{1}{n} \sum_{i=1}^{n} \min_{j \in \{1,\ldots,k\}} D(p_i : c_j)}$$

- k-means++: Pick uniformly at random at first seed c_1, and then iteratively choose the $(k-1)$ remaining seeds according to the following probability distribution:

$$\Pr(c_j = p_i) = \frac{D(p_i, \{c_1, \ldots, c_{j-1}\})}{\sum_{i=1}^{n} D(p_i, \{c_1, \ldots, c_{j-1}\})} \quad (2 \leq j \leq k).$$

- A randomized seeding initialization of k-means, can be further locally optimized using Lloyd's batch iterative updates, or Hartigan's single-point swap heuristic [13], etc.

Performance analysis of k-means++ [12]

Let κ_1 and κ_2 be two constants such that κ_1 defines the quasi-triangular inequality property:

$$D(x:z) \leq \kappa_1 \left(D(x:y) + D(y:z) \right), \quad \forall x, y, z \in \Delta^d,$$

and κ_2 handles the symmetry inequality:

$$D(x:y) \leq \kappa_2 D(y:x), \quad \forall x, y \in \Delta^d.$$

Theorem

The generalized k-means++ seeding guarantees with high probability a configuration C of cluster centers such that:

$$\boxed{E_D(\Lambda, C) \leq 2\kappa_1^2(1+\kappa_2)(2+\log k)E_D^*(\Lambda, k)}$$

Performance analysis of k-means++ in a metric space

In any metric space $(\mathcal{X}, d)$, the k-means++ wrt the squared metric distance d^2 is $16(2 + \log k)$-competitive.

Proof: Symmetric distance ($\kappa_2 = 1$). Quasi-triangular inequality property:

$$\begin{aligned} d(p,q) &\leq d(p,q) + d(q,r), \\ d^2(p,q) &\leq (d(p,q) + d(q,r))^2, \\ d^2(p,q) &\leq d^2(p,q) + d^2(q,r) + 2d(p,q)d(q,r). \end{aligned}$$

Let us apply the **inequality of arithmetic and geometric means**:

$$\sqrt{d^2(p,q)d^2(q,r)} \leq \frac{d^2(p,q) + d^2(q,r)}{2}.$$

Thus we have

$$d^2(p,q) \leq d^2(p,q)+d^2(q,r)+2d(p,q)d(q,r) \leq 2(d^2(p,q)+d^2(q,r)).$$

That is, the squared metric distance satisfies the 2-approximate triangle inequality, and $\kappa_1 = 2$.

Performance analysis of k-means++

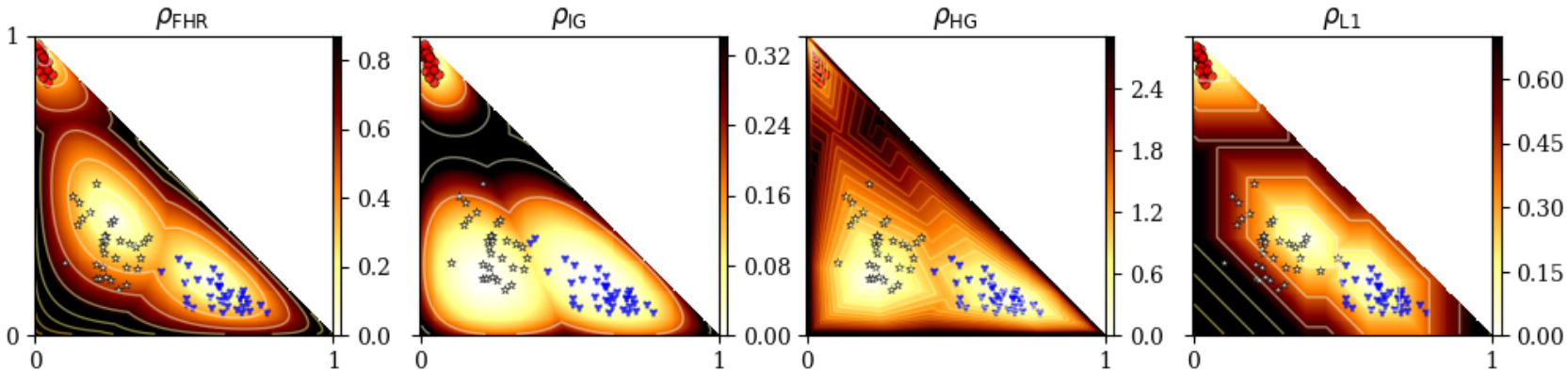
- In any normed space $(\mathcal{X}, \|\cdot\|)$, the k-means++ with $D(x, y) = \|x - y\|^2$ is $16(2 + \log k)$-competitive.
- In any inner product space $(\mathcal{X}, \langle\cdot,\cdot\rangle)$, the k-means++ with $D(x, y) = \langle x - y, x - y\rangle$ is $16(2 + \log k)$-competitive.
- Hilbert simplex geometry is isometric to a normed vector space [7] (Theorem 3.3):

$$(\Delta^d, \rho_{\mathrm{HG}}) \simeq (V^d, \|\cdot\|_{\mathrm{NH}})$$

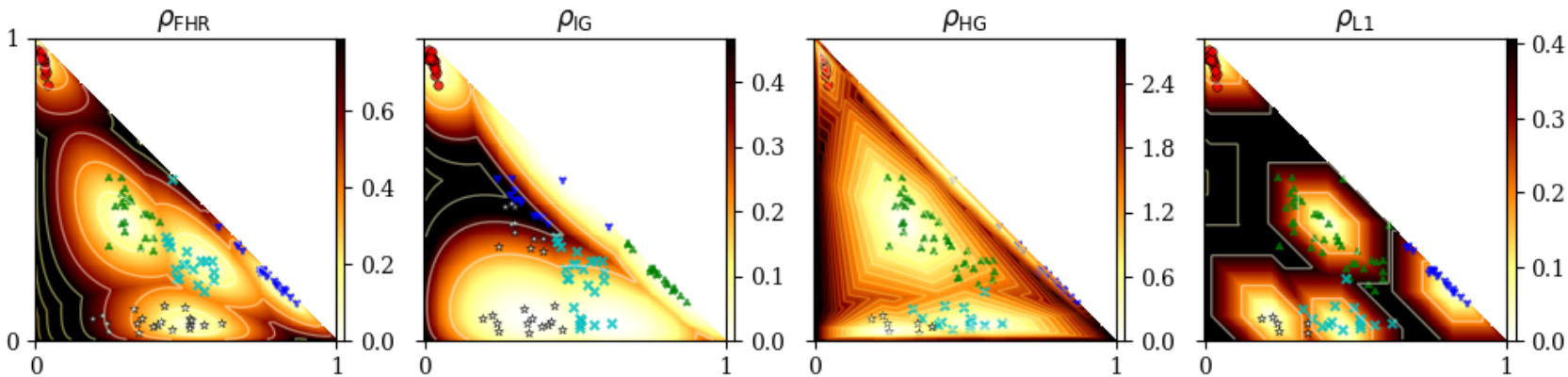
Theorem

k-means++ seeding in a Hilbert simplex geometry in fixed dimension is $16(2 + \log k)$-competitive.

Clustering in Δ_2 with k-means++



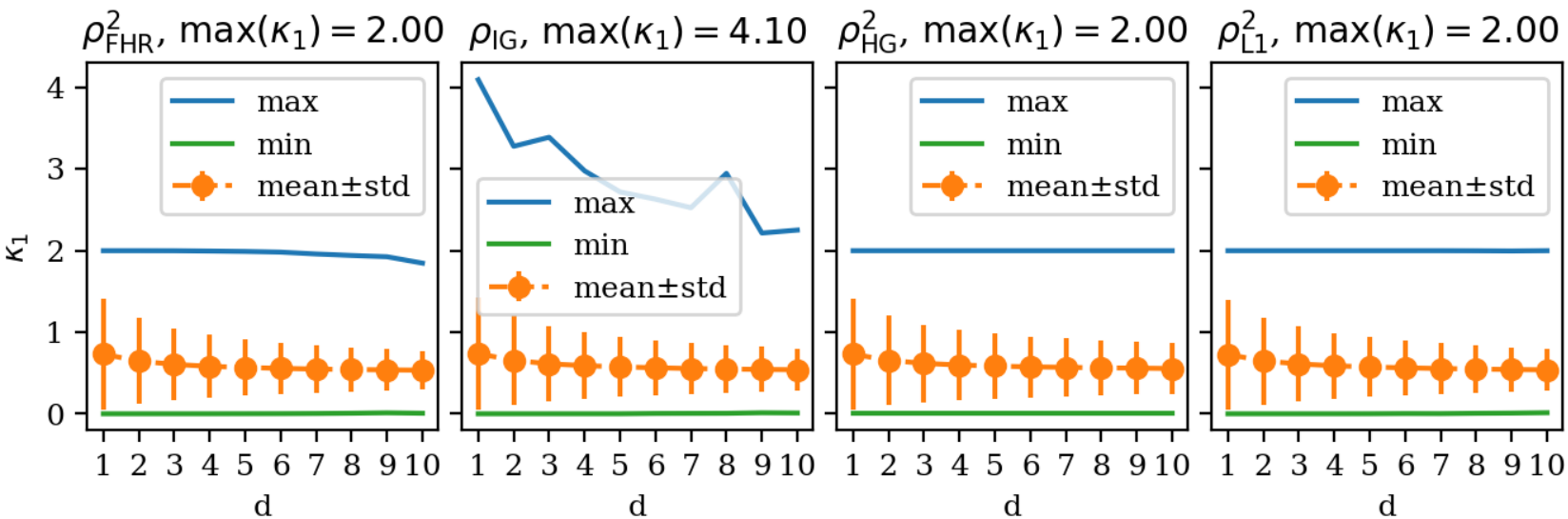
$k = 3$ clusters



$k = 5$ clusters

Experimental evaluation of κ_1

The maximum, mean, standard deviation, and minimum of κ_1 on 10^6 randomly generated tuples (x, y, z) in Δ^d for $d = 1, \ldots, 10$.



k-Center clustering

Cost function for a k-center clustering with centers C ($|C| = k$) is:

$$\boxed{f_D(\Lambda, C) = \max_{p_i \in \Lambda} \min_{c_j \in C} D(p_i : c_j)}$$

For Hilbert simplex clustering, consider the equivalent normed space:

$$\begin{aligned} r^*_{\mathrm{HG}} &= \min_{c \in \Delta^d} \max_{i \in \{1,\ldots,n\}} \rho_{\mathrm{HG}}(p_i, c), \\ r^*_{\mathrm{NH}} &= \min_{v \in V^d} \max_{i \in \{1,\ldots,n\}} \|v_i - v\|_{\mathrm{NH}}. \end{aligned}$$

k-Center clustering: farthest first traversal heuristic

Guaranteed approximation factor of 2 [5]

Algorithm : A 2-approximation of the k-center clustering for any metric distance ρ.

Data: A set Λ; a number k of clusters; a metric distance ρ.
Result: A 2-approximation of the k-center clustering

```
1 begin
2     c_1 ← ARandomPointOf(Λ);
3     C ← {c_1};
4     for i = 2, ··· , k do
5         c_i ← arg max_{p∈Λ} ρ(p, C);
6         C ← C ∪ {c_i};
7 Output C;
```

Smallest Enclosing Ball (SEB)

Given a finite point set $\{p_1, \ldots, p_n\} \in \Delta^d$, the SEB in Hilbert simplex geometry is centered at

$$c^* = \arg\min_{c \in \Delta^d} \max_{i \in \{1,\ldots,n\}} \rho_{\mathrm{HG}}(c, x_i),$$

with radius

$$r^* = \min_{c \in \Delta^d} \max_{i \in \{1,\ldots,n\}} \rho_{\mathrm{HG}}(c, x_i).$$

Decision problem can be solved by Linear Programming (LP), and optimization as a LP-type problem [11]
Do not scale well in high dimensions

Fast approximations for the smallest enclosing ball

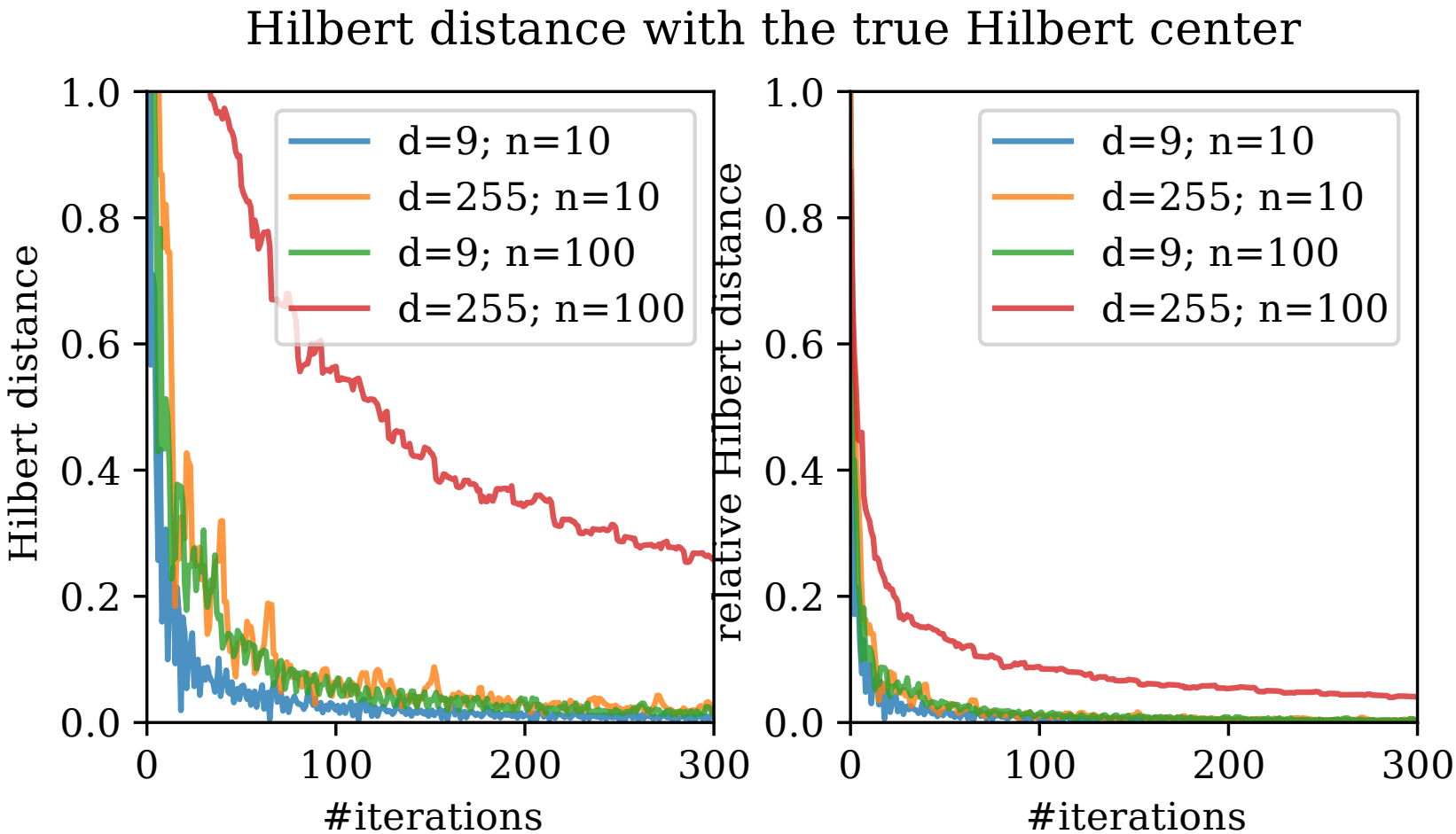
Start from an initial point, compute farthest point to current center, displace center along a geodesic to farthest point, repeat!

Algorithm 4: Geodesic walk for approximating the Hilbert minimax center, generalizing [11]

Data: A set of points $p_1, \cdots, p_n \in \Delta^d$. The maximum number T of iterations.
Result: $c \approx \arg\min_c \max_i \rho_{\mathrm{HG}}(p_i, c)$

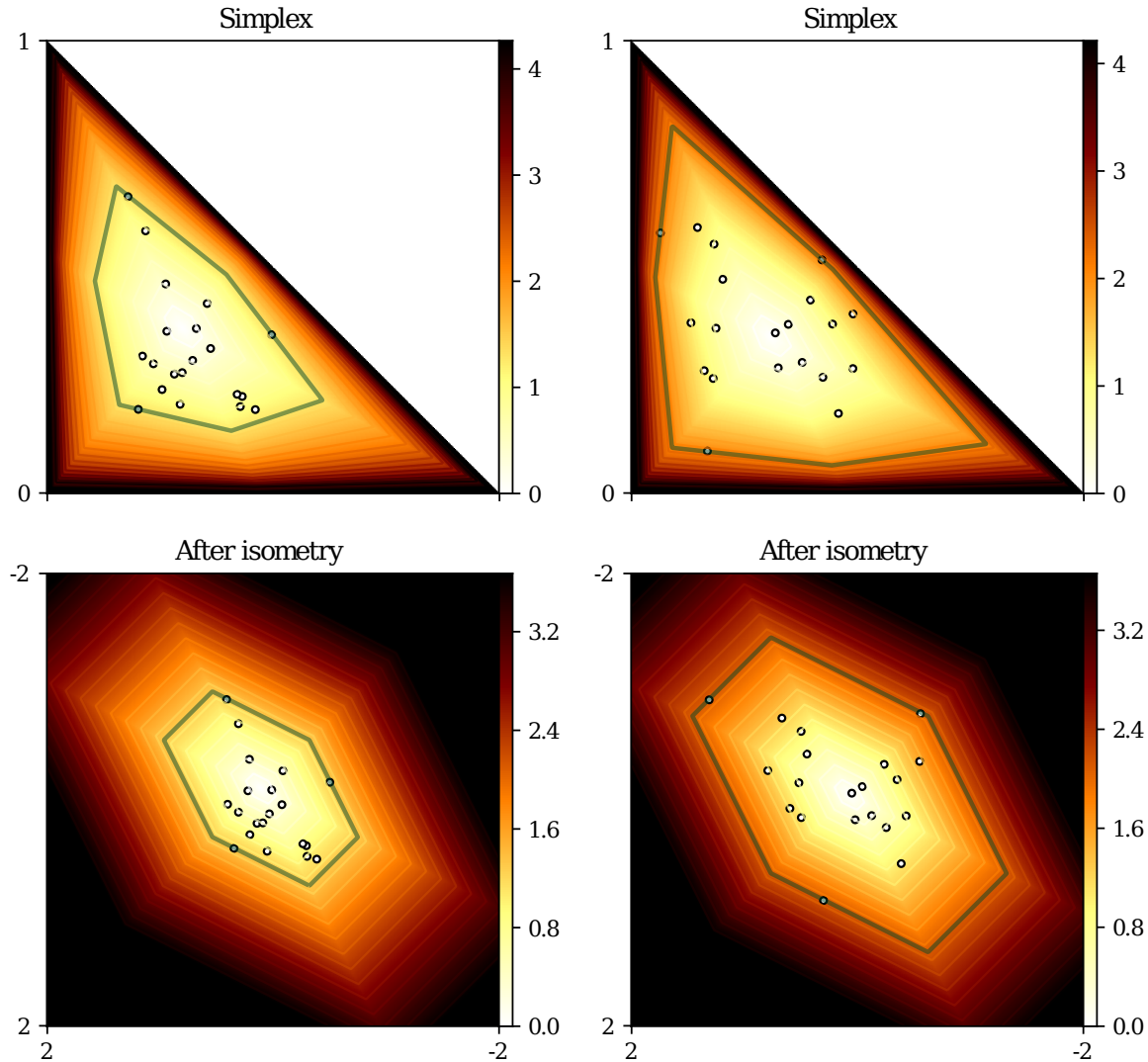
begin
  $c_0 \leftarrow \mathrm{ARandomPointOf}(\{p_1, \cdots, p_n\})$;
  for $t = 1, \cdots, T$ **do**
    $p \leftarrow \arg\max_{p_i} \rho_{\mathrm{HG}}(p_i, c_{t-1})$;
    $c_t \leftarrow c_{t-1} \#^{\rho}_{1/(t+1)} p$;
  Output c_T;

Convergence rate



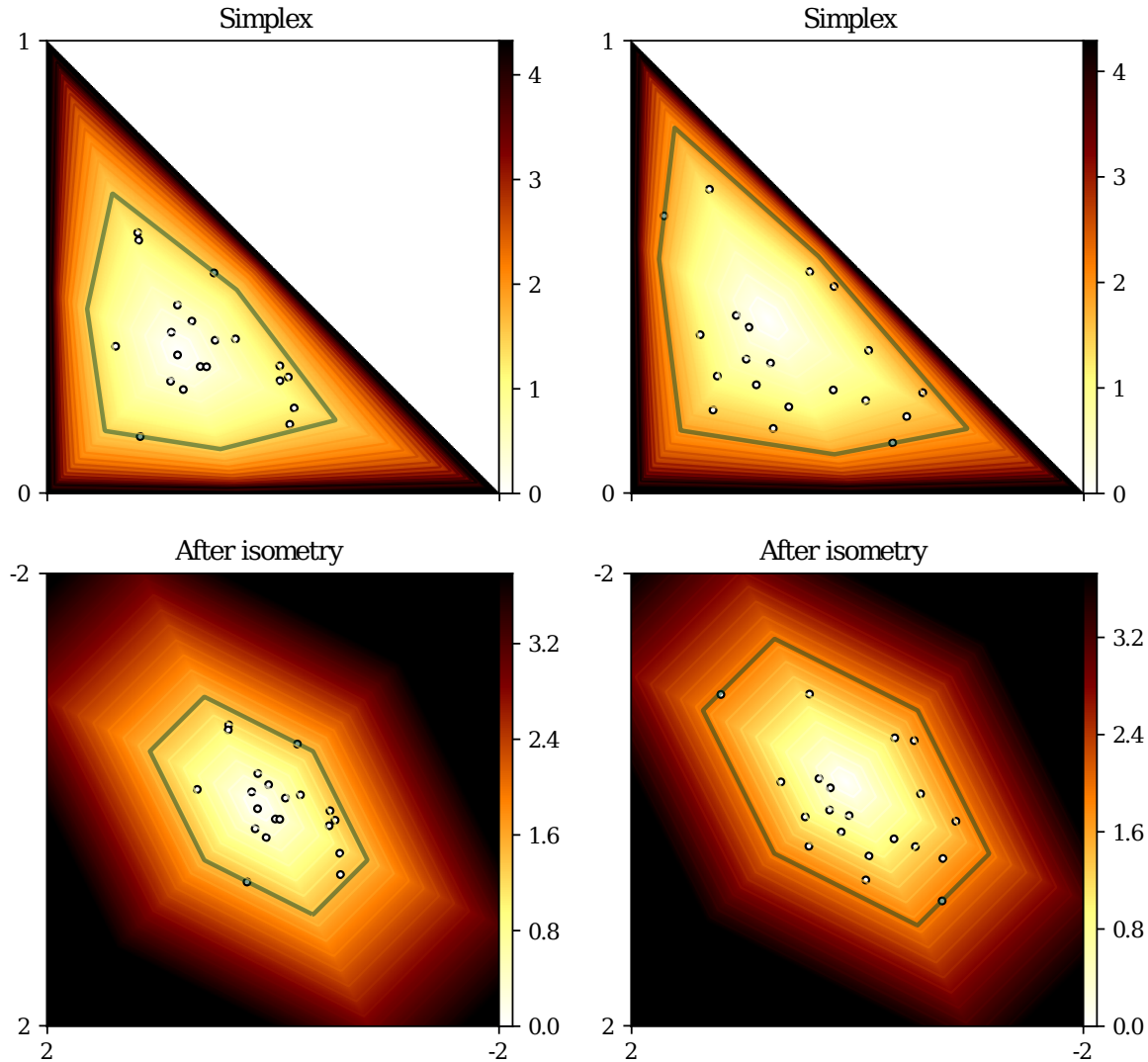
Hilbert distance between the current minimax center and the true center (left) or their Hilbert distance divided by the Hilbert radius of the dataset (right). The plot is based on 100 random points in Δ^9/Δ^{255}.

Examples of smallest enclosing balls in HG and HNG



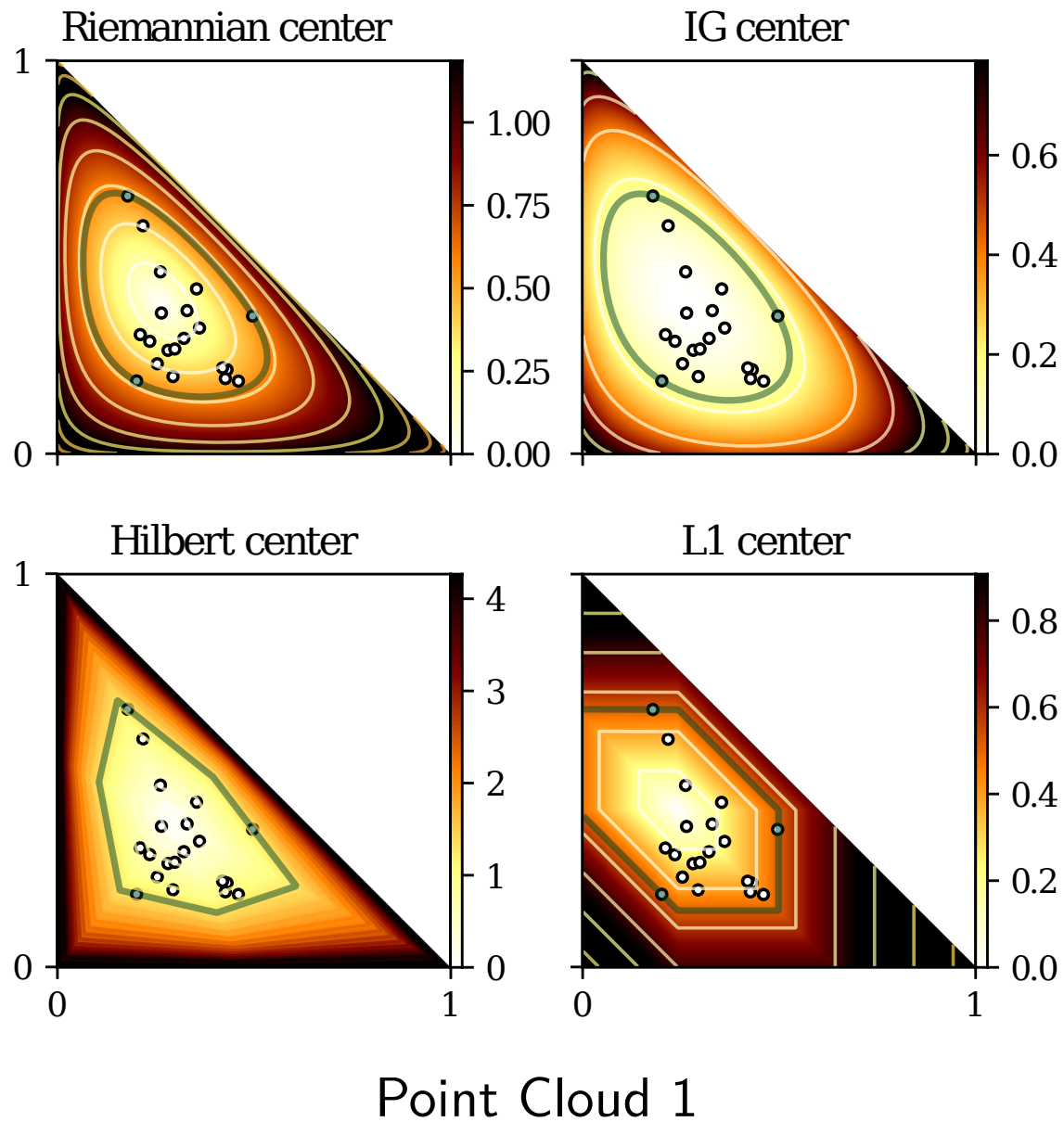
3 points on the border

Examples of smallest enclosing balls in HG and HNG

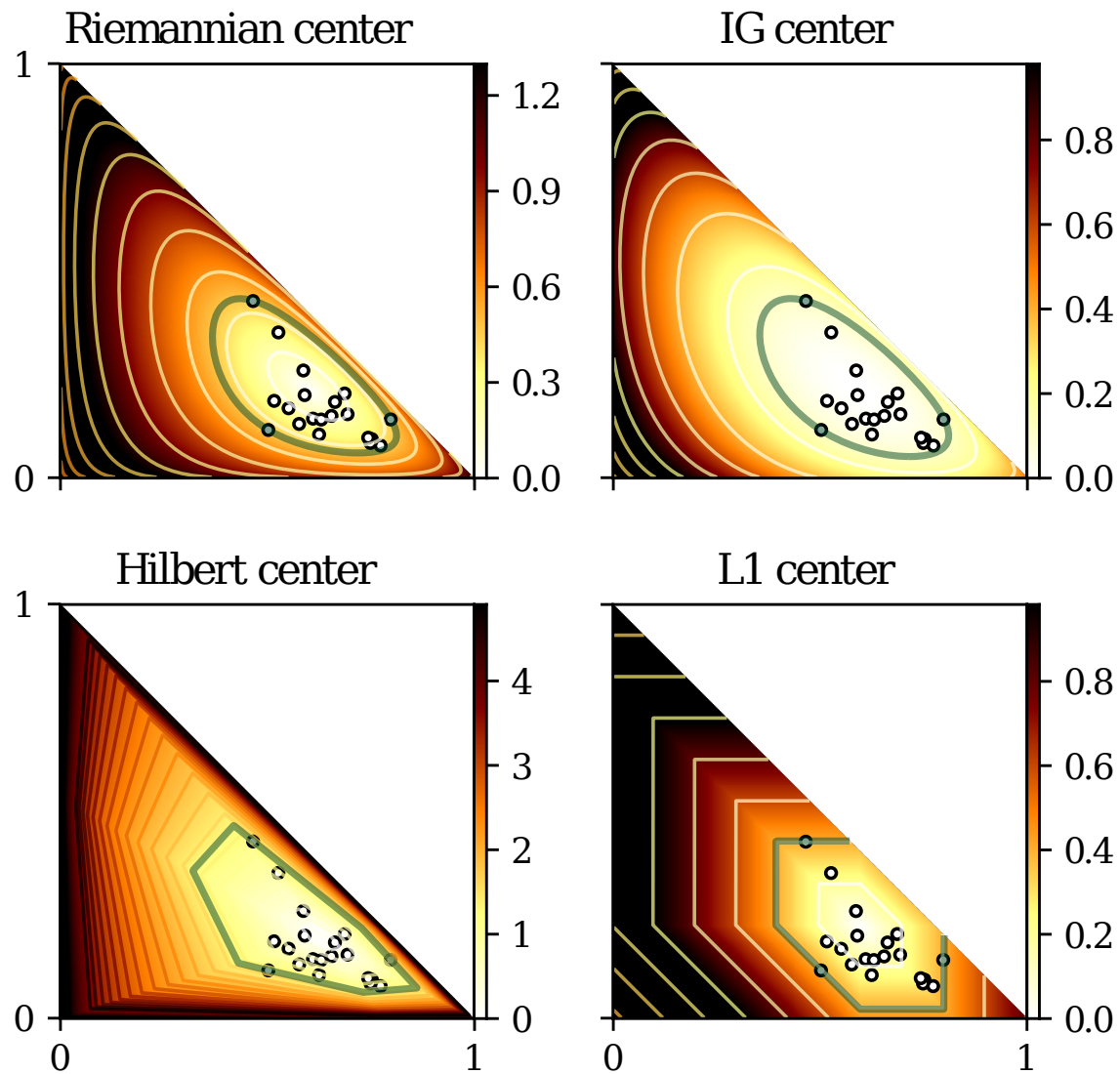


2 points on the border

Visualizing SEBs in the four geometries (1)

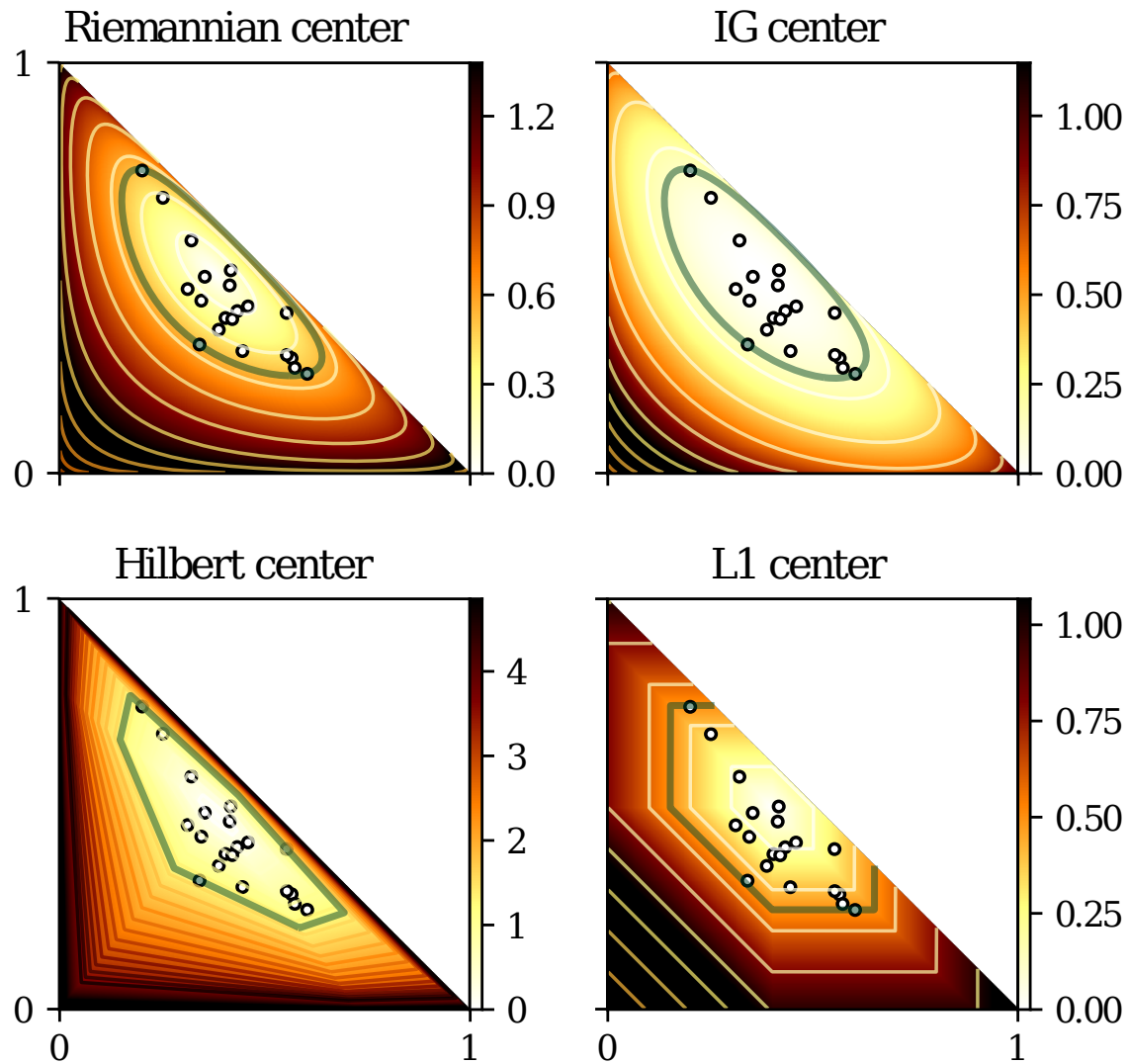


Visualizing smallest enclosing balls (2)



Point Cloud 2

Visualizing smallest enclosing balls (3)



Point Cloud 3

Quantitative experiments: Protocol

- pick uniformly a random center $c = (\lambda_c^0, \ldots, \lambda_c^d) \in \Delta^d$.
- any random sample $p = (\lambda^0, \ldots, \lambda^d)$ associated with c is independently generated by

$$\lambda^i = \frac{\exp(\log \lambda_c^i + \sigma\epsilon^i)}{\sum_{i=0}^d \exp(\log \lambda_c^i + \sigma\epsilon^i)},$$

  where $\sigma > 0$= noise level parameter, and each ϵ^i follows independently a standard Gaussian distribution (generator 1) or the Student's t-distribution (5 dof, generator 2).

- perform clustering based on the configurations $n \in \{50, 100\}$, $d \in \{9, 255\}$, $\sigma \in \{0.5, 0.9\}$, $\rho \in \{\rho_{\mathrm{FHR}}, \rho_{\mathrm{IG}}, \rho_{\mathrm{HG}}, \rho_{\mathrm{EUC}}, \rho_{\mathrm{L1}}\}$. The number of clusters k is set to the ground truth.
- repeat the clustering experiment based on 300 different random datasets. The performance is measured by the **normalized mutual information** (NMI, the larger the better).

Quantitative experiments of k-means++

k	n	d	σ	ρ_{FHR}	ρ_{IG}	ρ_{HG}	ρ_{EUC}	ρ_{L1}
3	50	9	0.5	0.76 ± 0.22	0.76 ± 0.24	$\mathbf{0.81 \pm 0.22}$	0.64 ± 0.23	0.70 ± 0.22
			0.9	0.44 ± 0.20	0.44 ± 0.20	$\mathbf{0.57 \pm 0.22}$	0.31 ± 0.17	0.38 ± 0.18
		255	0.5	0.80 ± 0.24	0.81 ± 0.24	$\mathbf{0.88 \pm 0.21}$	0.74 ± 0.25	0.79 ± 0.24
			0.9	0.65 ± 0.27	0.66 ± 0.28	$\mathbf{0.72 \pm 0.27}$	0.46 ± 0.24	0.63 ± 0.27
	100	9	0.5	0.76 ± 0.22	0.76 ± 0.21	$\mathbf{0.82 \pm 0.22}$	0.60 ± 0.21	0.69 ± 0.23
			0.9	0.42 ± 0.19	0.41 ± 0.18	$\mathbf{0.54 \pm 0.22}$	0.27 ± 0.14	0.34 ± 0.16
		255	0.5	0.82 ± 0.23	0.82 ± 0.24	$\mathbf{0.89 \pm 0.20}$	0.74 ± 0.24	0.80 ± 0.25
			0.9	0.66 ± 0.26	0.66 ± 0.28	$\mathbf{0.72 \pm 0.26}$	0.45 ± 0.25	0.64 ± 0.27
5	50	9	0.5	0.75 ± 0.14	0.74 ± 0.15	$\mathbf{0.81 \pm 0.13}$	0.61 ± 0.13	0.68 ± 0.13
			0.9	0.44 ± 0.13	0.42 ± 0.13	$\mathbf{0.55 \pm 0.15}$	0.31 ± 0.11	0.36 ± 0.12
		255	0.5	0.83 ± 0.15	0.83 ± 0.15	$\mathbf{0.88 \pm 0.14}$	0.77 ± 0.16	0.82 ± 0.15
			0.9	0.71 ± 0.17	0.70 ± 0.19	$\mathbf{0.75 \pm 0.17}$	0.50 ± 0.17	0.68 ± 0.18
	100	9	0.5	0.74 ± 0.13	0.74 ± 0.14	$\mathbf{0.80 \pm 0.14}$	0.60 ± 0.13	0.67 ± 0.13
			0.9	0.42 ± 0.11	0.40 ± 0.12	$\mathbf{0.55 \pm 0.15}$	0.29 ± 0.09	0.35 ± 0.11
		255	0.5	0.83 ± 0.14	0.83 ± 0.15	$\mathbf{0.88 \pm 0.13}$	0.77 ± 0.15	0.81 ± 0.15
			0.9	0.69 ± 0.18	0.69 ± 0.18	$\mathbf{0.73 \pm 0.17}$	0.48 ± 0.17	0.67 ± 0.18

generator 1

Quantitative experiments of k-means++

k	n	d	σ	ρ_{FHR}	ρ_{IG}	ρ_{HG}	ρ_{EUC}	ρ_{L1}
3	50	9	0.5	0.62 ± 0.22	0.60 ± 0.22	$\mathbf{0.71 \pm 0.23}$	0.45 ± 0.20	0.54 ± 0.22
			0.9	0.29 ± 0.17	0.27 ± 0.16	$\mathbf{0.39 \pm 0.19}$	0.17 ± 0.13	0.25 ± 0.15
		255	0.5	0.70 ± 0.25	0.69 ± 0.26	$\mathbf{0.74 \pm 0.25}$	0.37 ± 0.29	0.70 ± 0.26
			0.9	$\mathbf{0.42 \pm 0.25}$	0.35 ± 0.20	0.40 ± 0.19	0.03 ± 0.08	$\mathbf{0.44 \pm 0.26}$
	100	9	0.5	0.63 ± 0.22	0.61 ± 0.22	$\mathbf{0.71 \pm 0.22}$	0.46 ± 0.19	0.56 ± 0.20
			0.9	0.29 ± 0.15	0.26 ± 0.14	$\mathbf{0.38 \pm 0.20}$	0.18 ± 0.12	0.24 ± 0.14
		255	0.5	0.71 ± 0.26	0.69 ± 0.27	$\mathbf{0.75 \pm 0.25}$	0.31 ± 0.28	0.70 ± 0.27
			0.9	0.41 ± 0.26	0.33 ± 0.20	0.38 ± 0.18	0.02 ± 0.06	$\mathbf{0.43 \pm 0.26}$
5	50	9	0.5	0.64 ± 0.15	0.61 ± 0.14	$\mathbf{0.70 \pm 0.14}$	0.48 ± 0.14	0.57 ± 0.15
			0.9	0.31 ± 0.12	0.29 ± 0.12	$\mathbf{0.41 \pm 0.15}$	0.20 ± 0.09	0.26 ± 0.10
		255	0.5	0.74 ± 0.17	0.72 ± 0.17	$\mathbf{0.77 \pm 0.16}$	0.41 ± 0.20	0.74 ± 0.17
			0.9	0.44 ± 0.17	0.37 ± 0.16	0.44 ± 0.15	0.04 ± 0.06	$\mathbf{0.47 \pm 0.17}$
	100	9	0.5	0.62 ± 0.14	0.61 ± 0.14	$\mathbf{0.71 \pm 0.14}$	0.46 ± 0.13	0.54 ± 0.14
			0.9	0.30 ± 0.10	0.27 ± 0.11	$\mathbf{0.40 \pm 0.13}$	0.19 ± 0.08	0.25 ± 0.09
		255	0.5	0.73 ± 0.18	0.70 ± 0.18	$\mathbf{0.75 \pm 0.16}$	0.37 ± 0.20	0.73 ± 0.17
			0.9	0.43 ± 0.16	0.35 ± 0.14	0.41 ± 0.12	0.03 ± 0.06	$\mathbf{0.46 \pm 0.18}$

generator 2

Quantitative experiments of k-center

k	n	d	σ	ρ_{FHR}	ρ_{IG}	ρ_{HG}	ρ_{EUC}	ρ_{L1}
3	50	9	0.5	0.87 ± 0.19	0.85 ± 0.19	$\mathbf{0.92 \pm 0.16}$	0.72 ± 0.22	0.80 ± 0.20
			0.9	0.54 ± 0.21	0.51 ± 0.21	$\mathbf{0.70 \pm 0.23}$	0.36 ± 0.17	0.44 ± 0.19
		255	0.5	0.93 ± 0.16	0.92 ± 0.18	$\mathbf{0.95 \pm 0.14}$	0.89 ± 0.18	0.90 ± 0.19
			0.9	0.76 ± 0.24	0.72 ± 0.26	$\mathbf{0.82 \pm 0.24}$	0.50 ± 0.28	0.76 ± 0.25
	100	9	0.5	0.88 ± 0.17	0.86 ± 0.18	$\mathbf{0.93 \pm 0.14}$	0.70 ± 0.20	0.80 ± 0.20
			0.9	0.53 ± 0.20	0.49 ± 0.19	$\mathbf{0.70 \pm 0.22}$	0.33 ± 0.14	0.41 ± 0.18
		255	0.5	0.93 ± 0.16	0.92 ± 0.17	$\mathbf{0.95 \pm 0.13}$	0.88 ± 0.19	0.93 ± 0.16
			0.9	0.81 ± 0.22	0.75 ± 0.24	$\mathbf{0.83 \pm 0.22}$	0.47 ± 0.28	0.79 ± 0.22
5	50	9	0.5	0.82 ± 0.13	0.81 ± 0.13	$\mathbf{0.89 \pm 0.12}$	0.67 ± 0.13	0.75 ± 0.13
			0.9	0.50 ± 0.13	0.47 ± 0.13	$\mathbf{0.66 \pm 0.15}$	0.34 ± 0.11	0.40 ± 0.12
		255	0.5	$\mathbf{0.92 \pm 0.11}$	$\mathbf{0.91 \pm 0.12}$	$\mathbf{0.93 \pm 0.11}$	0.87 ± 0.13	$\mathbf{0.92 \pm 0.12}$
			0.9	0.77 ± 0.15	0.71 ± 0.17	$\mathbf{0.85 \pm 0.17}$	0.54 ± 0.19	0.74 ± 0.16
	100	9	0.5	0.83 ± 0.12	0.81 ± 0.13	$\mathbf{0.89 \pm 0.11}$	0.67 ± 0.11	0.76 ± 0.13
			0.9	0.48 ± 0.12	0.46 ± 0.12	$\mathbf{0.66 \pm 0.15}$	0.33 ± 0.09	0.39 ± 0.10
		255	0.5	$\mathbf{0.93 \pm 0.10}$	$\mathbf{0.92 \pm 0.11}$	$\mathbf{0.94 \pm 0.09}$	0.89 ± 0.11	0.92 ± 0.11
			0.9	0.81 ± 0.14	0.74 ± 0.15	$\mathbf{0.84 \pm 0.16}$	0.52 ± 0.19	0.79 ± 0.14

generator 1

Quantitative experiments of k-center

k	n	d	σ	ρ_{FHR}	ρ_{IG}	ρ_{HG}	ρ_{EUC}	ρ_{L1}
3	50	9	0.5	0.68 ± 0.22	0.67 ± 0.22	$\mathbf{0.80 \pm 0.20}$	0.48 ± 0.22	0.60 ± 0.22
			0.9	0.32 ± 0.18	0.29 ± 0.17	$\mathbf{0.45 \pm 0.21}$	0.20 ± 0.14	0.26 ± 0.15
		255	0.5	0.79 ± 0.24	0.75 ± 0.24	$\mathbf{0.82 \pm 0.22}$	0.13 ± 0.23	$\mathbf{0.81 \pm 0.24}$
			0.9	0.35 ± 0.27	0.35 ± 0.21	$\mathbf{0.42 \pm 0.19}$	0.00 ± 0.02	0.32 ± 0.30
	100	9	0.5	0.66 ± 0.22	0.65 ± 0.22	$\mathbf{0.79 \pm 0.21}$	0.45 ± 0.19	0.59 ± 0.20
			0.9	0.30 ± 0.16	0.28 ± 0.14	$\mathbf{0.42 \pm 0.19}$	0.20 ± 0.12	0.26 ± 0.14
		255	0.5	0.78 ± 0.25	0.76 ± 0.24	$\mathbf{0.82 \pm 0.21}$	0.05 ± 0.14	0.77 ± 0.27
			0.9	0.29 ± 0.28	0.29 ± 0.20	$\mathbf{0.39 \pm 0.20}$	0.00 ± 0.02	0.22 ± 0.25
5	50	9	0.5	0.69 ± 0.14	0.66 ± 0.14	$\mathbf{0.77 \pm 0.13}$	0.50 ± 0.13	0.61 ± 0.14
			0.9	0.34 ± 0.12	0.30 ± 0.12	$\mathbf{0.46 \pm 0.15}$	0.22 ± 0.09	0.28 ± 0.10
		255	0.5	$\mathbf{0.80 \pm 0.15}$	0.76 ± 0.15	$\mathbf{0.82 \pm 0.14}$	0.24 ± 0.23	$\mathbf{0.81 \pm 0.14}$
			0.9	0.42 ± 0.21	0.38 ± 0.16	$\mathbf{0.46 \pm 0.15}$	0.00 ± 0.02	0.39 ± 0.22
	100	9	0.5	0.66 ± 0.13	0.64 ± 0.14	$\mathbf{0.77 \pm 0.14}$	0.47 ± 0.13	0.57 ± 0.13
			0.9	0.31 ± 0.11	0.28 ± 0.10	$\mathbf{0.44 \pm 0.13}$	0.21 ± 0.08	0.25 ± 0.09
		255	0.5	$\mathbf{0.80 \pm 0.16}$	0.76 ± 0.15	$\mathbf{0.82 \pm 0.13}$	0.12 ± 0.17	$\mathbf{0.81 \pm 0.16}$
			0.9	0.32 ± 0.19	0.30 ± 0.15	$\mathbf{0.41 \pm 0.13}$	0.00 ± 0.01	0.26 ± 0.18

generator 2

Discussion

Experimentally, we observe that

$$\mathrm{HG} > \mathrm{FHR} > \mathrm{IG}$$

- HG even better for large amount of noise
- Intuitively, HG balls are more compact and better capture the clustering structure
- Not surprisingly, Euclidean geometry gets the poorest score!

Cramér-Rao vs Hammersley-Chapman-Robbins lower bounds for statistical estimations

- Cramér-Rao lower bound [8] for an unbiased estimator:

$$V_\theta[T(X)] \succ \mathcal{I}^{-1}(\theta)$$

The FIM is not defined for non-differentiable pdfs, and therefore the Cramér-Rao lower bound does not appply in that case.

- Better Hammersley-Chapman-Robbins Lower Bound [6, 3]:

$$\begin{aligned} V_\theta[T(X)] &\geq \sup_{\Delta} \frac{\Delta^2}{E_\theta\left[\left(\frac{p(x;\theta+\Delta)-p(x;\theta)}{p(x;\theta)}\right)^2\right]}, \\ &\geq \sup_{\Delta} \frac{\Delta^2}{\chi^2(P(x;\theta+\Delta):P(x;\theta))}. \end{aligned}$$

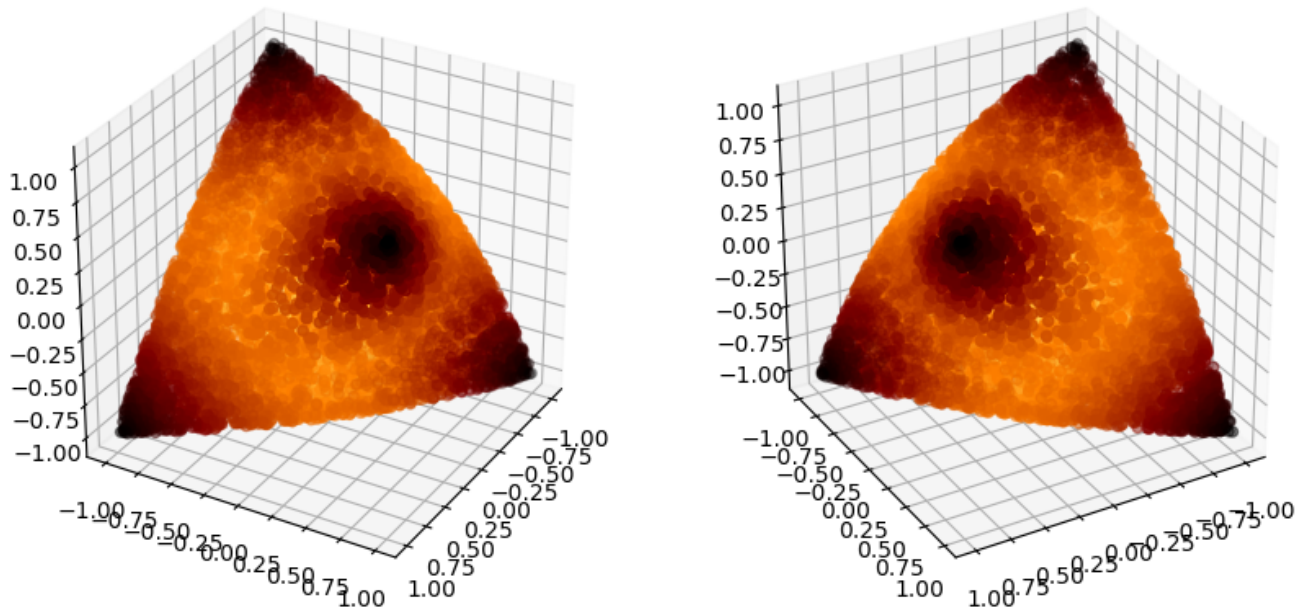
with $\chi^2(P:Q) = \int \left(\frac{\mathrm{d}P-\mathrm{d}Q}{\mathrm{d}Q}\right)^2 \mathrm{d}Q$

A Hilbert geometry for the elliptope, space of correlation matrices

Elliptope: Space of correlation matrices

$$\boxed{\mathcal{C}^d = \{C_{d\times d} \ : \ C \succ 0; C_{ii} = 1, \forall i\}}$$

$\mathcal{C}$ is convex: If $C_1, C_2 \in \mathcal{C}$, then $(1-\lambda)C_1 + \lambda C_2 \in \mathcal{C}$ for $0 < \lambda < 1$.



Distances between correlation matrices

- Hilbert distance in the elliptope:

$$\rho_{\mathrm{HG}}(C_1, C_2) = \left| \log \frac{\|C_1 - C_2'\| \|C_1' - C_2\|}{\|C_1 - C_1'\| \|C_2 - C_2'\|} \right|.$$

  with C_1' and C_2' the intersection correlation matrices with the boundary of the elliptope (no closed form, bisection method in practice)

- Fröbenius distance, L2 norm (Euclidean geometry)

- L1 norm

- Log-det divergence:

$$\rho_{\mathrm{LD}}(C_1, C_2) = tr(C_1 C_2^{-1}) - \log |C_1 C_2^{-1}| - d.$$

Experiments: k-means++ on correlation matrices

$n = 100$ 3×3 matrices with $k = 3$ clusters (points in $\mathcal{C}_3$, cluster of almost identical size)

Each cluster is independently generated according to

$$P \sim \mathcal{W}^{-1}(I_{3\times 3}, \nu_1),$$
$$C_i \sim \mathcal{W}^{-1}(P, \nu_2),$$

where $\mathcal{W}^{-1}(A, \nu)$ = inverse Wishart distribution with scale matrix A and ν degrees of freedom

C_i is a point in the cluster associated with P.

500 independent runs

ν_1	ν_2	ρ_{HG}	ρ_{EUC}	ρ_{L1}	ρ_{LD}
4	10	$\mathbf{0.62 \pm 0.22}$	0.57±0.21	0.56±0.22	0.58±0.22
4	30	$\mathbf{0.85 \pm 0.18}$	0.80±0.20	0.81±0.19	0.82±0.20
4	50	$\mathbf{0.89 \pm 0.17}$	0.87±0.17	0.86±0.18	0.88±0.18
5	10	$\mathbf{0.50 \pm 0.21}$	0.49±0.21	0.48±0.20	0.47±0.21
5	30	$\mathbf{0.77 \pm 0.20}$	0.75±0.21	0.75±0.21	0.75±0.21
5	50	$\mathbf{0.84 \pm 0.19}$	0.82±0.19	0.82±0.20	$\mathbf{0.84 \pm 0.18}$

Summary and conclusion [15]

- Introduced Hilbert's projective geometry for the space of multinomials (probability simplex) and the space of correlation matrices (elliptope)
- Gave a fast linear time algorithm for computing Hilbert distance in the probability simplex
- Hilbert simplex geometry is isometric to a normed vector space
- Benchmark experimentally four types of geometry for k-means and k-center clusterings in Hilbert simplex geometry
- Hilbert geometry experimentally well-suited for k-means and k-center clustering

References I

Ingemar Bengtsson and Karol Życzkowski.

Geometry of quantum states: an introduction to quantum entanglement.
Cambridge university press, 2017.

L. L. Campbell.

An extended čencov characterization of the information metric.
American Mathematical Society, 98(1):135–141, 1986.

Douglas G Chapman and Herbert Robbins.

Minimum variance estimation without regularity assumptions.
The Annals of Mathematical Statistics, pages 581–586, 1951.

N. N. Chentsov.

Statistical decision rules and optimal inference.
Monographs, American Mathematical Society, Providence, RI, 1982.

Teofilo F Gonzalez.

Clustering to minimize the maximum intercluster distance.
Theoretical Computer Science, 38:293–306, 1985.

JM Hammersley.

On estimating restricted parameters.
Journal of the Royal Statistical Society. Series B (Methodological), 12(2):192–240, 1950.

Bas Lemmens and Roger Nussbaum.

Birkhoff's version of Hilbert's metric and its applications in analysis.
Handbook of Hilbert Geometry, pages 275–303, 2014.

Frank Nielsen.

Cramér-Rao lower bound and information geometry.
In *Connected at Infinity II*, pages 18–37. Springer, 2013.

References II

Frank Nielsen.

An elementary introduction to information geometry.
ArXiv e-prints, August 2018.

Frank Nielsen, Boris Muzellec, and Richard Nock.

Classification with mixtures of curved Mahalanobis metrics.
In *IEEE International Conference on Image Processing (ICIP)*, pages 241–245, 2016.

Frank Nielsen and Richard Nock.

On the smallest enclosing information disk.
Information Processing Letters, 105(3):93–97, 2008.

Frank Nielsen and Richard Nock.

Total Jensen divergences: Definition, properties and k-means++ clustering, 2013.
arXiv:1309.7109 [cs.IT].

Frank Nielsen and Richard Nock.

Further heuristics for k-means: The merge-and-split heuristic and the (k, l)-means.
arXiv preprint arXiv:1406.6314, 2014.

Frank Nielsen and Laëtitia Shao.

On balls in a polygonal Hilbert geometry.
In *33st International Symposium on Computational Geometry (SoCG 2017)*. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017.

Frank Nielsen and Ke Sun.

Clustering in Hilbert simplex geometry.
CoRR, abs/1704.00454, 2017.